Chapter 1 – Categorical Data Analysis

Introduction: Statisticians, data scientists and data analysts analyze data all the time. Often, they analyze categorical data by looking at amounts, totals, percentages and decimal proportions.

Note about Terminology:

Percentages are a vital link to understanding categorical data. Most students think of percentages as a calculation of probability, like the probability of drawing an ace from a deck of cards. In statistics, we want to know the proportion of people or objects that have a certain characteristic in a data set. I find that if I ask my class to calculate a probability, they seem to understand the idea, but if I ask what is the proportion of people that want to purchase a particular car, they do not understand. Most students think of solving an equation when they hear the term "proportion". In statistics, a proportion is an amount divided by the total or a percentage divided by 100. Do not think of it as an equation you need to solve.

Though you can think of percentages and proportions as calculating a probability, we will focus on the more common statistics terminology of "proportion". Also, remember that though decimal proportions and percentages are equivalent, they are not the same thing. If a computer program asks for the sample proportion, it will say "error" if you put the percentage.

Decimal proportion = amount / total (or a percentage divided by 100)

Percentage = decimal proportion x 100%

Section 1A – Two Types of Data – Categorical and Quantitative

One of the most important factors when analyzing data is to determine what type of data you have and how many variables you are analyzing. Let us start with the type of data.

There are two general types of data, **categorical** and **quantitative**.

Categorical Data

Categorical data are generally labels that tell us something about the people or objects in the data set. For example, what country do they live in, what is the person's occupation, or what kind of pet they have?

Usually categorical data is made up of words (do you smoke - yes or no), but occasionally a number can be used as a category. For example, a zip code can be used instead of the place a person lives. The numbers "1" and "2" can be used instead of female and male.

Quantitative Data

Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. For example: a list of people's heights in inches, or their weights in kilograms, or a list of how many dogs are there in various animal shelters across Los Angeles. Notice in each of these cases the data is numerical and an average seems appropriate in the context. We can find the average height, the average weight, or the average number of dogs in animal shelters in Los Angeles.

Numbers used as categories

Remember, not all numeric data is quantitative. Ask yourself if the numbers are measuring or counting something and if an average would make sense. For example, a list of peoples zip codes are numbers but an average zip code would not really tell us anything. In addition, identity numbers like hospital ID numbers, student ID numbers or social security numbers are not measuring anything and an average would not make sense in the context so they are not quantitative.

Problem Set Section 1A

1. Open the bear data and classify each column of data as categorical or quantitative? If the data is quantitative, what are the units? If the data is categorical, list the different labels (variables) in that category.

2. Open the cereal data and classify each column of data as categorical or quantitative? If the data is quantitative, what are the units? If the data is categorical, list the different labels (variables) in that category.

3. Open the math 075-survey data fall 2015 and classify each column of data as categorical or quantitative? If the data is quantitative, what are the units? If the data is categorical, list the different labels (variables) in that category.

Section 1B – Proportions and Percentages

To analyze categorical data, we focus on exploring various types of percentages and compare them. In statistics, the decimal equivalent to a percentage is often called a "proportion".

How to calculate a decimal proportion

To find a decimal proportion you will need to find the amount divided by the total.

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

Counting how many people share a certain characteristic or even a total number of cars in a data set can take a long time in a big data set, however technology can help. Statistics software can count much quicker and easily than we can. In this section, we will assume we know the amount and the total.

Suppose a health clinic has seen 326 people in the last month and 41 of them had the flu. If we were analyzing their data, the first thing we would like to do is find what proportion of the patients have the flu. It is not a difficult calculation and can be done with a small calculator.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687$$

Should we round the answer? Proportions and Percentages are usually rounded to the three significant figures. Proportions are usually rounded to the thousandths place (3rd place to the right of the decimal).

Let us review rounding. We want to round the above answer to the thousandths place, which is the "5". Always look at the number to the right of the place you are rounding to. If the number to the right is 5-9, round up (add 1 to the place value). If the number is 0-4, round down (leave the place value alone). After rounding cut off the rest of the decimals.

Therefore, in the previous answer we want to round to the thousandths place (5). The number to the right of the 5 is a 7. So should we round up or down? If you said round up, you are correct. Therefore, we will add 1 to the place value and the 5 becomes a 6. Now we cut off the rest of the decimal and our approximate answer is 0.126.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687 \approx 0.126$$

Decimal proportions are vital in the analysis of categorical data, but many people have trouble understanding the implications of a decimal proportion like 0.126. That is why we often convert the proportion into a percentage.

How to convert a decimal proportion into a percentage

To convert a decimal proportion into a percentage, multiply by 100 and put on the "%" symbol. Think of it like taking 100% of the decimal proportion. When you multiply by 100, the decimal moves two places to the right. Some people prefer to move the decimal, but I find students make fewer errors when they just multiply by 100 with their calculator.

Percentage = Decimal Proportion x 100%

Look at our previous example of the number of cases of the flu at a health clinic. We used the amount and total to calculate the decimal proportion.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687 \approx 0.126$$

So what percentage of the patients had the flu? All we need to do is multiply the decimal proportion 0.126 by 100% to get the percentage equivalent.

Percentage = Decimal Proportion x 100% = 0.126 x 100% = 12.6%

So 12.6% of the patients at the health clinic were seen for the flu. This can be alarming information to the health clinic if that is an unusually high percentage.

Notice that the percentage still has three significant figures, but is rounded to the tenths place (one place to the right of the decimal). Rounding to the tenth of a percent is a common place to round percentages in statistics.

If you want to calculate the percentage directly from the categorical data, here is another formula you may use.

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

Important Note

There are three ways to describe the proportion for categorical data: fraction, decimal, and percentage. Notice for the flu data example above, we have the three ways of describing the data: the fraction 41/326, the decimal proportion 0.126, and the percentage 12.6%. All of them are equivalent. It is important to be comfortable with fractions, decimal proportions and percentages when describing categorical data. They are a foundation for more advanced categorical analysis later on.

Problem Set Section 1B

Directions: For each of the following questions, calculate a percentage (proportion). Write your answers as a fraction, decimal and as a percentage.

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

To convert proportion into percentage, multiply by 100%.

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

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

1. What proportion of the patients that were male? (Write your answer as a fraction, decimal and percentage).

2. What percent of the patients were female? (Write your answer as a fraction, decimal and percentage).

3. What proportion of the patients had Type A blood? (Write your answer as a fraction, decimal and percentage).

4. What percent of the patients had Type B blood? (Write your answer as a fraction, decimal and percentage).

5. What proportion of the patients had Type AB blood? (Write your answer as a fraction, decimal and percentage).

6. What percent of the patients had Type O blood? (Write your answer as a fraction, decimal and percentage).

7. What proportion of the patients were RH negative? (Write your answer as a fraction, decimal and percentage).

8. What percent of the patients were RH positive? (Write your answer as a fraction, decimal and percentage).

9. What proportion of the patients went to the emergency room (ER)? (Write your answer as a fraction, decimal and percentage).

10. What percent of the patients went to the Intensive Care Unit (ICU)? (Write your answer as a fraction, decimal and percentage).

11. What proportion of the patients went to the medical / surgical floor (Med/Surg)? (Write your answer as a fraction, decimal and percentage).

12. What percent of the patients went to same day surgery (SDS)? (Write your answer as a fraction, decimal and percentage).

Sometimes quantitative data may be classified into categories and then proportions (percentages) obtained. For example, look at the ages of the patients. Age is of course quantitative data (numbers that measure something), but we could classify the patients into three age groups: 30 or under, 31 to 64, 65 and above.

13. What percent of the patients were 30 years old or less? (Write your answer as a fraction, decimal and percentage).

14. What proportion of the patients were 31 to 64 years old? (Write your answer as a fraction, decimal and percentage).

15. What percent of the patients were 65 years old or more? (Write your answer as a fraction, decimal and percentage).

Sometimes we may want to know a proportion regarding two things being true about the person. For example, blood type and RH factor go together.

16. What proportion of the patients were "A negative"? (Write your answer as a fraction, decimal and percentage).

17. What percent of the patients were "O positive"? (Write your answer as a fraction, decimal and percentage).

Section 1C – Pie Charts and a Bar Charts with Technology

A quick way to count how many people or objects have a certain label is to create a Bar Chart or Pie Chart. There are many statistics software that we could use to create these graphs. They are useful to show the characteristics of categorical data.

Creating a Pie Chart with Raw Categorical Data and Statcato

A pie chart is a very useful graph and can give the count (or frequency) for each variable and the percentages for each variable.

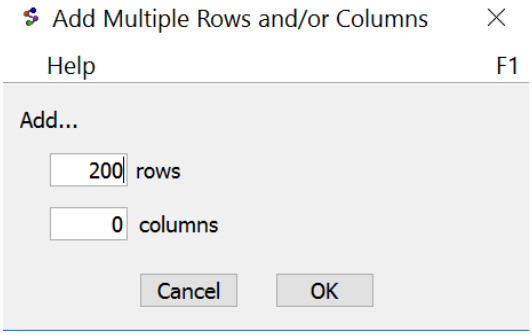
To create a pie chart with Statcato, open your excel spreadsheet. Copy and paste your column of categorical data from Excel into Statcato. Before pasting, be sure to click on the gray at the top of the column in Statcato, since titles must go in the gray. Now click on the graph menu at the top and then "pie chart". Click on "data values from a worksheet" and then under "data" put in the column. If your data is in the first column, you will click on "C1". If it is in the second column, you will click on "C2", and so on. Give the chart a title and click on "Show Legends" and "Show Values/Percentages for each Pie Sector". You can sort the graph by category or by frequency (counts). If you click on "sort by category", the pieces will be put in alphabetical order clockwise around the circle. If you click on "sort by frequency," then the chart will be organized from the smallest section to the largest section clockwise around the circle.

Graph Menu => Pie Chart => Data Values from a Worksheet => Sort by Categories or Frequencies, Show Legend, Show Values/Percentages

Let us look at an example.

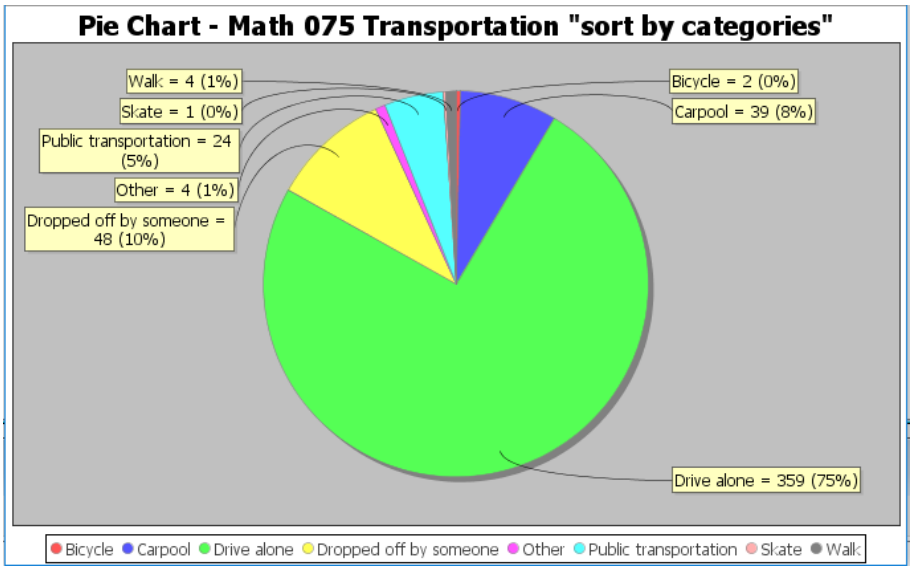
Open the Math 075 Survey Data from Fall 2015. The students were asked what type of transportation they take to get to college.

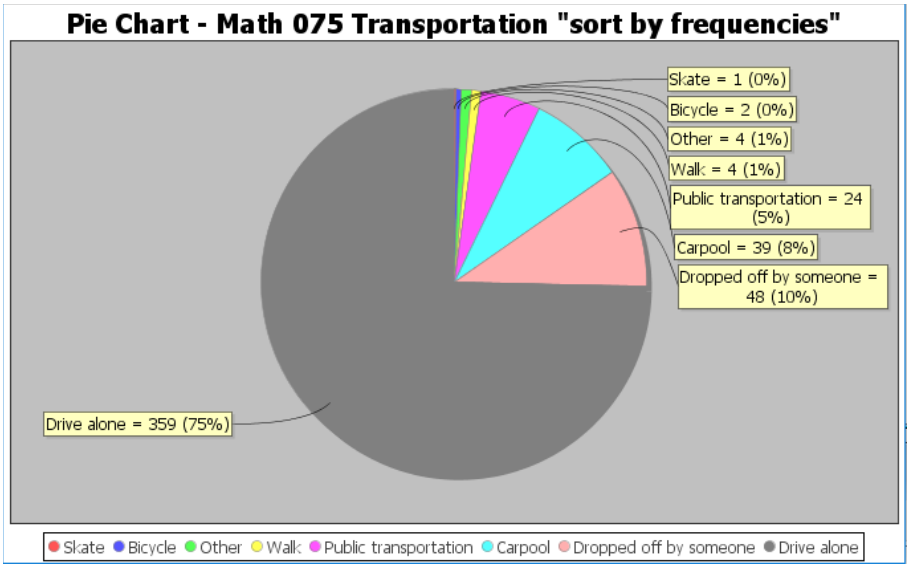
<u>Important Reminder</u>: If your data set is over 300 entries, you will need to add some rows to Statcato. The math 075-survey data had close to 500 students, so we will need to add some rows to the spreadsheet in Statcato before copy and pasting from Excel. (I added 200 more rows to Statcato before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



Once you have added enough rows in Statcato, copy and paste the column of data that says "Transportation" in Statcato. Do not forget to put the title in the gray cell at the top. Now go to the graph menu and make a

pie chart. We will show two versions of the graph. One if you sort by categories and the other if you sort by frequencies. That way you can see the difference and which one you like better. You can copy and paste the graph into a Word or Pages document, by going to the "graph" button on the left side of the graph and click on "copy graph to clipboard".





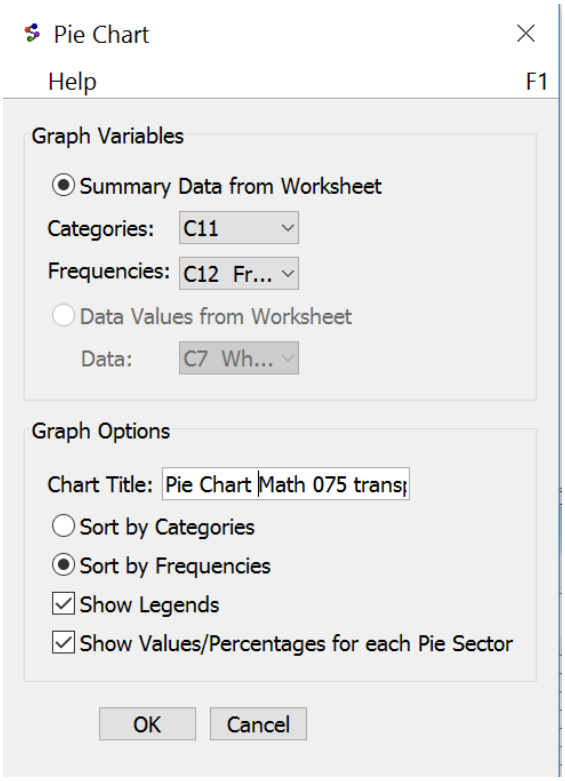
Notice at the touch of a button, the computer can tell us all of the counts (frequencies) and all of the percentages. We can answer all sorts of questions about how these students get to the college.

Creating a Pie Chart with Summary Data and Statcato

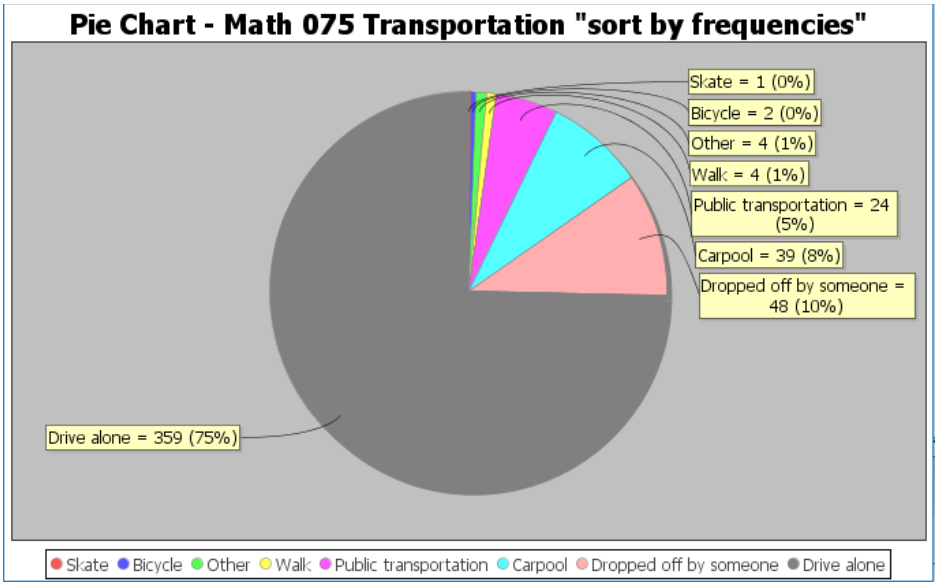
Categorical data is often given in summarized form with the variables and the counts. Statcato can also make a pie chart from summarized data. Suppose we do not have access to the raw categorical transportation data. Suppose we only knew the variable labels and the counts.

C11	C12
	Frequencies
Skate	1
Bicycle	2
Other	4
Walk	4
Public Transportation	24
Carpool	39
Dropped off	48
Drive Alone	359

Now go to the graph menu and then "pie chart". Click on "Summary Data from Worksheet". Give the columns for the categories and the columns for the frequencies.



Notice the pie chart looks the same.



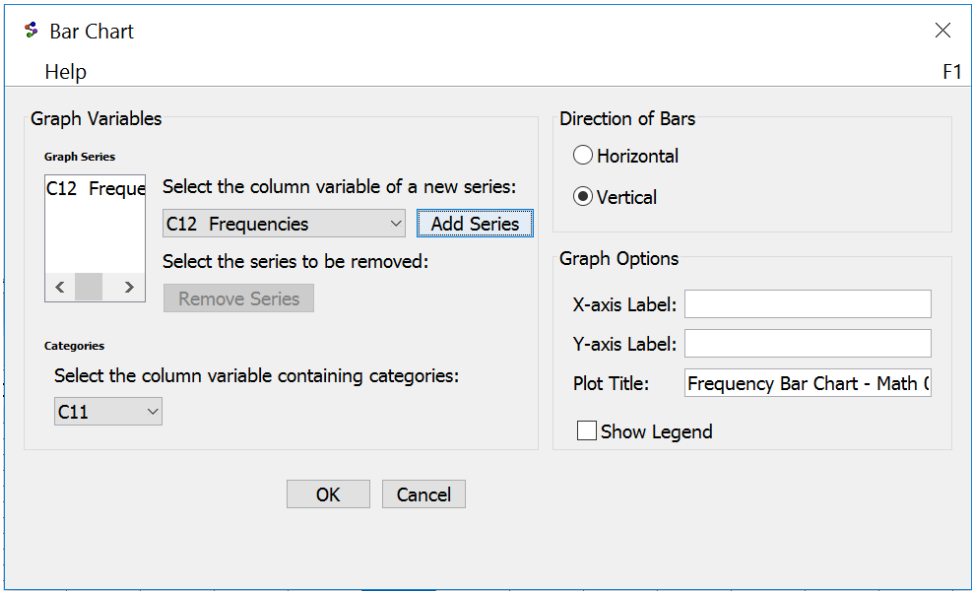
Create a Bar Chart with Statcato

We can also create a bar chart for categorical data. This shows a bar for each variable. The bar gives the count (frequency) or the percentage. In Statcato, it is a little challenging to make a bar chart. You will need the variable names in your data and the counts for each variable. I often make a pie chart first to get this information.

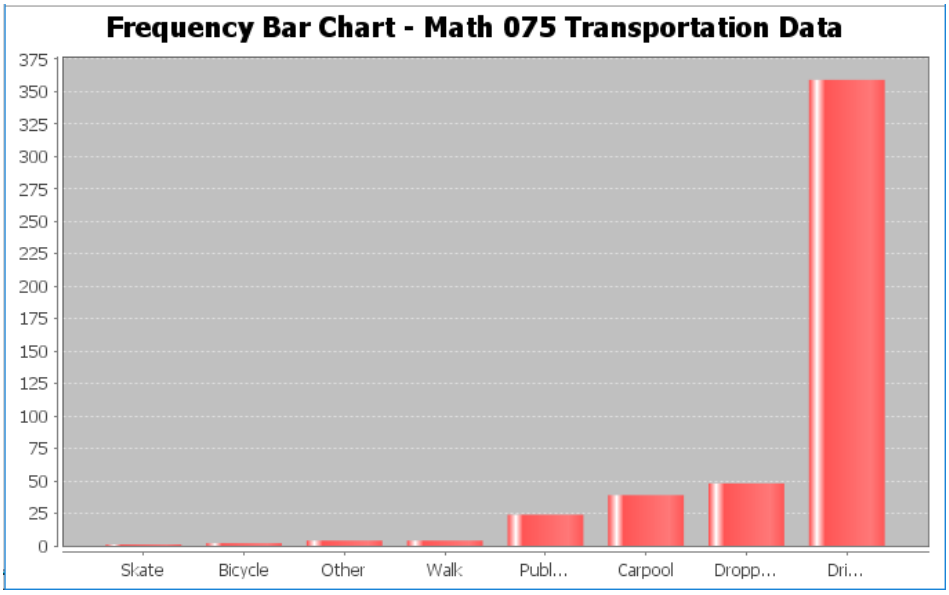
Type the variable names in one column of Statcato and the corresponding counts in the column next to it. Be careful that the counts go with the correct variable. For the transportation data, it will look something like this. I typed in my data into columns 11 and 12 but you can type them into any column you want. I also titled my counts the gray column as "frequencies". You can put the variables in any order you want. It is common for people to put them in alphabetical order or by order of frequency.

C11	C12
	Frequencies
Skate	1
Bicycle	2
Other	4
Walk	4
Public Transportation	24
Carpool	39
Dropped off	48
Drive Alone	359

Now go to the graph menu in Statcato and click on “Bar Chart”. Statcato will want to know what column has your variable names and the column that has your counts.



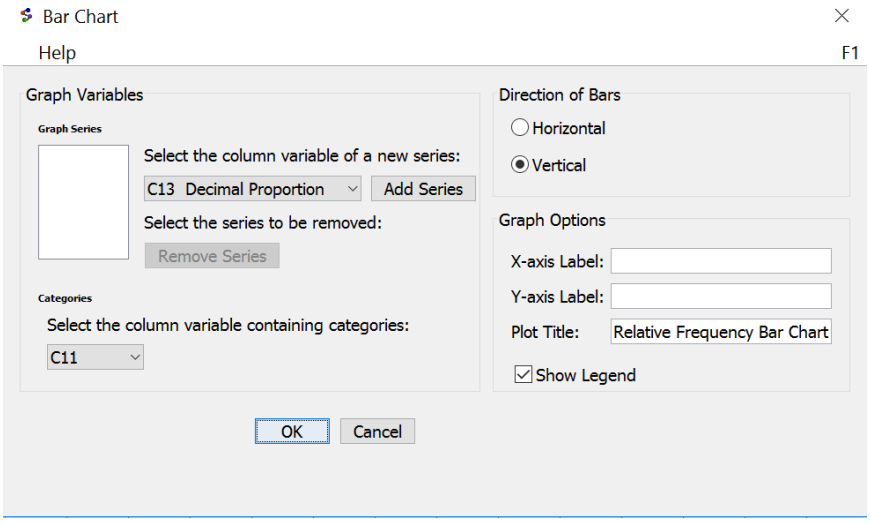
Under “Select the column variable of a new series”, pick the column with your counts (frequencies). Mine was in column 12. Now click “Add Series”. Under “Select the column variable containing categories” select the column that has your variable names. Mine was in column 11. Type in a title and “show legend” and press OK. You can make the bars vertical or horizontal as well. I used vertical in this example. Since this graph describes the frequencies for each variable, then this is often called a frequency bar chart.

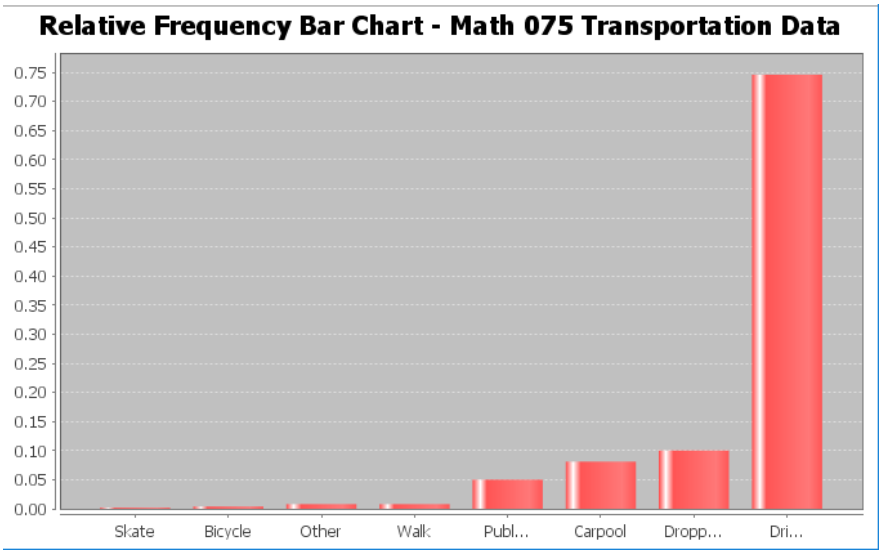


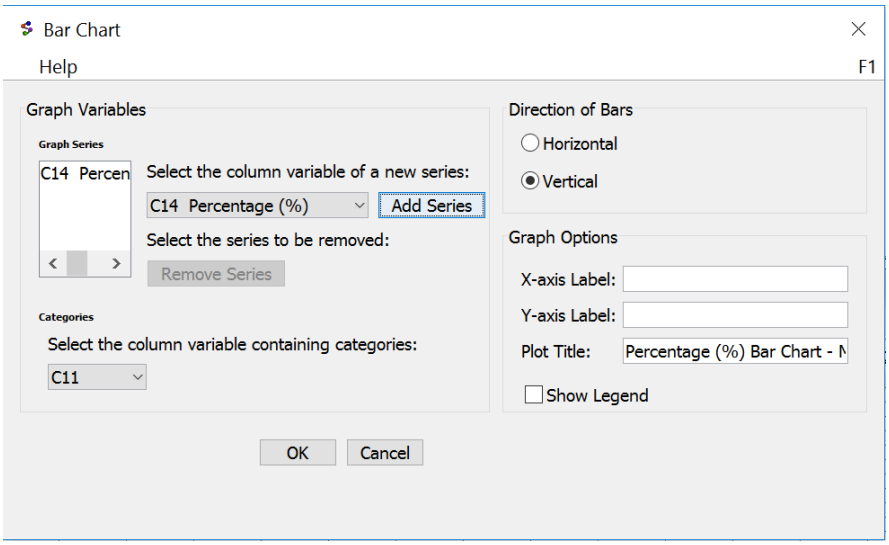
We can also make a bar chart to describe the decimal proportions or percentages for each variable as well. By dividing our counts by the total (481), we can get our decimal proportions. Multiplying by 100 gives the percentages. Note that you will not be able to put the "%" symbol on the numbers since it must be numerical data, but can label the chart that it is describing percentage. We typed the numbers into Statcato in another column.

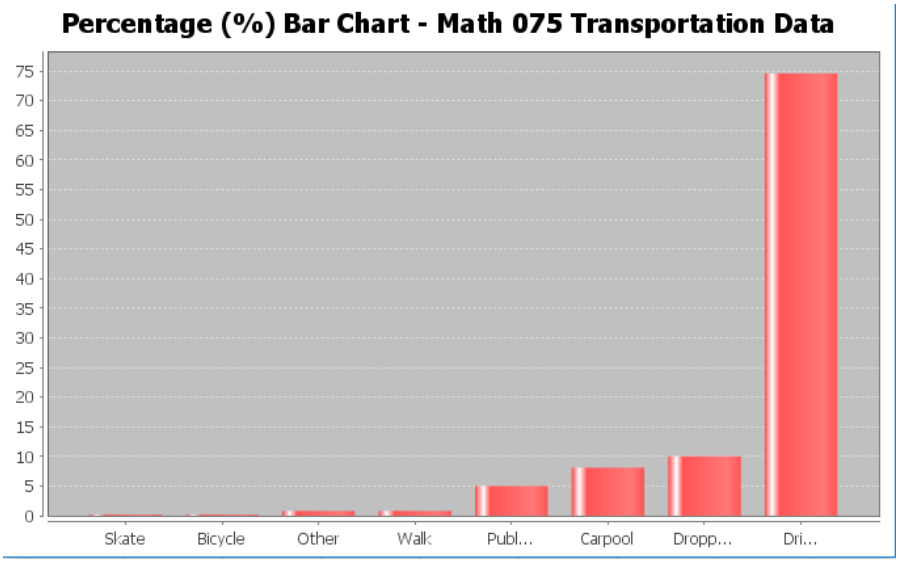
C11	C12	C13	C14
	Frequencies	Decimal Proportion	Percentage (%)
Skate	1	0.002	0.2
Bicycle	2	0.004	0.2
Other	4	0.008	0.8
Walk	4	0.008	0.8
Public Transportation	24	0.050	5.0
Carpool	39	0.081	8.1
Dropped off	48	0.100	10.0
Drive Alone	359	0.746	74.6

Now we can make a bar chart for the decimal proportions. This is often called a "relative frequency bar chart". We can also make bar chart for the percentages. This is often called a "percentage bar chart".




Relative Frequency Bar Chart - Math 075 Transportation Data
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Skate
Bicycle
Other
Walk
Publ...
Carpool
Dropp...
Dri...


Bar Chart
Help
F1
Graph Variables
Graph Series
C14 Percen
Select the column variable of a new series:
C14 Percentage (%)
Add Series
Select the series to be removed:
Remove Series
Categories
Select the column variable containing categories:
C11
Direction of Bars
Horizontal
Vertical
Graph Options
X-axis Label:
Y-axis Label:
Plot Title:
Percentage (%) Bar Chart - N
Show Legend
OK
Cancel



Problem Set Section 1C

Directions: Open the Math 075 Survey Data Fall 2015 and use Statcato to create the following graphs. Make a rough sketch of the Statcato graphs on a piece of paper or save them on a word document and answer the questions.

1. Look at the campus data. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document and answer the following questions. Were there more math 075 students at the Valencia campus or at the Canyon Country campus? How many math 075 students went to the Valencia campus? How many went to the Canyon Country campus? What percent of the 075 students went to the Valencia campus? What percent of the 075 students went to the Canyon Country campus?

2. Look at the gender data. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document and answer the following questions. Were there more female or male 075 students? How many of the 075 students were female? How many of the 075 students were male? What percent of the 075 students were female? What percent of the 075 students were male?

3. Look at the hair color data. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document and answer the following questions. Which hair color had the most students? Which hair color had the least? Give the amount of students (frequency) for each hair color. Give the percentage of students for each hair color.

4. Look at the month of birth data. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document and answer the following questions. Which month of birth had the most students? Which month had the least students? Give the amount of students (frequency) for each month. Give the percentage of students born in each month.

5. Look at the political party data. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document and answer the following questions. Which political party had the most students? Which political party had the least students? Give the amount of students (frequency) for each political party. Give the percentage of students in each political party.

Often Categorical Data is summarized with the amounts of each variable and total. Statcato can create bar charts and pie charts from summary data as well.

6. When a math 075 student asked COC students what their favorite coffee shop in Santa Clarita was, 41 said they preferred Starbucks, 27 said Coffee Bean, 19 said Peet's Coffee, and 9 said It's a Grind. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document. Give the percentage of students for each coffee house.

7. We looked at a sample of 83 retired NFL football players and found that only 18 of them were still doing ok financially, but 65 of them had gone bankrupt. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document and answer the following questions. What percent of the NFL players were bankrupt? What percent were not bankrupt? Does the percentage of bankrupt NFL players surprise you? Why?

8. We looked at a sample of 113 retired NBA basketball players and found that only 45 of them were still doing ok financially, but 68 of them had gone bankrupt. Make a pie chart with percentages and frequencies included, a regular bar chart with counts above each bar, and a percentage bar chart where each bar gives the percentage. Use the counts and percentages in the pie chart to make the bar charts. Make a rough sketch of the graphs on a piece of paper or save them on a word document and answer the following questions. What percent of the NBA players were bankrupt? What percent were not bankrupt? Does the percentage of bankrupt NBA players surprise you? Why?

Section 1D – Comparing Percentages from Multiple Groups

Statistics is based on the idea of answering questions. One of the most common questions that is often asked of a data analyst is to compare a categorical variable from multiple groups. Do men in the data have a higher percentage of Type 2 diabetes than women? Is the percentage of people that own guns lower in large cities than in rural communities? Which high schools in your community give students the best opportunity to get a scholarship to college?

These are all important questions that can be answered with technology and a good understanding of categorical data and percentages.

Note about populations: *At this point, we are learning to analyze data. For example, we can look at the percentage of men in the data set with Type 2 Diabetes verses the percentage of women. This gives us an idea about gender and diabetes but we should not apply that to all men or all women. It takes a much greater knowledge of statistical methods to apply data to millions of people. Your data set may not represent all men and all women on planet earth.*

Let us learn to think about questions we can answer from the data. Let us look at an example using the hospital data.

The data includes age, gender, blood type (A, B, AB, O), Rhesus factor (Rh + or Rh -) and part of the hospital (Medical/Surgical, Intensive Care Unit, Same Day Surgery, Emergency Room).

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

Example 1

Do patients with an age of 55 or older have a greater chance of being admitted to the emergency room than patients under 55?

Remember how to find a percentage.

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

To convert proportion into percentage, multiply by 100%.

Let us start by finding the total number of patients that are 55 or older and then see how many of them were admitted to the emergency room.

Note: This is often called a "Conditional Proportion" since we are only looking at those 55 or older and not everyone in the data set.

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

Total patients 55 and older: 11
How many of those 11 patients were admitted to the emergency room? 4

Decimal Proportion = 4/11 = 0.363636 ≈ 0.363
Percentage of older patients admitted to ER? 0.363 x 100% ≈ 36.3%

Now let us compare this percentage to patients younger than 55.

Total patients 54 or younger: 9
How many of those 9 patients were admitted to the emergency room? 1

Decimal Proportion = 1/9 = 0.111111 ≈ 0.111
Percentage of older patients admitted to ER? 0.111 x 100% ≈ 11.1%

So what does this tell us?

First, remember these percentages do not apply to all patients in every hospital, but we can see what this data suggests about the patients in this data set from this single hospital.

The percentage of patients 55 or older admitted to ER (36.3%) is definitely higher than the percentage of younger patients admitted to ER (11.1%). This is important information for this hospital and in particular the emergency room to know.

How can we tell if there is a significant difference between groups?
This is a very difficult question to answer. Many statisticians studied and worked on methods to determine significance. They invented hypothesis tests (significance tests), P-values, and many other ways to check significance. We are not at that level yet, but I find that taking a ratio of the percentages is a good way to compare.

$$\text{Percentage Ratio} = \frac{\text{Higher Percentage}}{\text{Lower Percentage}} \text{ or } \frac{\text{Higher Proportion}}{\text{Lower Proportion}}$$

If the ratio comes out to be around 2 or higher, that is usually a significant difference. If the ratio comes out around 1, that is usually not very significant. Remember, this is not the most accurate way to determine significance, but it can give us an idea.

Let us look at the ratio for the previous example.

Ratio = 36.3% / 11.1% ≈ 3.27

This tells us that patients age 55 or older in the data are over 3 times more likely to go to the emergency room than patients under 55 are. This seems significant. *(If the ratio had come out to be close to 1, then we might lean toward saying that it is not a significant difference between the groups.)*

Note: We can also calculate the ratio from the decimal proportions. Be careful to either compare the percentage or compare the decimal proportions. Do not compare a percentage to a decimal proportion.

Ratio using decimal proportions = 0.363 / 0.111 ≈ 3.27 (same correct answer)

However, 36.3% / 0.111 does not equal 3.27!!!

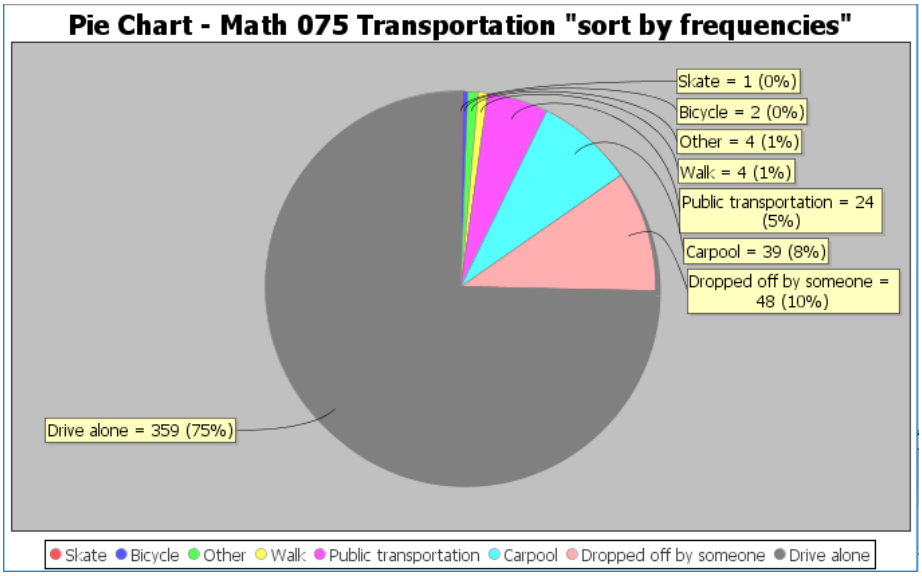
Using Technology

Remember, no data analyst counts frequencies and totals by hand, especially with large data sets. An easy way to compare percentages from one categorical data set is by making a pie chart. Here are the steps again for making a pie chart.

Graph Menu => Pie Chart => Data Values from a Worksheet => Sort by Categories or Frequencies, Show Legend, Show Values/Percentages

Example 2

Let us look again at the transportation data in the math 075-survey data fall 2015. We used Statcato to create the following pie chart.



We can answer many questions from this graph.

What percentage of students drove alone to school? What percentage of students were dropped off by someone. Calculate the percentage ratio. Is there a significant difference between the percentages?

Drive Alone ≈ 75%

Dropped off by someone ≈ 10%

Percentage Ratio = Higher % / Lower % = 75% / 10% = 7.5

It is probably obvious from the graph that the percentage of students that drive alone to college was significantly greater than the percentage that were dropped off. Notice the area of the circle for drive alone was so much greater. The percentage ratio confirms this. In fact, the percentage of students that drive alone was roughly 7.5 times greater than the percentage of those that are dropped off by someone.

Here is another question we can answer from this graph.

What percentage of students took public transportation to school? What percentage of students carpool? Calculate the percentage ratio. Is there a significant difference between the percentages?

Public Transportation ≈ 5%

Carpool ≈ 8%

Percentage Ratio = Higher % / Lower % = 8% / 5% ≈ 1.6

It is obvious from the graph that more students carpool than take public transportation, but it is not clear whether it was significantly greater. Let us look at the percentage ratio. The percentage of students that carpool was only 1.6 times greater than the percentage that take public transportation. This is a borderline case. It may be significant since it is closer to two than to one, but it is not as significant as the drive alone / dropped off example.

Note: If you want to compare percentages involving more than one column of categorical data, you often need to use technology to create a "two-way table". We will get to that example in the next chapter.

Problem Set Section 1D

Use the hospital data below to answer the following questions.

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

1. Is the percentage of male patients that go to ICU higher or lower than the percentage of female patients that go to ICU? Is it a significant difference?

2. Is the percentage of female patients that go to same day surgery (SDS) higher or lower than the percentage of male patients that go to same day surgery? Is it a significant difference?

3. Is the percentage of male patients that go to the Med/Surg floor higher or lower than the percentage of female patients that go to Med/Surg? Is it a significant difference?

4. Is the percentage of patients 55 or older that have type "O" blood higher or lower than the percentage of patients under 55 years old? Is it a significant difference?

5. Is the percentage of patients 55 or older that have type "A" blood higher or lower than the percentage of patients under 55 years old? Is it a significant difference?

6. Is the percentage of patients 55 or older that go to same day surgery higher or lower than the percentage of patients under 55 years old? Is it a significant difference?

7. Is the percentage of patients 55 or older that go to intensive care (ICU) higher or lower than the percentage of patients under 55 years old? Is it a significant difference?

8. Compare the four blood types (A, B, AB and O). Which blood type has the highest percentage of being RH negative? Which blood type had the lowest percentage of being RH negative? Is it a significant difference?

9. Compare the four blood types (A, B, AB and O). Which blood type has the highest percentage of being RH positive? Which blood type had the lowest percentage of being RH positive? Is it a significant difference?

Now open the math 075 Survey Data Fall 2015 in Excel. Copy and paste the gender column and the transportation column. (Do not forget to add some rows to Statcato before copy and pasting the data.)

Adding Rows in Statcato:

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

Creating a Pie chart in Statcato:

Graph Menu => Pie Chart => Data Values from a Worksheet => Sort by Categories or Frequencies, Show Legend, Show Values/Percentages

10. Look at the campus data. Make a pie chart with percentages and frequencies included and answer the following questions. What percent of the students went to Valencia? What percent of the 075 students went to the Canyon Country campus? Calculate the percentage ratio. Was it a significant difference?

11. Look at the gender data. Make a pie chart with percentages and frequencies included and answer the following questions. What percent of the students were female? What percent of the 075 students were male? Calculate the percentage ratio. Was it a significant difference?

12. Look at the hair color data. Make a pie chart with percentages and frequencies included and answer the following questions. What percentage of the students have brown hair? What percentage of the students have black hair? Calculate the percentage ratio for brown hair and black hair. Was there a significant difference between the percentage of students with brown hair and black hair?

13. Look at the hair color data. Make a pie chart with percentages and frequencies included and answer the following questions. What percentage of the students have blond hair? What percentage of the students have red hair? Calculate the percentage ratio for the students with blond hair and red hair. Was there a significant difference between the percentage of students with blond hair and red hair?

14. Look at the month of birth data. Make a pie chart with percentages and frequencies included and answer the following questions. What percent of the students were born in March? What percent of the students were born in October? Calculate the percentage ratio. Was there a significant difference between the percentages for March and October?

15. Look at the month of birth data. Make a pie chart with percentages and frequencies included and answer the following questions. What percent of the students were born in January? What percent of the students were born in June? Calculate the percentage ratio. Was there a significant difference between the percentages for January and June?

16. Look at the political party data. Make a pie chart with percentages and frequencies included and answer the following questions. What percent of the students were republican? What percent of the students were democrat? Calculate the percentage ratio. Was there a significant difference between the percentage of democrat and republican students?

17. Look at the political party data. Make a pie chart with percentages and frequencies included and answer the following questions. What percent of the students were republican? What percent of the students were independent? Calculate the percentage ratio. Was there a significant difference between the percentage of students that are republican and independent political party?

Section 1E – Using Percentage Data

If you pick up any newspaper or magazine or click on any news or sports link online, you are likely to see information summarized with percentages.

How can we use these percentages to give us a better understanding of the categorical data?

One of the most common uses of percentages is to estimate amounts from a total. Before we can do this, we need to be able to convert the percentage back into its decimal proportion equivalent. The "%" symbol means to divide by 100. Even the word "percent" refers to "per" (divide) and "cent" (100). (Think of 100 cents in a dollar.)

Convert a Percentage into a Decimal Proportion

To convert a percentage into decimal form: Remove the % symbol and divide by 100. (Or move the decimal point two places to the left)

Example 1

Convert 29.5% into a decimal proportion.

All we need to do is remove the % symbol and divide by 100.

29.5% = 29.5 / 100 = 0.295

Example 2

Convert 0.97% into a decimal proportion. (This is less than 1%)

All we need to do is remove the % symbol and divide by 100.

0.97% = 0.97 / 100 = 0.0097

Some students prefer to move the decimal point two places to the left. This is fine as well, though I find students make more mistakes with decimal point moving than with dividing a number by 100 with their calculator. Look at this example.

Example 3

Convert 5% into a decimal proportion.

Many students do not know where to move the decimal because there is no decimal shown. (They need to remember that 5% is the same as 5.0%)

A better way is to remove the % symbol and divide by 100.

5% = 5 / 100 = 0.05

Estimating an amount from percentage information

Recall the following formula.

Decimal Proportion = Amount / Total

If you do a little algebra and multiply both sides of that formula by the Total, you get the following formula.

Amount = Decimal Proportion x Total

In other words, to find an amount, convert the percentage into a decimal proportion and then multiply by the total. This is a common use of percentage information and a great way to bring meaning to articles that you read.

To summarize: To take a percentage of a total, convert the percentage into a decimal proportion by dividing by 100. Then multiply the proportion times the total.

Example 4

According to the Center for Disease Control (CDC), about 32% of Americans have hypertension (high blood pressure). According to suburbanstats.org, Tulsa Oklahoma has approximately 603,403 people living in it. If the CDC is correct and 32% of Americans have hypertension, then how many people do we expect to have hypertension in Tulsa?

Step 1: Convert 32% into a decimal proportion.

32% = 32 / 100 = 0.32

Step 2: Multiply the decimal proportion by the total.

Amount of people with hypertension = 0.32 x 603403 = 193088.96 ≈ 193,089

So approximately 193 thousand people in Tulsa have high blood pressure. This is vital information for hospitals and doctors in the Tulsa, Oklahoma area.

Problem Set Section 1E

Converting a Percentage into a Decimal Proportion: *Remove the % symbol and divide by 100. (Or move the decimal point two places to the left)*

Directions: Convert the following percentages into decimal proportions. (Remove the % symbol.)

1. 19%
2. 31.9%
3. 5.8%
4. 0.87%
5. 93.75%
6. 0.041%
7. 3.025%
8. 7.24%
9. 3%
10. 0.7%

Finding an amount by taking a percentage of a total: *Convert the Percentage into a decimal proportion, and then multiply by the total. (Amount = Decimal Proportion x Total)*

11. According to an article by CBS news, approximately 15% of Americans still do not have health insurance.
    a) Convert 15% into a decimal proportion by dividing by 100.
    b) Use part (a) to answer the following: Approximately 78,300 people live in Chino Hills CA. If the CBS article were correct, how many people in Chino Hills would we expect to not have health insurance? Round your answer to the ones place.

12. According to an article online, about 30% of Americans own guns.
    a) Convert 30% into a decimal proportion by dividing by 100.
    b) Use part (a) to answer the following: About 305,700 people live in Stockton CA. If the article was accurate, then approximately how many people in Stockton do we expect to own a gun? Round your answer to the ones place.

13. An article by the American Diabetes Association estimates that as of 2012, about 9.3% of Americans have diabetes.

a) Convert 9.3% into a decimal proportion by dividing by 100.

b) Use part (a) to answer the following: College of the Canyons has approximately 18,400 students. If the percentage were correct, how many COC students would we expect to have diabetes? Round your answer to the ones place.

14. According to a news report by www.nielsen.com, about 15.9% of Americans struggle with hunger.

a) Convert 15.9% into a decimal proportion by dividing by 100.

b) Use part (a) to answer the following: Lancaster CA has approximately 161,000 people living in it. If the percentage from the Nielsen report is accurate, then how many people in Lancaster CA may be struggling with hunger? Round your answer to the ones place.

15. According to an article by the Autism Society, about 1.47% of people in the U.S. have autism. The article also stated that the percentage is increasing every year and that Autism is one of the fastest growing disorders in the U.S.

a) Convert 1.47% into a decimal proportion by dividing by 100.

b) Use part (a) to answer the following: Van Nuys, CA has approximately 136,400 people living in it. If the percentage by the Autism Society is correct, how many do we expect to have autism?

16. Find 3 articles online or in a newspaper or magazine that have percentages in them. Give the name of the article and the author as well as the percentage information and topic.

Chapter 1 Review

Here is a list of important ideas in this chapter.

- You should be able to distinguish between categorical data and quantitative data.
- You should be comfortable with the following terms: Frequency (count), Amount, Total, Decimal Proportion, Percentage

- You should be able to calculate a decimal proportion for a category from the frequency (amount) and the total.

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

- You should be comfortable converting a decimal proportion into a percentage and a percentage into a decimal proportion.
  Percentage => Decimal Proportion: *Remove percentage symbol and divide by 100.*
  Decimal Proportion => Percentage: *Multiply by 100 and put on the % symbol.*
- You should be comfortable converting a decimal proportion into a percentage and a percentage into a decimal proportion.
  Percentage => Decimal Proportion: *Remove % symbol and divide by 100.*
  Decimal Proportion => Percentage: *Multiply by 100 and put on the % symbol.*
- You should be able create various graphs for categorical data with technology including a pie chart, frequency bar chart, relative frequency (or percentage) bar chart.
- You should be able to estimate an amount from a percentage and a total.
  Convert the percentage into a decimal proportion by dividing by 100.
  Amount = Decimal Proportion x Total
- You should be more comfortable reading and understanding articles online or in print that involve percentages.

Problem Set Chapter 1 Review

Directions: Show your work and circle your answers. You will need a scientific calculator. Formulas are given below.

(#1-4) Classify the following variables as Categorical or Quantitative.

1. The amount of money spent by customers in restaurants across the San Fernando Valley.

2. Whether or not a person uses Marijuana.

3. The types of frogs in Florida.

4. The number of cattle on various cattle ranches in Nebraska.

(#5-7) To convert a percentage into a decimal proportion: Divide by 100 and take off the % symbol

5. Convert 3.85% into the equivalent decimal proportion.

6. Convert 92.6% into the equivalent decimal proportion.

7. Convert 0.51% into the equivalent decimal proportion.

(#8-10) To convert a decimal proportion into a percentage: Multiply by 100 and put on the % symbol

8. Convert the decimal proportion 0.558 into a percentage.

9. Convert the decimal proportion 0.0032 into a percentage.

10. Convert the decimal proportion 0.093 into a percentage.

(#11-12) Missy works for a shoe store and is wondering what percent of her customers prefer Adidas shoes. She asked 47 customers what their favorite shoe was and 17 said Adidas.

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

11. What is the decimal proportion of the customers that prefer Adidas? Round your answer to the thousandths place (3rd decimal to the right of the decimal point)

12. What percent of the customers prefer Adidas? Round your percentage answer to the tenths place (1st decimal to the right of the decimal point)

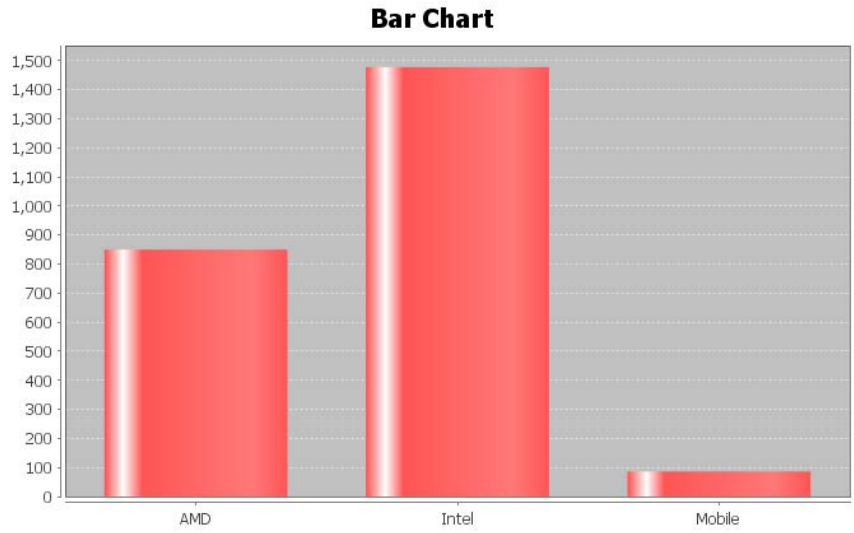
(#13-14) According to an article by www.who.int, people with HIV are highly susceptible to Tuberculosis. In fact, they say that approximately 33.3% of HIV deaths are from Tuberculosis.

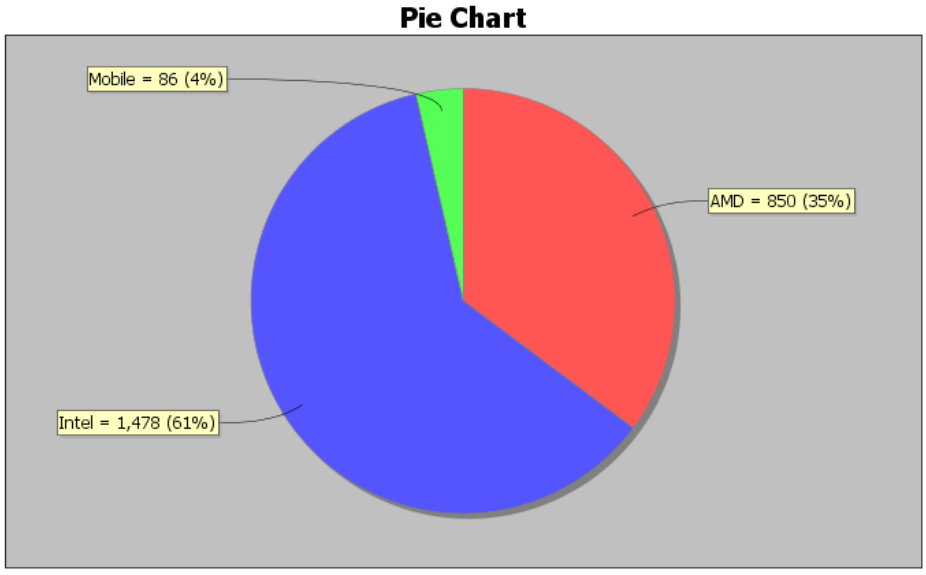
13. Convert 33.3% into a decimal proportion. (Divide by 100 and take off the % symbol.)

14. If a hospital has 58 HIV deaths, how many do we expect to be from Tuberculosis? (Round your answer to the ones place)

$$\text{Amount} = \text{Decimal Proportion} \times \text{Total}$$

(#15-18). Three of the largest producers of computer CPU's (central processing units) worldwide are AMD, Intel, and Mobile. Use the bar plot and pie chart to answer the following questions.





15. How many different processors are made by Intel?

16. How many different processors are made by AMD?

17. What percentage of the CPU's are made by Mobile?

18. What percentage of the CPU's are made by AMD?

19. What percentage of the CPU's are made by Intel?

20. Which of the three companies makes the most CPU's?

21. Which of the three companies makes the least CPU's?

22. Calculate the percentage ratio for Intel and AMD. Is there a significant difference between the percentages of processors made by the two companies? Explain why.

Project Chapter 1 - Categorical Data Analysis Group Poster

Directions: *The class will be separated into groups. Each group is required to pick a "team name" for their group and analyze one categorical data set from the math 075-survey data fall 2015, create a poster summarizing their findings, and present the poster to other students in the class.*

Each group will have a different topic and will pick one of the following data sets to present it to their classmates: 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

The Poster Should Have

- **Group/Team Name**
- **First and Last Name of each team members on the poster**
- **A few sentences explaining why this data is important or interesting to your group?**
- **Counts and Percentages Listed**
- **Regular Bar Plot**
- **Relative Frequency Bar Plot**
- **Pie Chart**
- **Does your group think that the percentages are significantly different?**
- **Do any of the percentages seem unusual or surprising?**
- **Can your group think of any reasons why the percentages are different?**
- **Decorate Poster**

Presentation

Make sure each person on the team understands the poster and can present your findings. Bring your poster to a designated presentation area in the classroom and hang or tape your poster to a wall. One person at a time will present the poster. We will then rotate so that each member of the team gets to present. Everyone else will listen to

presentations and give feedback with sticky notes.
(Be Nice!)