Section 4E – Categorical Relationships: Contingency Tables

Vocabulary

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Contingency Table: Also called a two-way table. This table summarizes the counts when comparing two different categorical data sets each with two or more variables.

Marginal Percentage (Marginal Proportion): A single percentage or proportion without any conditions. In a contingency table, this can found with numbers in the margins.

Conditional Percentage (Conditional Proportion): The percentage or proportion calculated from a particular group or if a particular condition was true. These are the very important when studying categorical relationships.

Joint Percentage (Joint Proportion): A percentage or proportion involving two variables being true about the person or object, but does not have a condition. There are generally two types (AND, OR).

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 contingency (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" or "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. Remember that just because there is a relationship, that does not prove causation. There may be many other factors involved.*

To analyze categorical data we need to know the counts (frequencies) for each categorical variable. This particularly important when you are studying categorical relationships. No data scientist or statistician finds the frequencies by hand. They use computer programs to make a contingency table (or two-way table).

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Creating a contingency table with raw data and StatKey

Let us look at an example. Go to www.matt-teachout.org and click on the math 140 survey data fall 2015. We want to explore the relationship between the campus a person goes to and their political party.

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. 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 fresh excel spreadsheet. 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 Excel sheet. Otherwise, StatKey will not accept it. 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".

To make a contingency table with StatKey, go to www.lock5stat.com and click the "StatKey" button. Now click on "Two Categorical Variables" under the "Descriptive Statistics and Graphs" menu. Then click on the "edit data" button. Copy both columns together in your excel spreadsheet and paste them into StatKey. Check the "raw data" box and the "data has header row" box and push "OK".

Counts Table Switch Variables

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions Row | Column | Overall

This is called a contingency table. Notice that we a lot of frequency information. Notice 66 is in the Democratic column and Valencia row, so 66 Math 140 students are both Democrat and go to the Valencia campus. Similarly, 18 math 140 students are both Republican and go to the Canyon Country campus.

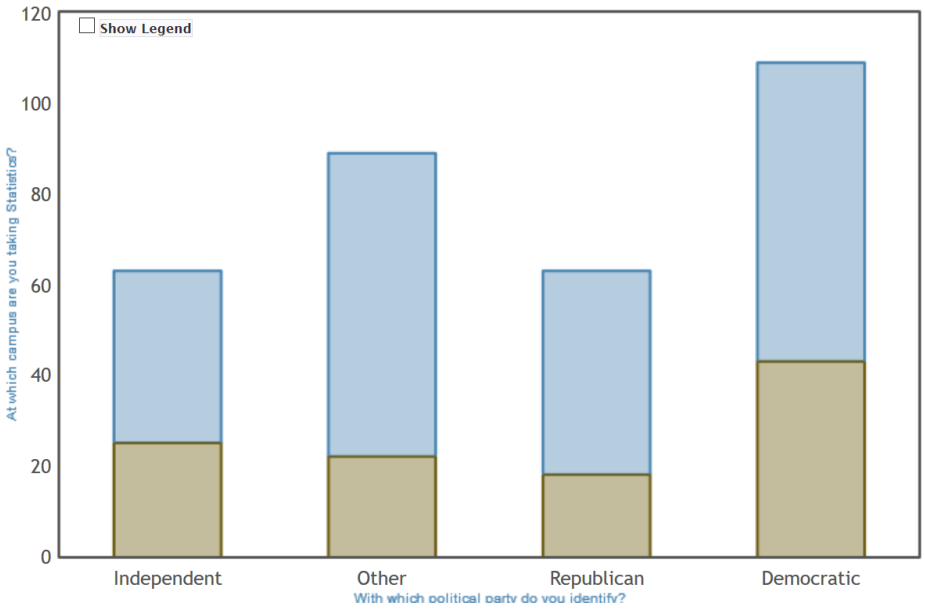
The size of a contingency table is the number of rows by the number of columns. Totals are not included. This table has two rows (CCC and Valencia) and four columns (Independent, Other, Republican, and Democratic), so this is a "2 by 4" or "2×4" contingency table.

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

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

StatKey **Descriptive Statistics for Two Categorical Variables**

Custom Dataset | Show Data Table | Edit Data | Upload File | Change Column(s)



The "proportion" buttons are particularly useful. If we click on the "overall" proportion button. The computer calculates the intersection (AND) percentages for the entire data set. If we click on the "row" proportion button it gives conditional percentages for the rows. If we click on the "column" proportion button it gives the conditional percentages for the columns. We will discuss these more later, but these are very useful.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.231	0.204	0.167	0.398	1
Valencia Campus	0.176	0.31	0.208	0.306	1
Total	0.194	0.275	0.194	0.336	1

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.397	0.247	0.286	0.394	0.333
Valencia Campus	0.603	0.753	0.714	0.606	0.667
Total	1	1	1	1	1

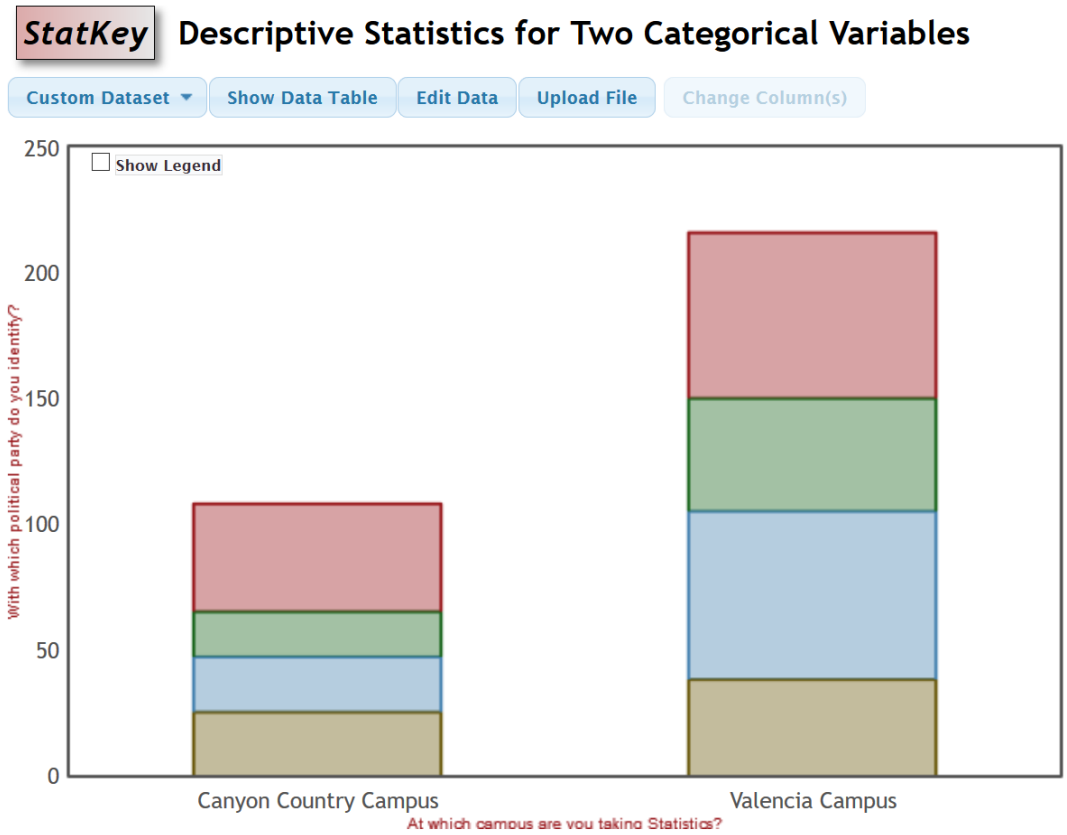
Another feature is the "switch variables" button. Clicking on this button will switch the rows and columns.

Counts Table Switch Variables

With which political party do you identify? \ At which campus are you taking Statistics?	Canyon Country Campus	Valencia Campus	Total
Independent	25	38	63
Other	22	67	89
Republican	18	45	63
Democratic	43	66	109
Total	108	216	324

Proportions Row | Column | Overall

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18



Notice that you can click on any section in the graph and it will highlight the count it came from in the contingency table. In addition, you can click on the proportion buttons to calculate and compare various proportions.

Creating a contingency table with summary counts and StatKey

Let us look at the same example again. As we said in the last section, categorical data is often not given in raw form. Sometimes a person may give you the summary counts (frequencies). In that case, you already have the contingency table, yet it is good to be able to put that into StatKey to create the stacked bar chart and use the switch variable and proportion features. To put in a contingency table into StatKey, go to www.lock5stat.com and click the "StatKey" button. Now click on "Two Categorical Variables" under the "Descriptive Statistics and Graphs" menu. Then click on the "edit data" button. Type in the table as seen below. Note that there should be a space after every comma and the totals are not included. There should also be a "[blank]" in the upper left corner. Uncheck the "raw data" box and check the "data has header row" box and push "OK". Notice this gives us the exact same table and graphs as if we had used the raw data.

[blank], Independent, Other, Republican, Democratic
Canyon Country Campus, 25, 22, 18, 43
Valencia Campus, 38, 67, 45, 66

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Creating a contingency table with raw data and Statcato

You can also create a contingency table with Statcato. Copy and paste the ordered pair categorical data into a fresh excel spreadsheet. Make sure to clean the data and delete out any rows with missing values. Since this data set is over 300 values, go to the "edit" menu, "add multiple rows" and add another 100 rows. When that is done, copy and paste the two columns one at a time into Statcato. Statcato does not copy and paste multiple columns at the same time very well. It is best to copy and paste one at a time. Now 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

If we use the campus and political party data from the previous example, we get the following from Statcato. Notice it gives the counts (frequencies), totals (All), and the intersection percentages (AND).

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Calculating Marginal Percentages

Marginal Percentages are percentages that involve only one variable and do not have a condition. They get their name because the amount and total are found in the margins (totals). Let us look at a couple examples. Remember a proportion and percentage can be found from the amount (frequency) and the total.

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

Percentage = $\frac{Amount\ (Frequency)}{Total} \times 100\%$

Example: Find the proportion and percentage of the math 140 students are democrat. Notice we need to find the amount of democrats and the total number of students. The amount of democrats will be in total part of the democrat row or column. The total number of students is often called the grand total and is found in the bottom right of the table.

Proportion = $\frac{Amount\ (Frequency)}{Total} = \frac{109}{324} \approx 0.336$

Percentage = proportion $\times$ 100% = 0.336 $\times$ 100% = 33.6%

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

It is always better to use technology when you can instead of calculating something by hand. We could had found the proportion with StatKey by clicking on the "overall" proportion button. Statcato had this percentage already calculated as well. Notice the democrat data is summarized as a column. In both StatKey and Statcato, we need to look at the total in the democratic column to get the proportion. We can then convert the answer into a percentage or proportion as needed.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Example: Use the tables above to give the proportion and percentage of the Math 140 students that attended the Canyon Country campus. Look for the Canyon Country campus data. Notice it is in the first row. So the number we are looking for is at the end of the first row under "total" or "All".

Proportion of Math 140 students at the Canyon Country campus ≈ 0.333
Percentage of Math 140 students at the Canyon Country campus ≈ 33.3%

Calculating Joint Percentages

There are two types of joint percentages. The first type is the percentage of the total that has two things true about the person. We often call this the intersecting percentage or "AND". The second type is the proportion or percentage of the total that has either one of two things true about the person. This is sometimes called the union percentage or "OR". Intersecting percentages means that both things must be true about the person or object. Let us look at a few examples. Remember a proportion and percentage can be found from the amount (frequency) and the total.

Example: Find the proportion and percentage of the math 140 students that are both democrat AND attend the Valencia campus. Both things must be true about the person. In an "AND" (intersection) proportion, the amount can be found in the cell where the column and row meet. We will still use the "grand total" in the lower right corner as the total, since we need to include everyone in the data set. Look at the where the democratic column meets the Valencia row. There are 66 students that have both characteristics. This is the amount we need. The grand total is still 324 so here is the proportion and percentage calculation. Round your answer to three significant figures.

"AND" Proportion = $\frac{Frequency\ in\ intersection\ cell}{Grand\ Total} = \frac{66}{324} \approx 0.2037037 \approx 0.204$

"AND" Percentage = proportion × 100% = 0.204 × 100% = 20.4%

Again, we could have used technology to get that answer. We could had found the proportion with StatKey by clicking on the "overall" proportion button. Statcato had this percentage already calculated as well. Both times, we need to look at the cell where the Democratic column meets the Valencia row.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Example: Use the tables above to give the proportion and percentage of the Math 140 students that both attend the Canyon Country campus AND are Republican. Look for where the Canyon Country campus row meets the Republican Column.

Proportion of Math 140 students at the Canyon Country campus AND Republican ≈ 0.056
Percentage of Math 140 students at the Canyon Country campus AND Republican ≈ 5.6%

Example: Now calculate the proportion and percentage of Math 140 students that either are at the Valencia campus OR are Democratic. This means we need to include anyone that was Democrat regardless of campus and include anyone at the Valencia campus regardless of the political affiliation. This is a more difficult calculation. Here is a couple common formulas for "OR" (union) percentages.

"OR" (Union) Proportion = $\frac{(Row\ Total + Column\ Total - Intersection\ Cell)}{Grand\ Total} = \frac{(216 + 109 - 66)}{324} = \frac{(259)}{324} \approx 0.79938 \approx 0.799$

It is better to use technology if we can. StatKey and Statcato printouts can help us calculate the "OR" (union) proportion or percentage.

"OR" (Union) Proportion = Row Total Proportion + Column Total Proportion − Intersection Cell Proportion

Notice these proportions are given in the StatKey table we can use them to calculate the "OR" proportion.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

"OR" (Union) Proportion = Row Total Proportion (Valencia) + Column Total Proportion (Democratic) − Intersection Cell Proportion (where Valencia and Democratic meet)

= 0.667 + 0.336 – 0.204 = 0.799

(We can convert this proportion to a percentage if needed. Percent = Proportion × 100% ≈ 0.799 × 100% ≈ 77.9%)

"OR" (Union) Percentage = Row Total % + Column Total % − Intersection Cell %

Notice these percentages are given in the Statcato table we can use them to calculate the "OR" percentage.

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

"OR" (Union) Percentage = Row Total % (Valencia) + Column Total % (Democratic) − Intersection Cell % (where Valencia and Democratic meet) = 66.7% + 33.6% − 20.4% = 79.9%

Conditional Proportions and Percentages

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 140 students are Democrat?

First notice that this is not a joint proportion. It does NOT ask for the percentage of all students that are both Democrat 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.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Counts Table Switch Variables

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions Row Column Overall

Notice that the Canyon Country Campus counts are in the first row. So we should only use numbers in the first row. 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 democrats in the Canyon Country row. In other words the intersection cell frequency.

Conditional Proportion = $\frac{Amount\ in\ Intersection\ Cell\ (Canyon\ Country\ meets\ Democratic)}{Row\ or\ Column\ Total\ (Row\ total\ Canyon\ Country)} = \frac{43}{108} \approx 0.398148 \approx 0.398$

We can use the "row" and "column" proportion buttons in StatKey to find this conditional proportion. Since the condition is a row, we should click the "row" proportion button.

Proportions Row Column Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.231	0.204	0.167	0.398	1
Valencia Campus	0.176	0.31	0.208	0.306	1
Total	0.194	0.275	0.194	0.336	1

Notice the answer we are looking for is given in the intersecting cell. If we restrict ourselves to considering only the Canyon Country students, 0.398 or 39.8% of them are democrat.

Example: What proportion of the republican math 140 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 republican students ("out of"). Since the condition is being republican, we should only use numbers in the republican column. The total will now be the total number of republicans and the amount will be the amount of republicans that attend the Valencia campus.

Counts Table Switch Variables

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions Row Column Overall

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Conditional Proportion = $\frac{Amount\ in\ Intersection\ Cell\ (Republican\ meets\ Valencia)}{Row\ or\ Column\ Total\ (column\ total\ Republican)} = \frac{45}{63} \approx 0.7142857 \approx 0.714$

We can also use StatKey to find what proportion of Republican Math 140 students attend the Valencia campus. Notice our condition is now republican ("out of"). This is a column so I will click the "column" proportion button in StatKey.

Proportions Row | Column | Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.397	0.247	0.286	0.394	0.333
Valencia Campus	0.603	0.753	0.714	0.606	0.667
Total	1	1	1	1	1

Notice now we want to restrict ourselves to the Republican column. The conditional proportion we are looking for is 0.714 or 71.4%.

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 Lakers winning if LeBron James plays ≈ 75%
Chances of Lakers winning if LeBron James does not play ≈ 40%

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.

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

Chances of Lakers winning if it snows in Nebraska ≈ 75%
Chances of Lakers winning if it does not snow in Nebraska ≈ 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 = Condition is NOT related to the categorical variable

Significantly Different Conditional Percentages = 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 republicans from the Canyon Country campus to the percentage of democrats 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 republicans from the Canyon Country campus to the percentage of republicans from the Valencia campus. That will give us information about the relationship. Conditional percentage analysis is the basis behind the Chi-Squared test statistic we will learn in chapter 5.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18