

## Chapter 3 – 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.

We can study relationships between two categorical variables like texting while driving and having a car accident. We can also study relationships between two quantitative variables like the weight of a person and their blood pressure. A third relationship we can study is the relationship between a categorical variable and a quantitative variable. For example, we can study the relationship between the type of job you have and your annual income. 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 advanced statistics students do not understand the Chi-Square distribution because they lack the foundational understanding of contingency tables and analyzing the 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, while "relationship" and "association" are used for two categorical variables or for a categorical - quantitative relationship study.*

*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 3A – Contingency Tables with Technology

When studying relationships between categorical variables, we start by creating a contingency table. Some people refer to this table as a “two-way” table, but contingency table is more common. A contingency 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 variable with two options (male and female). Blood type is another categorical variable with four options (A, B, AB, and O). To make a contingency 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 contingency table with raw data and StatKey**

Let us look at an example. Go to [www.matt-teachout.org](http://www.matt-teachout.org) or Canvas and click on the “math 075 survey data fall 2015”. We want to explore the relationship between the campus a person goes to and their political party.



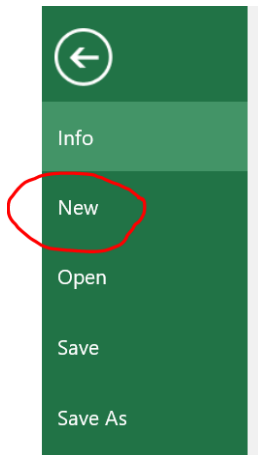
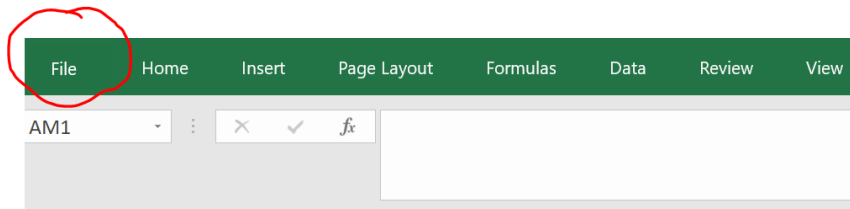
A	B	C	D	E
Campus	Gender	Age (in years)	Month of birthday	Weight (in pounds)
Canyon Country Campus	Female	20	4	144
Canyon Country Campus	Female	19	3	120
Canyon Country Campus	Female	50	10	135
Canyon Country Campus	Male	22	1	155
Canyon Country Campus	Female	25	6	125
Valencia Campus	Male	18	10	180
Valencia Campus	Male	20	5	155
Valencia Campus	Male	19	6	172
Valencia Campus	Female	19	5	135
Valencia Campus	Female	17	10	149
Valencia Campus	Female	18	11	106
Valencia Campus	Male	19	4	165
Valencia Campus	Male	18	12	250

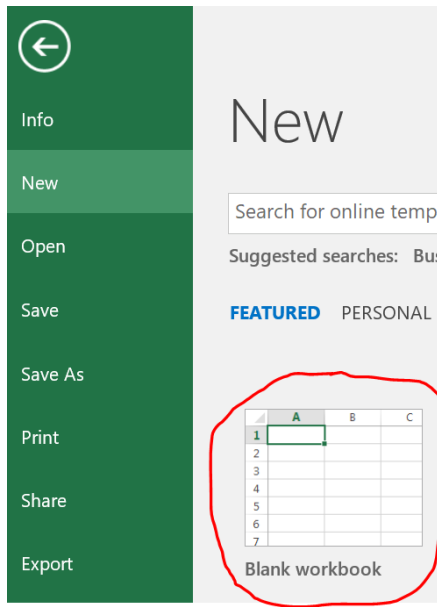


*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

Z	AA	AB
Number alcoholic beverages (per week on average)	Political Party	Math Intimidation? (on scale of 1-10)
0	Democratic	10
0	Democratic	7
1	Democratic	8
0	Other	4
6	Other	6
0	Other	3
0	Republican	3
0	Republican	7
0	Democratic	7
0	Republican	5
0	Democratic	7
0	Democratic	1
0	Democratic	2
0	Republican	7
0	Other	6
0	Independent	7

First, we will need to check the data. When exploring relationships between two data sets, the data needs to be ordered pair. This usually means the data came from the same people. In this data values in the same row came from the same math 075 pre-stat student. We also need to be careful of blanks. This means a person did not answer one or both of the questions. Start by copy and pasting the campus data and political party data into a new spreadsheet. In excel, you would go to the "file" menu on the top left corner and then press "new" and click on the "blank workbook".





A good rule of thumb is never mess up an original data set. Always copy and paste into a new excel file if you want to change things. The two columns of categorical data need to be in next to each other in the new spreadsheet. Otherwise, StatKey will not accept it. The larger the data set, the more difficult it is to copy and paste columns of data. In Excel, if you hold your cursor at the title at the top of the column you will see it turn into a downward arrow ↓. When you see the downward arrow, left click and it will highlight the entire column. You can also click and drag, but the larger the data set, the longer it will take to drag to the very bottom. I prefer the downward arrow method. Once the data set is highlighted, hold the control key down and your keyboard and push “C”. Control C is an easy way to copy. You can also right click and push copy. Your new spreadsheet should now look like this. Be careful. Did you copy and paste all of the data? Your columns should go all the way to row 460!

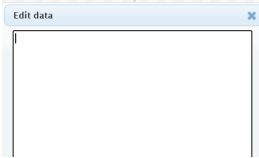
	A	B
1	<b>Campus</b>	<b>Political Party</b>
2	Canyon Country Campus	Democratic
3	Canyon Country Campus	Democratic
4	Canyon Country Campus	Democratic
5	Canyon Country Campus	Other
6	Canyon Country Campus	Other
7	Valencia Campus	Other
8	Valencia Campus	Republican
9	Valencia Campus	Republican
10	Valencia Campus	Democratic
454	Canyon Country Campus	Democratic
455	Canyon Country Campus	Democratic
456	Canyon Country Campus	Republican
457	Canyon Country Campus	Independent
458	Canyon Country Campus	Democratic
459	Canyon Country Campus	Democratic
460	Canyon Country Campus	Independent
461		

Go through the data and make sure there are no blanks. If there is a blank, delete that entire row. If you remember from chapter 1, this is called non-response bias. This process of deleting out missing cells is sometimes called “cleaning the data”. This data did not seem to have any blanks that needed deleting.



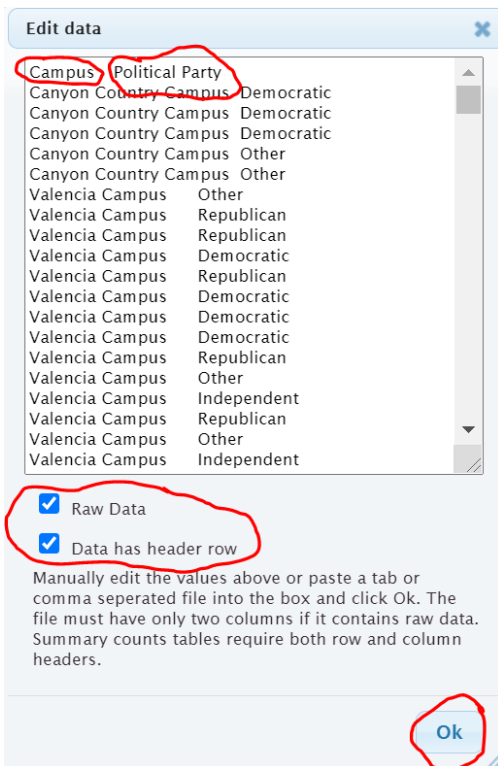
*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

To make a contingency table with StatKey, go to [www.lock5stat.com](http://www.lock5stat.com) and click the “StatKey” button. Now click on “Two Categorical Variables” under the “Descriptive Statistics and Graphs” menu. Click on the “edit data” button. If there is data already there, push “Control A” on your keyboard and then delete. Make sure your cursor is at the very top of the edit data field.



Now we will copy and paste our data into StatKey. Remember to hold your cursor right above the column you want to copy until you see the downward arrow and then left click. Hold the control key down on your keyboard and do the same thing for the second column. Now push “Control C”. Both columns are now copied together.

Then go back to the “edit data” field in the “Two Categorical Variables” menu in StatKey and paste the columns into StatKey.



It is important to know what type of data you have put in. The data we pasted is not a list of the summary counts. This data is the actual column of words. We call that “raw categorical data”. So we will need to check the box that says “raw data”. It is also to note whether the titles are included in the data we pasted. We see the titles “campus” and “political party” at the top of our data. StatKey refers to titles as “header rows”. Since our titles are there, we will need to check the box that says “data has header row”. Now push “OK”.



Note: If the data did not have the titles, we should uncheck the box that says “data has a header row”. If this was summary counts of our categorical data and not the actual column of words, we would also uncheck the box that says “raw data”.

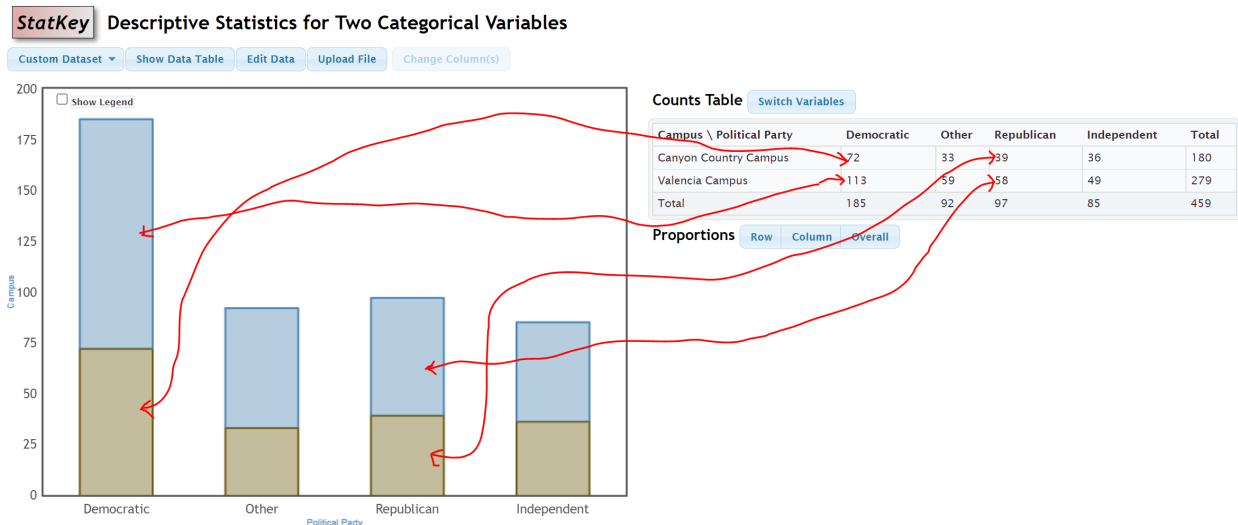
You should now see the contingency table. StatKey refers to the table as a “Counts Table”. Notice StatKey has counted all of the categorical variables for us. We know there were 72 democrats that went to the Canyon Country campus. We know there was 58 republican students that went to the Valencia campus. We know there was a total of 92 students that identified as supporting a political party other than democrat, republican or independent. We also see the grand total of 459 students. (Remember your columns of data had 460 rows. That is because the first row was the title.)

**Counts Table** [Switch Variables](#)

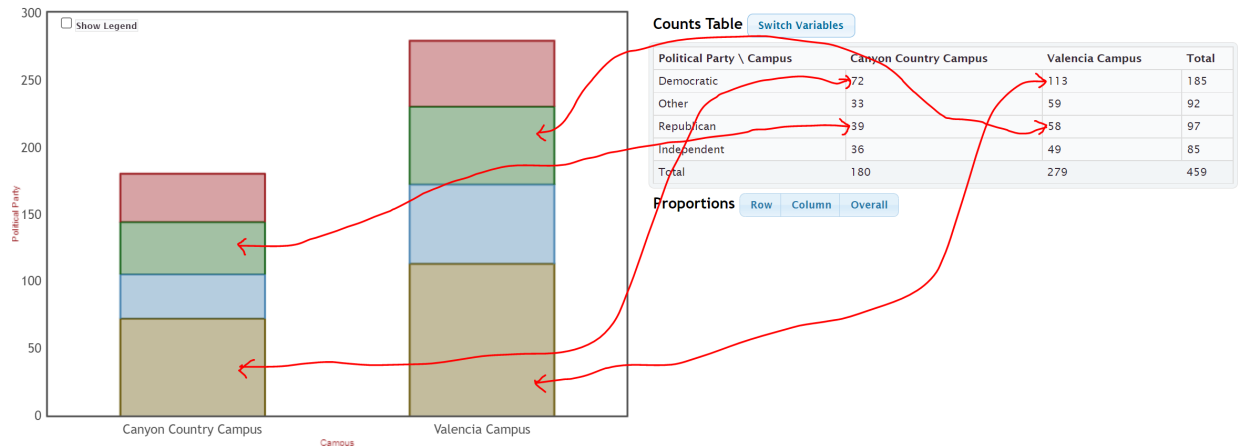
Campus \ Political Party	Democratic	Other	Republican	Independent	Total
Canyon Country Campus	72	33	39	36	180
Valencia Campus	113	59	58	49	279
Total	185	92	97	85	459

The size of a contingency table is the number of rows by the number of columns. Totals are not included. This table has two rows (CCC and Valencia) and four columns (Independent, Other, Republican, and Democratic), so this is a “2 by 4” or “2x4” contingency table.

StatKey has several cool features with the contingency table. Notice it has created a stacked bar chart. This stacked bar chart gives a visual representation of a contingency table. Notice if you place your cursor on any section of the graph the corresponding count lights up in the contingency table.



Another feature is the “switch variables” button in StatKey. Clicking on this button will switch the rows and columns. So the rows will now be political party and the columns will be campus. The stacked bar chart will also adjust to the new switched contingency table. Notice all of the counts are the same.



Most statistics software programs can make contingency tables. Here is what the contingency tables would look like in Statcato. Notice the counts are identical to the tables with StatKey.

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	72	36	33	39	180
Valencia Campus	113	49	59	58	279
All	185	85	92	97	459

	Canyon Country Campus	Valencia Campus	All
Democratic	72	113	185
Independent	36	49	85
Other	33	59	92
Republican	39	58	97
All	180	279	459

### Putting an existing contingency table into StatKey

Suppose you have an existing contingency table that you want to put into StatKey in order to create the stacked bar chart. You can go to “Two categorical Variables” under the “Descriptive Statistics and Graphs” menu. Click on edit data and type in the contingency table. This is no longer “raw data”. It is the counts of the categories. So we will NOT check the box that says “raw data”. Every program has a different way of typing in data. StatKey uses commas.

#### Example

Earlier in this section we created a contingency table by counting patients in terms of their gender and blood type. We can type this contingency table into StatKey using commas. It must be typed in a certain way though.

	Type A	Type B	Type AB	Type O



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*



Female	1	1	2	5
Male	4	1	0	6

Here is what we would type. We need [blank] in the top left corner corresponding to the empty cell in the top left corner. Comma means we are going to the next cell in that row. There should only be one space after the comma. You also want to avoid typing other punctuation marks. Never type in the totals. Those are calculated automatically. This is not raw categorical data. That would be columns of words. This is a contingency table with summary counts. So we should NOT check the box that says "Raw Data". It does have titles at the top, so we will check the box that says "Data has header row".

[blank], Type A, Type B, Type AB, Type O

Female, 1, 1, 2, 5

Male , 4, 1, 0, 6

Edit data
✕

```
[blank], Type A, Type B, Type AB, Type O
Female, 1, 1, 2, 5
Male , 4, 1, 0, 6
```

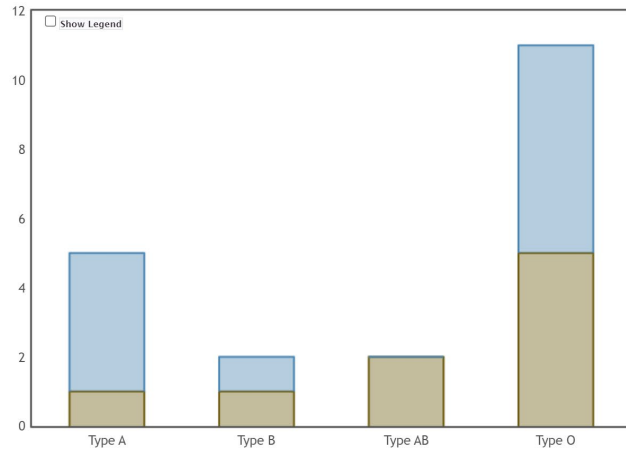
Raw Data
  **Data has header row**

Manually edit the values above or paste a tab or comma seperated file into the box and click Ok. The file must have only two columns if it contains raw data. Summary counts tables require both row and column headers.

Ok



Once we press OK, we have the contingency table in StatKey and the stacked bar chart is automatically created. Notice that even though we did not type them, the totals are now there.

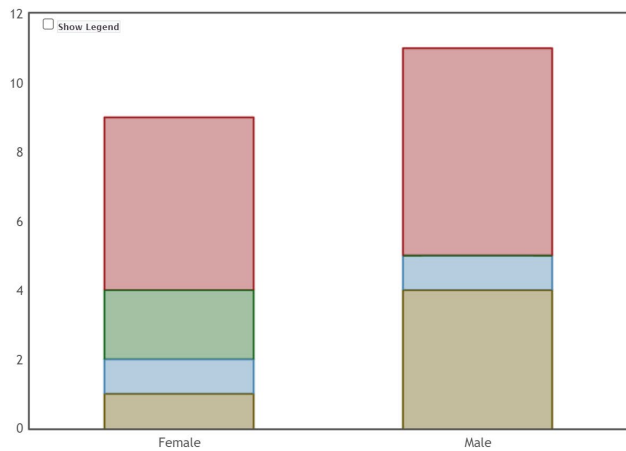


Counts Table [Switch Variables](#)

undefined \ undefined	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	20

Proportions [Row](#) [Column](#) [Overall](#)

We can also press the “Switch Variables” button to switch the blood type to the rows and gender to the columns if we prefer.



Counts Table [Switch Variables](#)

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

Proportions [Row](#) [Column](#) [Overall](#)



### Practice Problems Section 3A

Directions #1-4: Here is some data taken from the medical records department at a local hospital. The data includes 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).

Gender	Blood Type	Rh Factor	Floor
M	A	-	SDS
M	O	+	ER
F	AB	+	Med/Surg
M	O	-	ICU
F	O	+	SDS
F	O	+	Med/Surg
M	A	+	SDS
F	O	+	Med/Surg
F	O	+	ER
M	B	+	SDS
F	A	-	Med/Surg
M	O	+	ICU
M	A	+	Med/Surg
F	O	-	SDS
F	B	+	ICU
M	O	+	ER
F	AB	-	ER
M	O	+	SDS
M	O	+	Med/Surg
M	A	+	ER

1. Create a contingency table that we could use to compare Rh factor (Rh+ or Rh-) to blood type (A,B,AB or O). Make the rows represent the Rh factor and the columns represent the blood type. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?
2. Create a contingency table that we could use to compare gender to the part of the hospital the patient went to. Make the rows represent gender and the columns represent the part of the hospital. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?
3. Create a contingency table that we could use to compare the Rh factor (Rh+ or Rh-) to the part of the hospital the patient went to (SDS, ER, MedSurg, ICU). Make the rows represent the Rh factor and the columns represent the part of the hospital. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?
4. Create a contingency table that we could use to compare the part of the hospital (SDS, ER, MedSurg, ICU) to the blood type (A,B,AB or O). Make the rows represent the blood type and the columns represent the part of the hospital. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?



Directions #5-8: Open the “Math 075 Survey Data Fall 2015” in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Use StatKey to create a contingency table and stacked bar chart for the following variables. Make a rough sketch of the stacked bar chart and table on your paper. Then use the table and graph to answer the questions.

Directions for creating contingency table with StatKey:

- Open the “Math 075 Survey Data Fall 2015”. Copy and paste the two columns next to each other in a new spreadsheet. Then copy both columns together.
- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”. Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then paste in your two columns of data. Check the box that says “Raw Data”. If your data has a title, check the box that says “Data has a header row”. Then push “OK”. If your rows and columns are backward, push the “Switch Variables” button.

5. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for campus (Valencia or Canyon Country) and at least one tattoo (yes or no). Let the rows represent tattoo status and let the columns represent the campus.

- a) Draw a sketch of the contingency table including titles and totals.
- b) Draw a sketch of the stacked bar chart.
- c) What was the grand total?
- d) How many total students went to the Valencia campus?
- e) How many total students have at least one tattoo?
- f) How many students both did not have a tattoo and went to the Canyon Country campus?

6. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for contact lenses or glasses (yes or no) and hair color (brown, black, blond(e), red, other). Let the rows represent glasses/contacts status and the columns represent hair color.

- a) Draw a sketch of the contingency table including titles and totals.
- b) Draw a sketch of the stacked bar chart.
- c) What was the grand total?
- d) How many total students need contacts or glasses?
- e) How many total students have brown hair?
- f) How many students both did not need glasses and have black hair?

7. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for texting while driving (yes or no) and being in a car accident (yes or no). Let the car accident status represent the rows and texting while driving represent the columns.

- a) Draw a sketch of the contingency table including titles and totals.
- b) Draw a sketch of the stacked bar chart.
- c) What was the grand total?
- d) How many total students said that do not text and drive? Do you believe them?



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

- e) How many total students have not been in a car accident?
- f) How many students both text and drive and have been in a car accident?

8. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for live with parents (yes or no) and political party (democrat, republican, independent, other). Let the political party represent the rows and living with parents status represent the columns.

- a) Draw a sketch of the contingency table including titles and totals.
  - b) Draw a sketch of the stacked bar chart.
  - c) What was the grand total?
  - d) How many total students do not live with their parents?
  - e) How many total students identify with independent political party?
  - f) How many students are both democrat and live with their parents?
- 



## Section 3B – Marginal and Joint Percentages from Contingency Tables

Analyzing two categorical data sets involves not only creating contingency 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 (sample size).

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

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

### 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 contingency table created with StatKey from the Fall 2015 Math 075 Survey data. This table describes the relationship between smoking and political party for Math 075 pre-stat students.

#### Counts Table [Switch Variables](#)

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

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 table.

#### Example 1

What percentage of the pre-stat 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. Where do we find the amount of pre-stat students that smoke cigarettes? Smoking cigarettes (yes) is a row, so we should look in the margin at the total for that row.



## Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

Notice the amount and the grand total are found in the margins where the totals are. This is why this is often called a “marginal proportion” or a “marginal percentage”. Notice the marginal percentage only involves one variable (smoking) and does not include political party.

Proportion of students that smoke = Amount of Smokers ÷ Grand Total =  $33 \div 459 = 0.07189524 \approx 0.072$

Percentage of students that smoke  $\approx 0.072 \times 100\% = 7.2\%$

### Example 2

What percentage of the pre-stat students identified as other political party?

Notice we are looking at all of the pre-stat students, so we should use the grand total again as our total.

Where will we find the amount of pre-stat students that support “other” political party? Other political party is a column so we will have to look at the total for that column.

## Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

Proportion of students that support other political party = Amount of other political party ÷ Grand Total =  $92 \div 459 = 0.200435729 \approx 0.200$

Percentage of students that are other political party  $\approx 0.200 \times 100\% = 20.0\%$

Notice we only looked at one variable (other political party), and the amount of students that identified as other political party and the grand total were both found in the margins. So this is again a “marginal percentage”.

Note: Some students may ask why we did not write the answer as 0.2 or 20%. These are equivalent to 0.200 and 20.0%, but these answers tell us that the answer was rounded to three significant figures.

### Formula

**Single Variable Marginal Proportion = Total for Row or Column ÷ Grand Total**



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

## 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 (“AND Joint %)

What percentage of all the pre-stat students both smoked cigarettes 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 only republicans. We need the amount of smoking republicans. This is a classic “AND” proportion since both things need to be true about the student.

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

#### Counts Table

[Switch Variables](#)

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

Notice that to find the smoking republicans, we need to look where the republican column meets the yes smoking row. This is why “AND” proportions are often referred to as an intersection.

Proportion of pre-stat students that both smoke cigarettes and are republican =

$$\text{amount of smoking republicans} \div \text{grand total} = 7 \div 459 \approx 0.015250544 \approx 0.015$$

Percentage of pre-stat students that both smoke and are republican  $\approx 0.015 \times 100\% \approx 1.5\%$

#### Formula

“AND” Intersecting Proportion = Amount where row and column intersect  $\div$  Grand Total

### Example 4 (“OR” Joint %)

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



*This chapter is from [Introduction to Data Analysis](#), first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](#) - 10/1/2017*



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. This is why an OR joint proportions are often referred to as a union. When calculating an “OR” joint proportion, you will need to do some adding to find the amount.

**Counts Table** [Switch Variables](#)

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

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

$$= (90 + 9 + 10 + 7 + 7) \div 459 = 123 \div 459 \approx 0.267973856 \approx 0.268$$

Percentage of students that either smoke or are republican  $\approx 0.268 \times 100\% \approx 26.8\%$

**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 (97), we would have gotten 130 as our amount. This would be wrong. The correct amount was 123. Adding the row and column totals gives you the wrong answer because we would have added the 7 smoking republicans twice.

Here are some other formulas that may be used to calculate an OR (union) proportion.

**Formulas**

“OR” Union Proportion = Add up all of the values in the row or column without using totals  $\div$  Grand Total

“OR” Union Proportion = (Row Total + Column Total – Intersection amount)  $\div$  Grand Total

“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion

In the previous example here is how we could have used the other formulas to get the same answer.

What proportion of the pre-stat students either smoke cigarettes or are republican?

**Counts Table** [Switch Variables](#)

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

“OR” Union Proportion = (Row Total + Column Total – Intersection amount)  $\div$  Grand Total

$$= (97 + 33 - 7) \div 459 = 123 \div 459 = 123 \div 459 \approx 0.267973856 \approx 0.268$$

Percentage of students that either smoke or are republican  $\approx 0.268 \times 100\% \approx 26.8\%$



Notice we got the same answer as before.

**“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion**

= Proportion Smoke + Proportion Republican – Proportion that smoke and are Republican

$$= \frac{33}{459} + \frac{97}{459} - \frac{7}{459} \approx 0.072 + 0.211 - 0.015 = 0.268 = 26.8\%$$

Notice we got the same answer as before. This formula is particularly useful, especially when a statistics program calculates the marginal and intersecting proportions for you.

## Calculating Marginal and Joint Proportions with StatKey

StatKey can calculate the marginal and intersecting proportions for you. Under the “Counts table” (Contingency Table) you will see a “Proportions” menu. Click the button that says “Overall”. We put the smoking and political party columns from the Math 075 Summary Data Fall 2015 into StatKey.

Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

**Proportions** Row Column Overall ←

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

When you click the Overall button, StatKey calculates the marginal proportions and the “AND” intersecting proportions. However, it does not calculate the “OR” union proportions.

Our first example in this section asked what percentage of the pre-stat students smoke cigarettes. Yes (smoking) is a row in this table so we just need to look at the margin (end of the row) to find the answer. Notice it says ) 0.072 or 7.2%. This is the same answer we calculated earlier in the section.



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

## Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

## Proportions Row Column Overall

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

Earlier in this section we asked what percentage of pre-stat students both smoke cigarettes and are republican. To find this answer we just need to go to where Yes (smoking) and Republican intersect. Notice the answer is given as 0.015 or 1.5%. This is again the same answer we calculated earlier in the section.

## Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

## Proportions Row Column Overall

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

Earlier in the section we wanted to find out what percentage of pre-stat students either smoke cigarettes or are republican. StatKey does not calculate "OR" (union) proportions, but we can use the proportions calculated and the following formula.

**"OR" Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting "AND" Proportion**  
 = Proportion Smoke + Proportion Republican – Proportion that smoke and are Republican



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

## Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

**Proportions** Row Column Overall ←

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

The proportion that smoke will be at the end of the Yes (smoke) row. The proportion of republicans will be at the bottom of the "Republican" column. The AND proportion will be where the smoking row and republican column intersect. Notice we got the same answer as before.

$$= 0.072 + 0.211 - 0.015 = 0.268 = 26.8\%$$

Note: Categorical data is often given to a data scientist as a contingency table with summary counts. Most data scientists do not calculate things by hand. Recall that in section 3A, we learned we can type in an existing contingency table into StatKey using commas. Typing the table into StatKey allows us to not only have access to the stacked bar chart, but also the proportion button that can calculate proportions automatically for us.

### Formulas

**Single Variable Marginal Proportion = Total for Row or Column ÷ Grand Total**

**"AND" Intersecting Proportion = Amount where row and column intersect ÷ Grand Total**

**"OR" Union Proportion = Add up all of the values in the row or column without using totals ÷ Grand Total**

**"OR" Union Proportion = (Row Total + Column Total – Intersection amount) ÷ Grand Total**

**"OR" Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting "AND" Proportion**



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

## Practice Problems Section 3B

### Formulas

Single Variable Marginal Proportion = Total for Row or Column  $\div$  Grand Total

“AND” Intersecting Proportion = Amount where row and column intersect  $\div$  Grand Total

“OR” Union Proportion = Add up all of the values in the row or column without using totals  $\div$  Grand Total

“OR” Union Proportion = (Row Total + Column Total – Intersection amount)  $\div$  Grand Total

“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion

To convert a proportion into a percentage, multiply by 100%.

Directions #1-12: The following contingency 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. Show your work.



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459

### Basic Marginal Proportions

1.
  - a) How many students have at least one tattoo?
  - b) What proportion of the students have a tattoo?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students have a tattoo.  
(Show how you calculated the answer. Round your percent to the tenths place.)
  
2.
  - a) How many students prefer Facebook?
  - b) What proportion of the students prefer Facebook?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students prefer Facebook?  
(Show how you calculated the answer. Round your percent to the tenths place.)
  
3.
  - a) How many students do not have a tattoo?
  - b) What proportion of the students do not have a tattoo?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students do not have a tattoo?  
(Show how you calculated the answer. Round your percent to the tenths place.)
  
4.
  - a) How many students prefer Instagram?
  - b) What proportion of the students prefer Instagram?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students prefer Instagram?  
(Show how you calculated the answer. Round your percent to the tenths place.)



### Joint Proportions “AND”

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459

5. a) How many students both have a tattoo and prefer Facebook?  
b) What proportion of the students both have a tattoo and prefer Facebook?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)  
c) What percentage of the students both have a tattoo and prefer Facebook?  
(Show how you calculated the answer. Round your percent to the tenths place.)
6. a) How many students both do not have a tattoo and prefer Instagram?  
b) What proportion of the students both do not have a tattoo and prefer Instagram?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)  
c) What percentage of the students both do not have a tattoo and prefer Instagram?  
(Show how you calculated the answer. Round your percent to the tenths place.)
7. a) How many students both do not have a tattoo and prefer Snapchat?  
b) What proportion of the students both do not have a tattoo and prefer Snapchat?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)  
c) What percentage of the students both do not have a tattoo and prefer Snapchat?  
(Show how you calculated the answer. Round your percent to the tenths place.)
8. a) How many students both have a tattoo and prefer “Other” social media?  
b) What proportion of the students both have a tattoo and prefer “Other” social media?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)  
c) What percentage of the students both have a tattoo and prefer “Other” social media?  
(Show how you calculated the answer. Round your percent to the tenths place.)

### Joint Proportions “OR”

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459



9. a) How many total students either have a tattoo or prefer Facebook?  
(Show how you calculated the answer.)
- b) What proportion of the students either have a tattoo or prefer Facebook?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students either have a tattoo or prefer Facebook?  
(Show how you calculated the answer. Round your percent to the tenths place.)
10. a) How many students either do not have a tattoo or prefer Instagram?  
(Show how you calculated the answer.)
- b) What proportion of the students either do not have a tattoo or prefer Instagram?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students either do not have a tattoo or prefer Instagram?  
(Show how you calculated the answer. Round your percent to the tenths place.)
11. a) How many students prefer either Twitter or Snapchat?  
(Show how you calculated the answer.)
- b) What proportion of the students prefer either Twitter or Snapchat?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students prefer either Twitter or Snapchat?  
(Show how you calculated the answer. Round your percent to the tenths place.)
12. a) How many students either have a tattoo or prefer "Other" social media?  
(Show how you calculated the answer.)
- b) What proportion of the students either have a tattoo or prefer "Other" social media?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students either have a tattoo or prefer "Other" social media?  
(Show how you calculated the answer. Round your percent to the tenths place.)

13. Copy and paste the gender and month data taken columns from the "Bear" data into StatKey. Use StatKey to calculate the following. Let gender represent the rows and the month data taken represent the columns.

Directions for creating contingency table with StatKey from Raw Data:

- Open the "Math 075 Survey Data Fall 2015". Copy and paste the two columns next to each other in a new spreadsheet. Then copy both columns together.
- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Click the "Edit Data" button. Push "Control A" and "Delete" on your keyboard to delete out any existing data. Then paste in your two columns of data. Check the box that says "Raw Data". If your data has a title, check the box that says "Data has a header row". Then push "OK".
- Click on the "Overall" proportions button and use the proportions provided to answer the questions.



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*



- a) What proportion of the bears had data taken in September? Convert the proportion into a percentage.
- b) What proportion of the bears were female? Convert the proportion into a percentage.
- c) What proportion of the bears were both female and had data taken in September? Convert the proportion into a percentage.
- d) What proportion of the bears were either female or had data taken in September? Use the following formula and your answers from parts (a), (b) and (c). Convert the proportion into a percentage.

**“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion**

14. Type in the following contingency table into StatKey and use the “Overall Proportions” button in StatKey to calculate the following proportions.

Directions for putting a contingency table into StatKey:

- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”.
- Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then type in the contingency table with commas as seen below. Do NOT check the box that says “Raw Data”. Check the box that says “Data has a header row”. Then push “OK”.
- Click on the “Overall” proportions button and use the proportions provided to answer the questions.

Contingency Table (Credit Card by Server)

[Blank], Cash, Credit Card

Server A, 39, 21

Server B, 50, 15

Server C, 17, 15

- a) What proportion of the bills were paid with cash? Convert the proportion into a percentage.
- b) What proportion of the bills had server B as the server? Convert the proportion into a percentage.
- c) What proportion of the bills were both served by server B and paid in cash? Convert the proportion into a percentage.
- d) What proportion of the bills were either served by server B or paid in cash? Use the following formula and your answers from parts (a), (b) and (c). Convert the proportion into a percentage.

**“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion**

---

### Section 3C – Conditional Percentages from Contingency Tables and Categorical Relationships

Conditional Proportions and Percentages



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

Conditional proportions and percentages are the key to understanding categorical relationships. A condition is thought of as prior knowledge about the person or situation that may change the percentage. Let us say that the Los Angeles Lakers have a 75% chance of beating the Phoenix Suns. If the Lakers best player LeBron James does not play, will that change the percentage? Of course. Knowing that LeBron James will not play is called a condition.

In contingency tables, a condition involves restricting to one particular group before you calculate the percentage.

Example:

What percentage of the Canyon Country campus Math 075 pre-stat students prefer Twitter as their favorite social media?

First notice that this is not a joint proportion. It does NOT ask for the percentage of all students both prefer Twitter and go to the Canyon Country campus.

The key is to identify which group we are restricting ourselves to. In other words, what is the condition? Look for words that say “if” or “given this is true” or “out of”. This designates the condition. In this example, notice that the problem said “of the Canyon Country students”. That means that we are supposed to only look at the Canyon Country students when we find our amount (frequency) and total. A commonly used method for calculating conditional percentages from a contingency table is to circle the row or column that has your condition (Canyon Country). Then only use numbers in that row or column.

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

Notice that the Canyon Country Campus counts are in the first row. So we should highlight or circle the first row and only use numbers in the first row when we calculate. We should not use the grand total anymore. We need the total number of students that attend the Canyon Country campus. In other words, the total from our condition. The amount will be the number of students that prefer Twitter in the Canyon Country row. In other words the intersection cell frequency.

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

$$\text{Conditional Proportion} = \frac{\text{Amount in Intersection Cell (Canyon Country meets Twitter)}}{\text{Row or Column Total (Row total Canyon Country)}} = \frac{35}{180} \approx 0.19444444 \approx 0.194$$

$$\text{Conditional Percentage} = \text{Conditional Proportion} \times 100\% = 0.194 \times 100\% = 19.4\%$$

So 19.4% of the Canyon Country pre-stat students prefer Twitter as their favorite social media.

We can also have StatKey calculate conditional proportions for us by using the “row” and “column” proportion



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

buttons. We need to ask ourselves if the condition is a row or a column? In the last question we were restricting ourselves to only Canyon Country students. This is a row. Since the condition is a row, we should click the “row” proportion button. If the condition had been a column, we would have clicked on the “Column” proportion button.

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

**Proportions** Row Column Overall

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	0.194	0.289	0.239	0.2	0.078	1
Valencia Campus	0.201	0.305	0.201	0.208	0.086	1
Total	0.198	0.298	0.216	0.205	0.083	1

Notice that all of the rows add up to 1 (100%). This confirms that the computer is calculating the conditional proportions for the rows. We are looking for the proportion of Canyon Country pre-stat students that prefer Twitter. Notice the answer we are looking for is given in the intersecting cell. If we restrict ourselves to considering only the Canyon Country students, 0.194 or 19.4% of them prefer twitter. This is the same answer we got earlier in the section.

Example:

What proportion of the Snapchat math 075 pre-stat students attend the Valencia campus?

To answer this we need to recognize that we are no longer considering all the students. We are restricting our proportion to considering only the Snapchat students (“out of”). Notice that the student that prefer Snapchat are in a column. Since the condition is preferring Snapchat, we should only use numbers in the Snapchat column to calculate the proportion. Notice we highlighted the numbers in Snapchat column. The total will now be the total number of Snapchat students and the amount will be the amount of Snapchat students that attend the Valencia campus.

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

$$\text{Conditional Proportion} = \frac{\text{Amount in Intersection Cell (Snapchat meets Valencia)}}{\text{Row or Column Total (column total Snapchat)}} = \frac{58}{94} \approx 0.617021276 \approx 0.617$$

$$\text{Conditional Percentage} = \text{Conditional Proportion} \times 100\% = 0.617 \times 100\% = 61.7\%$$



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

We can also use StatKey to find the proportion of the Snapchat math 075 pre-stat students attend the Valencia campus. Notice our condition is now Snapchat (“out of”). This is a column so I will click the “column” proportion button in StatKey.

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

**Proportions** Row Column Overall

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	0.385	0.38	0.434	0.383	0.368	0.392
Valencia Campus	0.615	0.62	0.566	0.617	0.632	0.608
Total	1	1	1	1	1	1

Notice that when we click the “Column” proportion button, all of the columns add up to 1 (100%). This lets us know that StatKey has calculated all of the conditional proportions for the columns.

The conditional proportion we are looking for is where Snapchat and Valencia intersect. 0.617 or 61.7%. Notice this is the same answer as our earlier calculation.

Note: Categorical data is often given to a data scientist as a contingency table with summary counts. Most data scientists do not calculate things by hand. Recall that in section 3A, we learned we can type in an existing contingency table into StatKey using commas. Typing the table into StatKey allows us to not only have access to the stacked bar chart, but also the proportion button that can calculate proportions automatically for us.

Relationship Principle

Let us go back to the LeBron James example. The key to understanding categorical relationships is to judge how close or far apart conditional percentages are.

- Chances of Los Angeles Lakers beating the Phoenix Suns if LeBron James plays  $\approx 75\%$
- Chances of Los Angeles Lakers beating the Phoenix Suns if LeBron James does not play  $\approx 25\%$

These percentages are significantly different, so it tells us that the condition of LeBron James playing in the game is related to the Lakers winning.

Note: Does this mean that LeBron playing is the only factor that CAUSES the Lakers to win? No. Remember related (associated) does NOT prove cause and effect. There are many confounding variables that go into the Lakers winning or losing. (Health of LeBron, Health of other players, the team the Lakers are playing, home game or away game, Number of games played, etc...) We can say that LeBron James playing is related to the Lakers winning, but the data does not prove that LeBron James playing is the only factor that causes the Lakers to win.



Let us look at another example using the Lakers chances of beating the Phoenix Suns.

Chances of Lakers winning if it snows in Nebraska  $\approx 75\%$

Chances of Lakers winning if it does not snow in Nebraska  $\approx 75\%$

These percentages are not significantly different, so it tells us that the condition of snowing in Nebraska is not related to the Lakers winning. The condition does not matter.

**Relationship Principle:**

**Close Conditional Percentages from same variable = Condition is NOT related to the categorical variable**

**Significantly Different Conditional Percentages from same variable = Condition IS related to the categorical variable**

*Note: You cannot compare any conditional percentages you want. They must be the same variable for the percentage and from different groups (different condition). You cannot compare the percentage of Snapchat students from the Canyon Country campus to the percentage of Twitter from the Valencia campus. They are not the same thing and will likely have very different percentages regardless of the relationship. Compare the percentage of Snapchat students from the Canyon Country campus to the percentage of Snapchat students from the Valencia campus. That will give us information about the relationship. Conditional percentage analysis is the basis behind the Chi-Square test statistics in more advanced statistics classes.*

Example:

Look at the following conditional proportions StatKey calculated based on the rows (Canyon Country campus and Valencia campus).

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

**Proportions** Row Column Overall

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	0.194	0.289	0.239	0.2	0.078	1
Valencia Campus	0.201	0.305	0.201	0.208	0.086	1
Total	0.198	0.298	0.216	0.205	0.083	1

We can only compare Twitter to Twitter, Instagram to Instagram, Facebook to Facebook, Snapchat to Snapchat, Other to Other. Notice the proportions look very close. This gives us the idea that a pre-stat students favorite social media may not be related to the campus they go to. If the proportions were significantly different, that may indicate a relationship.



## Practice Problems Section 3C

### Formulas

Conditional Proportion = Intersection of the row and column  $\div$  Row or column total for Condition

Circle the row or column that has the condition. Use only numbers in that row or column when calculating the conditional proportion.

To convert a proportion into a percentage, multiply by 100%.

Directions #1-4: The following contingency 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. Show your work.

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459

- How many total students have at least one tattoo?
  - How many students both have a tattoo and prefer Instagram?
  - What proportion of the tattoo students prefer Instagram? (Show how you calculated the answer.)
  - What percentage of the tattoo students prefer Instagram? (Show how you calculated the answer.)
- How many total students prefer Twitter?
  - How many students both do not have a tattoo and prefer Twitter?
  - What proportion of the Twitter students do not have a tattoo? (Show how you calculated the answer.)
  - What percentage of the Twitter students do not have a tattoo? (Show how you calculated the answer.)
- How many total students do not have a tattoo?
  - How many students both do not have a tattoo and prefer Facebook?
  - What proportion of the no tattoo students prefer Facebook? (Show how you calculated the answer.)
  - What percentage of the no tattoo students prefer Facebook? (Show how you calculated the answer.)
- How many total students prefer Snapchat?
  - How many students both have a tattoo and prefer Snapchat?



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

- c) What proportion of the Snapchat students have a tattoo? (Show how you calculated the answer.)
- d) What percentage of the Snapchat students have a tattoo? (Show how you calculated the answer.)

Directions #5-8: The following contingency table was created from the Math 075 Survey Data Fall 2015 and describes the campus the student attended and the type of transportation they took to get to school. Use the table to find the given proportions and percentages. Show your work.

Campus \ Transportation type to campus	Drive alone	Public transportation	Dropped off by someone	Carpool	Walk	Other	Bicycle	Skate	Total
Canyon Country Campus	138	7	14	15	1	4	1	0	180
Valencia Campus	203	17	32	22	3	0	1	1	279
Total	341	24	46	37	4	4	2	1	459

5. a) How many total students went to the Canyon Country campus?  
 b) How many students both drive alone and went to the Canyon Country campus?  
 c) What proportion of the Canyon Country campus students drove alone to school? (Show how you calculated the answer.)  
 d) What percentage of the Canyon Country campus students drove alone to school? (Show how you calculated the answer.)
6. a) How many total students were dropped off by someone?  
 b) How many students were both dropped off and went to the Canyon Country campus?  
 c) What proportion of the dropped off students went to the Canyon Country campus? (Show how you calculated the answer.)  
 d) What percentage of the dropped off students went to the Canyon Country campus? (Show how you calculated the answer.)
7. a) How many total students went to the Valencia campus?  
 b) How many students both carpool and went to the Valencia campus?  
 c) What proportion of the Valencia campus students carpool to school? (Show how you calculated the answer.)  
 d) What percentage of the Valencia campus students carpool to school? (Show how you calculated the answer.)
8. a) How many total students used public transportation to school?  
 b) How many students both used public transportation and went to the Valencia campus?



- c) What proportion of the public transportation students went to the Valencia campus?  
(Show how you calculated the answer.)
- d) What percentage of the public transportation students went to the Valencia campus?  
(Show how you calculated the answer.)

9. Copy and paste the gender and month data taken columns from the “Bear” data into StatKey. Use StatKey to calculate the following. Let bear gender represent the rows and month data was taken represent the columns.

Directions for creating contingency table with StatKey from Raw Data:

- Open the “Math 075 Survey Data Fall 2015”. Copy and paste the two columns next to each other in a new spreadsheet. Then copy both columns together.
- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”. Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then paste in your two columns of data. Check the box that says “Raw Data”. If your data has a title, check the box that says “Data has a header row”. Then push “OK”. The rows should be gender and the columns should be month data taken. If it is not, simply click the “Switch Variables” button.
- Click on the “Row” proportions button and use the conditional row proportions to answer the questions.

- a) What proportion of the female bears were measured in August?
- b) What proportion of the male bears were measured in August?
- c) Compare your answer in letter (a) to your answer in letter (b). Do the proportions look close or significantly different?
- d) What proportion of the female bears were measured in October?
- e) What proportion of the male bears were measured in October?
- f) Compare your answer in letter (d) to your answer in letter (e). Do the proportions look close or significantly different?
- g) Do your answers in letters (c) and (f) indicate that the bear gender may be related to what month the bears are measured in? Explain your answer.
- h) If data indicated that bear gender was related to the month the bears were measured in, would that prove that the gender of the bear causes the bear to be measured in a certain month? Explain your answer.

10. Type in the following contingency table into StatKey and use the “Overall Proportions” button in StatKey to calculate the following proportions.

Directions for putting a contingency table into StatKey:

- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”.
- Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then type in the contingency table with commas as seen below. Do NOT check the box that says “Raw Data”. Check the box that says “Data has a header row”. Then push “OK”. The rows should be the servers and the columns should be the type of payment. If they are not, simply push the “switch variables” button.
- Click on the “Column” proportions button and use the conditional column proportions provided to answer the questions.





Contingency Table (Credit Card by Server)

[Blank], Cash, Credit Card

Server A, 39, 21

Server B, 50, 15

Server C, 17, 15

- a) What proportion of the credit card customers were served by server A?
  - b) What proportion of the credit card customers were served by server B?
  - c) What proportion of the credit card customers were served by server C?
  - d) Do your answers in parts (a), (b) and (c) seem close or significantly different?
  - e) Does your answer in part (d) indicate that paying with a credit card may be related to who the server is?  
Explain your answer.
- 



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

### Chapter 3 Review Sheet

Here is a list of important ideas in this chapter.

- Be comfortable creating and analyzing contingency 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 3 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

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}}$$

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

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

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}}$$

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



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?
- 



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

## Project Chapter 3 – Categorical Relationships

**Directions for Online Classes:** *This will be an individual project. Each student will chose two columns of categorical data to analyze from the “Math 075 Survey data Fall 2015” and create a poster summarizing their findings. Students can chose two from the following columns of data: Tattoo, Texting While Driving, Favorite Social Media, Transportation to School, Car Accident, Cigarettes, Eat Breakfast, Glasses/Contacts, High School in Santa Clarita, Living with parents*

*After submitting the project to their instructor, students will then go to the “Chapter 3 Project Class Discussion” in Canvas and discuss their findings with other students in the class.*

### The Individual Poster Should Have

- The poster does not have to be extremely large.
- Your first and last name on the poster
- What two columns of data did you pick?
- Explain why this data is important or interesting to you?
- Copy and paste the two columns of data next to each other in a new spreadsheet. Then copy both columns together. Go to StatKey at [www.lock5stat.com](http://www.lock5stat.com) and click on “Two Categorical Variables” under the “Descriptive Statistics and Graphs” menu. Then paste the two columns of data together under “edit data”. Remember to check the box that says “raw data” before pushing “OK”. Also check “header row” if data has a title.
- Copy the “Counts Table” table from StatKey onto your poster in large letters. (Label this table as your “Contingency Table”.) Pick out a few counts on this table and explain them.
- Draw the stacked bar chart onto your poster.
- Click the “Overall” button where it says Proportions in StatKey. Copy the “Overall Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Overall Proportions Table”.) Pick out a proportion under totals. Explain what variable the computer is finding the marginal proportion of and explain how the computer calculated it. Pick out one proportion in the middle of the table. Explain what two variables the computer is finding the “AND” joint proportion for and explain how the computer calculated it.
- Click the “Row” button where it says Proportions in StatKey. Copy the “Row Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Conditional Row Proportions Table”.) In the “Row proportion table”, compare proportions that are in the same column. Do they look close or significantly different? What does this indicate about whether or not the two columns of data you chose are related or not?
- Decorate Poster

Now take a picture of your poster project and submit the picture to your instructor in Canvas.

After submitting the picture of the poster, go to the discussion menu in Canvas and complete the “Chapter 3 Project Discussion”. You will be discussing your findings with other students in the class.

---



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

**Directions for Face to Face Classes:** The class will be broken up into groups of three or four. Each group will pick a team name and one of the following pairs of 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

#### The Group Poster Should Have

- The poster does not have to be extremely large.
- Your first and last name of everyone in your group should be on the poster.
- Which two columns of categorical data did you chose?
- Explain why this data is important or interesting to your group?
- Copy and paste the two columns of data next to each other in a new spreadsheet. Then copy both columns together. Go to StatKey at [www.lock5stat.com](http://www.lock5stat.com) and click on “Two Categorical Variables” under the “Descriptive Statistics and Graphs” menu. Then paste the two columns of data together under “edit data”. Remember to check the box that says “raw data” before pushing “OK”. Also check “header row” if data has a title.
- Copy the “Counts Table” table from StatKey onto your poster in large letters. (Label this table as your “Contingency Table”.) Pick out a few counts on this table and explain them.
- Draw the stacked bar chart onto your poster.
- Click the “Overall” button where it says Proportions in StatKey. Copy the “Overall Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Overall Proportions Table”.) Pick out a proportion under totals. Explain what variable the computer is finding the marginal proportion of and explain how the computer calculated it. Pick out one proportion in the middle of the table. Explain what two variables the computer is finding the “AND” joint proportion for and explain how the computer calculated it.
- Click the “Row” button where it says Proportions in StatKey. Copy the “Row Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Conditional Row Proportions Table”.) In the “Row proportion table”, compare proportions that are in the same column. Do they look close or significantly different? What does this indicate about whether or not the two columns of data you chose are related or not?
- Decorate 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. Then rotate and have another person from the group present. Keep rotating till each person in every group has presented. Each presentation should take a few minutes. Make sure audience rotates to new posters as well.



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*