Introduction to Categorical & Quantitative Data Analysis

Vocabulary

Data: Information in all forms.

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.

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A population value, which is sometimes calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be. For example, a population mean average or a population percentage.

Introduction

We learned that, in order to learn about the world around us, we need to collect and analyze data. Our goal is to understand populations. Sometimes we can collect data from everyone in the population (census) and sometimes we can only collect data from a small subgroup of the population (sample). Either way, once we have the data, we need to be able to analyze it. This chapter focuses on the basics of data analysis. If you remember, there are two types of data, quantitative (numerical measurements) and categorical (labels). We analyze quantitative data very differently than categorical data, so it is always vital to ask yourself a couple key questions.

- Was the data collected correctly, either an unbiased census or an unbiased large random sample?
- Is the data quantitative or categorical?
- Is their one data set or are we trying to analyze relationships between two data sets?

We will learn about rules for judging sample sizes in the next few chapters. This chapter focuses on being able to analyze the sample data or census data you have.

When analyzing data we rely on numbers calculated from the data that can help us understand the key features of the data set. If these numbers were calculated from a sample, they are called statistics. If these numbers are calculated from an unbiased census, they are called parameters. Most of the time, we only have sample data, so it is vital to understand and explain statistics.

Note on calculation: We live in the age of "big data". No one today calculates statistics by hand, especially for a data set of ten-thousand values. Even a sample of one-hundred can be overwhelming to calculate. Statisticians and data scientists rely on computers to calculate statistics. The focus should be on understanding the meaning and correct use of the statistic, not on calculating by hand with a calculator.

Section 1E – Categorical Data Analysis

Vocabulary

Percentage (%): An amount out of 100. For example if 72 out of every one-hundred employees opts to use a company's HMO insurance, we would say that 72% of the employees are using the HMO insurance.

Proportion: The decimal equivalent of a percentage. To calculate, divide the percentage by 100 and remove the percent symbol.

Proportion and Percentage Conversions

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".

To convert a decimal proportion into a percentage, we multiply the proportion by 100%. This moves the decimal point two places to the right. Do not forget to add the % symbol.

Example: Convert 0.047 into a percentage.

$0.047 \times 100\% = 4.7\%$

To convert a percentage into a decimal proportion, we divide by 100 and remove the percentage symbol. This moves the decimal two places to the left. Do not forget to remove the % symbol.

Example: Convert 52.9% into a decimal proportion.

$52.9\% = 52.9 \div 100 = 0.529$

Calculating Proportions and Percentages from Categorical Data

In order to calculate a decimal proportion from categorical data, you will need to find the amount (count, frequency) and divide by the total.

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

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.

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

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 value you are rounding. 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.

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

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

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

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.

<u>Calculating a Frequency (Count) from a Percentage</u>

How to calculate a count (frequency) from a percentage or proportion. Sometimes a percentage is given in a scientific report or in an article. For more advanced proportion analysis, the computer programs usually require the actual count (frequency). So it is important to be able to find the frequency from percentage information.

Start by converting the percentage into a proportion.

Proportion = Percentage ÷ 100 (and remove the percent symbol %).

Now multiply the proportion times the total to get the amount (frequency). This often called taking a “percentage of a total”. It is important to round your answer to the ones place since is the number of people or objects that have a certain characteristic.

Count (Frequency) = Decimal Proportion × Total.

Example

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.

Bar Charts and Pie Charts

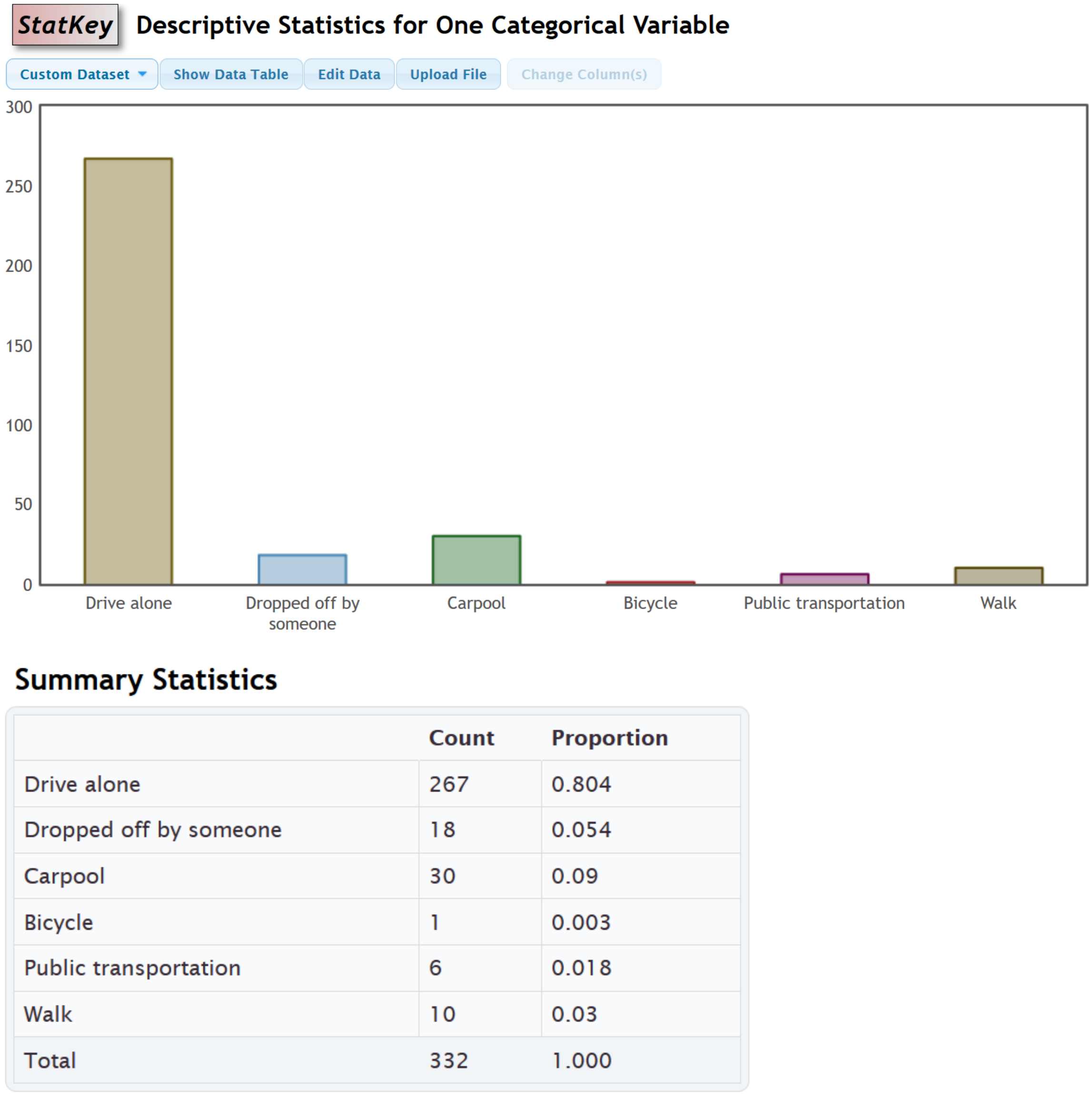
A quick way to count how many people or objects have a certain label is to create a Bar Chart or Pie Chart. There are many different statistics software that we could use to create these graphs. They are useful to show the characteristics of categorical data.

Creating a Bar Chart with Raw Data and StatKey

StatKey does not create pie charts, but does have a nice bar chart feature. It not only creates the bar chart from the raw data but also calculates the counts (frequencies) from each category as well as the decimal proportions.

To make a bar chart with raw data, go to www.lock5stat.com and click on the “StatKey” button. Now click on “one categorical variable” under the descriptive statistics and graphs button. If you have raw categorical data, click the “edit data” tab and paste your raw categorical data into StatKey. Make sure to check “raw data” at the bottom. If your data has a title, also check “data has a header row”. No click “OK”.

For example, I copied and pasted the “transportation data” from the Math 140 Fall 2015 survey data at www.matt-teachout.org into StatKey and created the bar chart. Notice it not only created the graph, but also gave me the counts (frequencies) and the decimal proportions.



Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

Creating a Bar Chart with Summary Data and StatKey

Categorical data is often summarized by the counts for each variable. When a data analyst receives categorical data to analyze, if my not be in raw form. Often it is just the counts (frequencies). In that case, when you go to the "edit data" button, you will need to type in the variables and counts as shown below. Uncheck the "raw data" box at the bottom and push "OK". Note that you need only one space after the comma and do not type in the totals. Notice you will get the exact same graphs, counts and proportions as shown above.

Response, Frequency
Drive alone, 267
Dropped off by someone, 18
Carpool, 30
Bicycle, 1
Public Transportation, 6
Walk, 10

Creating a Pie Chart with Raw Categorical Data and Statcato

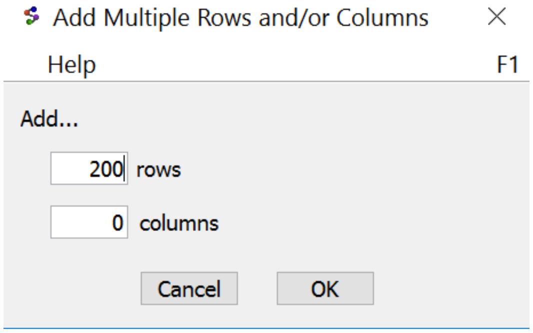
A pie chart is a very useful graph and can give the count (or frequency) for each variable and the percentages for each variable.

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 use the same example, and open the transportation data from "Math 140 Survey Data" from fall 2015.

Important Reminder: If your data set is over 300 entries, you will need to add some rows to Statcato. The math 140-survey data had close to 350 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.)


Add Multiple Rows and/or Columns

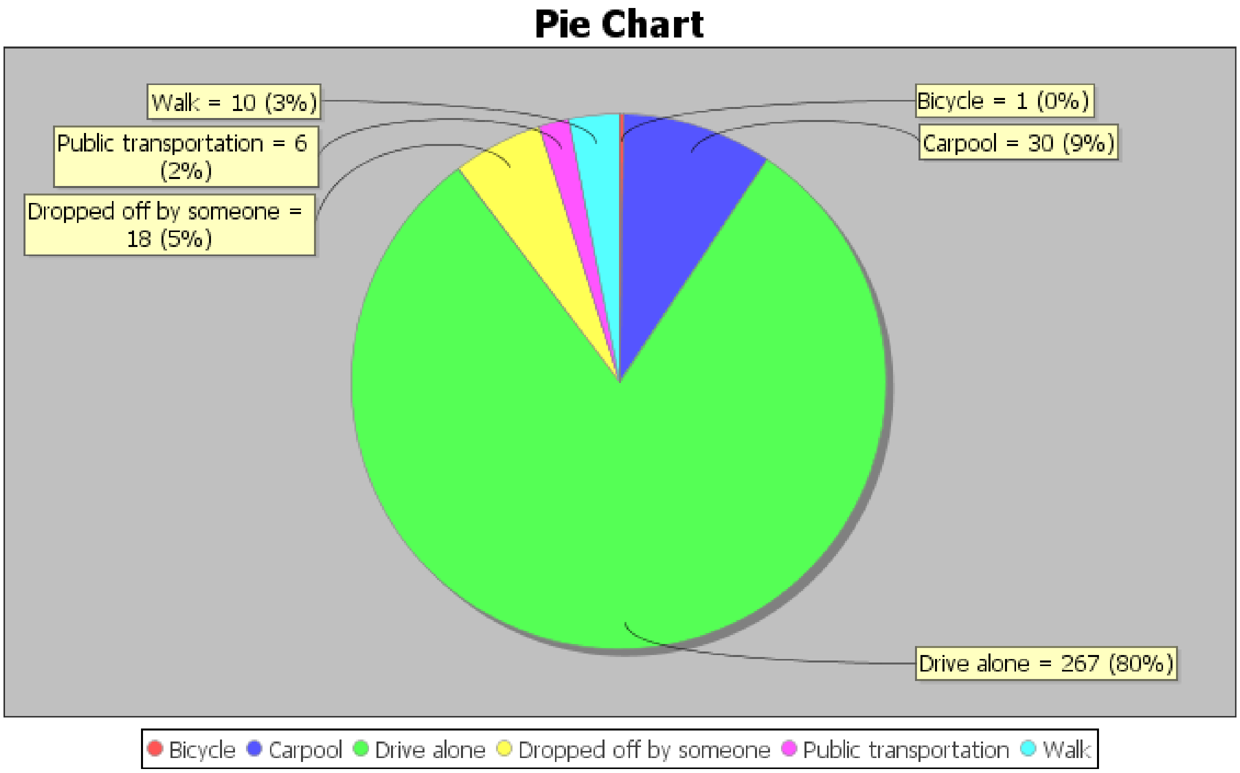
Help F1

Add...

200 rows

0 columns

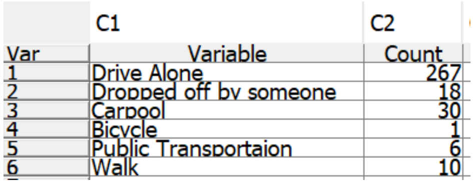
Once you have added enough rows in Statcato, copy and paste the column of data that says "Transportation" in Statcato. Do not forget to put the title in the gray cell at the top. Now go to the graph menu and make a pie chart. We will show two versions of the graph. One if you sort by categories and the other if you sort by frequencies. That way you can see the difference and which one you like better. The following graph was sorted by categories. Notice it gives the same counts as StatKey, though the proportions have been converted into percentages and rounded to two significant figures. 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.

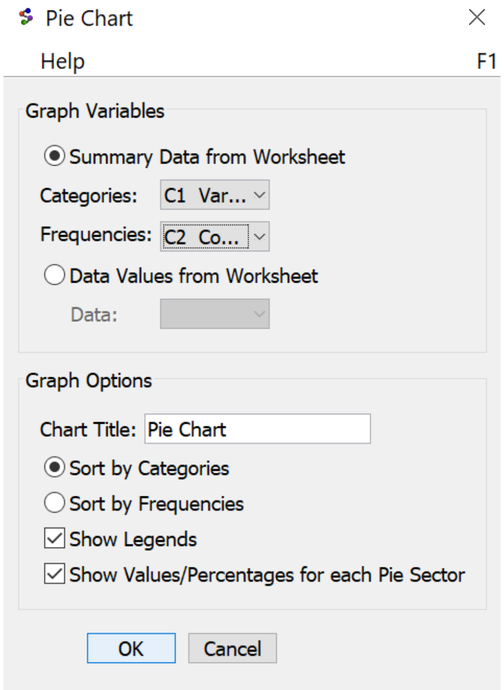
<u>Creating a Pie Chart or Bar Chart with Summary Data and Statcato</u>

Categorical data is often given in summarized form with the variables and the counts. Statcato cannot make bar charts from raw data, but it can make a bar chart from summary 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. First type in the variables and counts (frequencies) into two columns of Statcato. We will use the transportation data again. Note that titles like "variable" or "count" must be typed in the gray where it says "Var".



	C1	C2
Var	Variable	Count
1	Drive Alone	267
2	Dropped off by someone	18
3	Carpool	30
4	Bicycle	1
5	Public Transportaion	6
6	Walk	10

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.



Pie Chart ×

Help F1

Graph Variables

● Summary Data from Worksheet

Categories: C1 Var...

Frequencies: C2 Co...

○ Data Values from Worksheet

Data:

Graph Options

Chart Title: Pie Chart

● Sort by Categories

○ Sort by Frequencies

☑ Show Legends

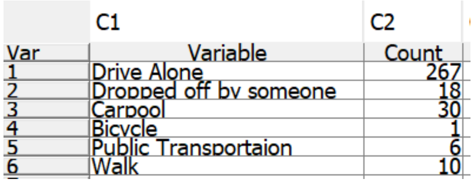
☑ Show Values/Percentages for each Pie Sector

OK Cancel

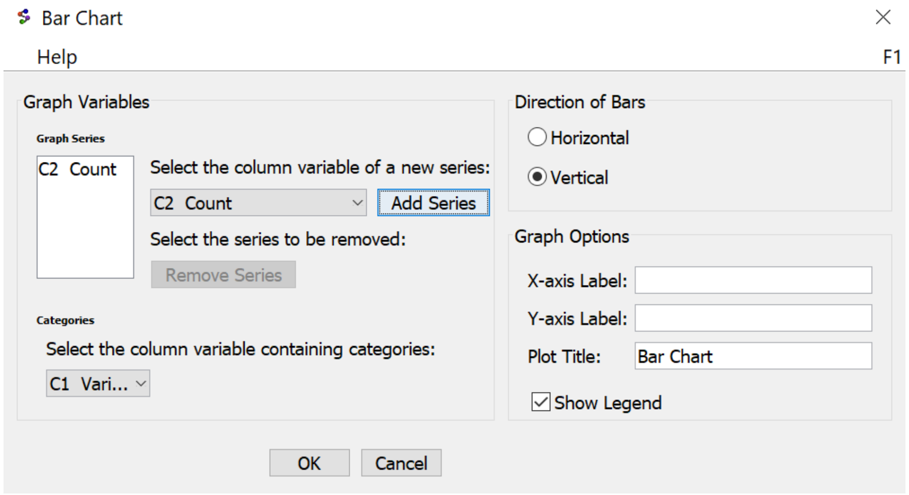
Notice the pie chart looks the same as the one we created with raw data.

We can also create a bar chart from summary categorical data. Again, type in the summary counts and variables into two columns of Statcato. Then 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 2. Now click “Add Series”. Under “Select the column variable containing categories” select the column that has your variable names. Mine was in column 1. 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.



	C1	C2
Var	Variable	Count
1	Drive Alone	267
2	Dropped off by someone	18
3	Carpool	30
4	Bicycle	1
5	Public Transportaion	6
6	Walk	10



Graph Variables

Graph Series

C2 Count

Select the column variable of a new series: C2 Count

Select the series to be removed:

Categories

Select the column variable containing categories: C1 Vari...

Direction of Bars

○ Horizontal

● Vertical

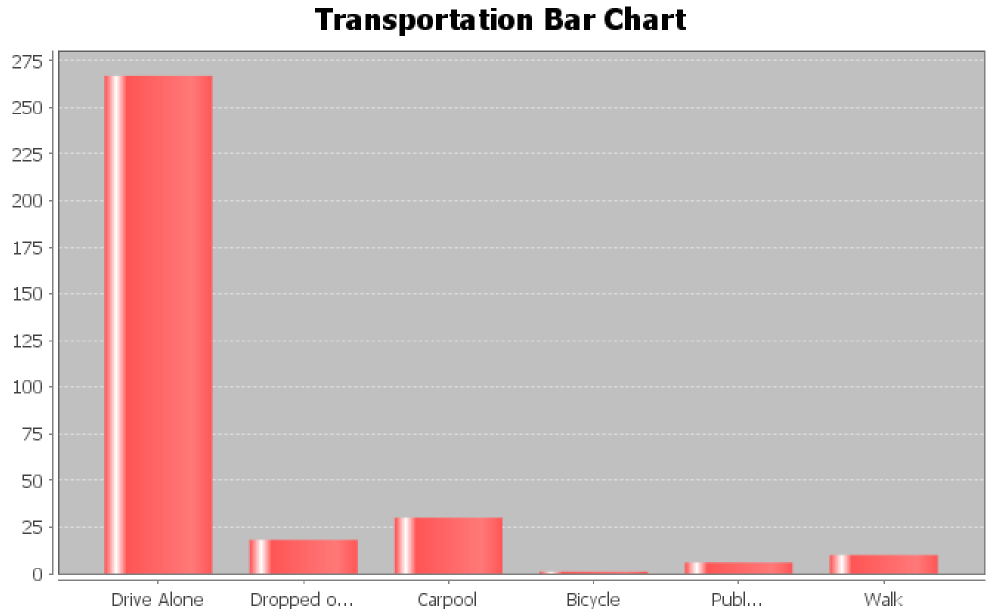
Graph Options

X-axis Label:

Y-axis Label:

Plot Title: Bar Chart

☑ Show Legend



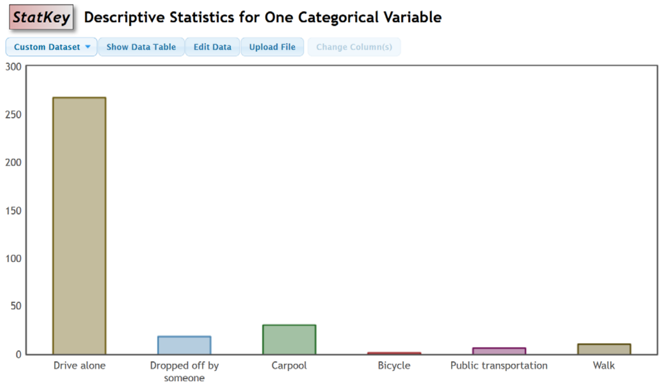
Comparing Percentages

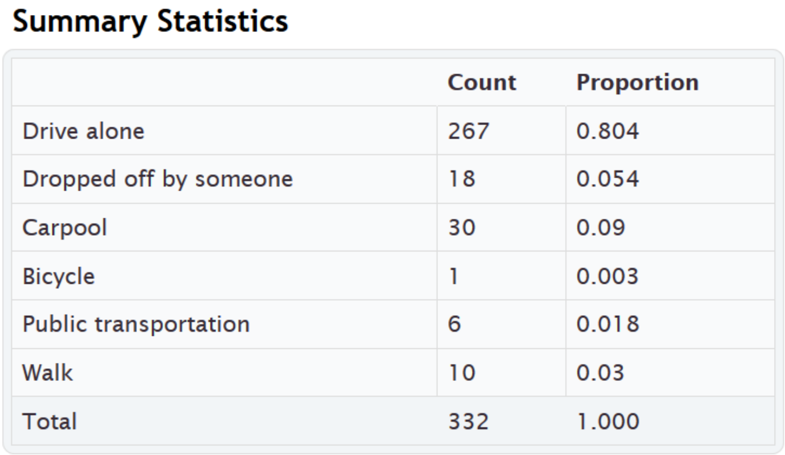
Sometimes we want to compare categorical variables and see if one variable has a significantly higher proportion or percentage than another. To compare proportion or percentages, many people often calculate the “percentage of increase”. There are three different ways of calculating the percentage of increase. Any of these formulas give the same answer.

$$\text{Percent of Increase} = \frac{(Higher\ Proportion - Lower\ Proportion)}{Lower\ Proportion} \times 100\%$$

$$\text{Percent of Increase} = \frac{(Higher\ \% - Lower\ \%)}{Lower\ \%} \times 100\%$$

For example, let us look at the transportation bar chart found with StatKey. Suppose we want to compare the percentage of math 140 students that carpool verse the percentage that were dropped off. We can calculate the percent of increase from the counts, proportions or percentages. It is important to recognize which is the lower count (frequency) and which is the higher count. In this case, the number of students that carpool was higher than the number of students that were dropped off. The key question is was it significantly higher.





Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

We can calculate the percent of increase from either the proportions or the percentages.

Percent of Increase = $\frac{(Higher\ Proportion - Lower\ Proportion)}{Lower\ Proportion} \times 100\% = \frac{(0.09-0.054)}{0.054} \times 100\% \approx 66.7\%$

Percent of Increase = $\frac{(Higher\ \% - Lower\ \%)}{Lower\ \%} \times 100\% = \frac{(9\%-5.4\%)}{5.4\%} \times 100\% \approx 66.7\%$

Notice this tells us that the proportion of students that carpool is 66.7% higher than the proportion that are dropped off. This difference seems statistically significant.

Note: In chapter 3 and chapter 4, we will learn how to use confidence intervals, test statistics, and P-values to determine significant differences. These are generally more accurate than the percent of increase calculation.

Statistical Significance verses Practical Significance

Sometimes when there is a statistically significant difference, it does not necessarily mean it is of practical use. In the last example, we saw that the number of students that carpool was a 66.7% higher than the number of students that are dropped off. Does this mean that college should make a special parking lot for all of the Math 140 students that carpool? Probably not. We are only talking about a difference of 12 total students a semester. College of the Canyons has thousands of students. So even though the percent of increase is significant, the data is not really of practical use in the sense that I would be careful of making huge decisions from the 66.7%.

Binomial Proportions with Statcato *(Optional Topic)*

Sometimes we want to know a percentage or proportion associated with a categorical event happening multiple times. One example of this is called a binomial proportion. A binomial proportion can be calculated from categorical data with only two outcomes (winning or losing, smoking or not, drinking alcohol or not). These are often referred to as "success" and "failure". The individuals must be independent of each other and the event (success) percentage *(p)* must be the same all the time. To calculate a binomial percentage, you will need a computer program and three bits of information, the number of events (number of successes), the event proportion *(p),* and the sample size *(n).*

Example

Categorical data often has a requirement of at least 10 success and at least 10 failures. Suppose we collect a random sample of 72 people and ask them whether they smoke cigarettes or not. Is 72 a large enough data set? Are we likely to get 10 or more people that smoke and 10 or more people that do not smoke? We can use Statcato to calculate this binomial percentage. According to the center for disease control, about 15.5% of adults in the U.S. smoke cigarettes.

Probability (percentage) of 10 or more people smoking =?

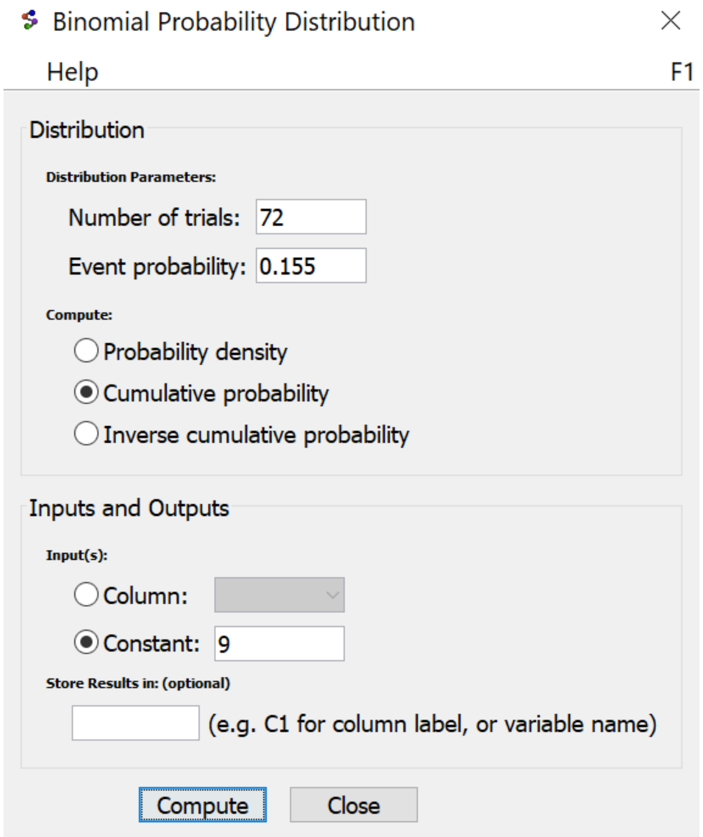
Number of Trials = Sample Size *(n)* = 72
Number of Events *(X)* = 10
Event Probability *(p)* = 0.155

Calculating binomial percentages can be challenging. Here is the formula that computer programs use.

Binomial Probability of X events: P(X) = $C(n,x)p^X(1-p)^{n-X}$

The problem with this formula is we have to calculate it for X = 10, X = 11, X = 12, … , X = 72 and then add all the proportions together. That is very difficult. It is best to let a computer program do the heavy lifting.

Open Statcato and click on "Calculate" menu. Then click on "probability distributions" and "binomial". Statcato is limited in the sense that it only calculates binomial percentage for either equal to (probability density) or less than or equal to (cumulative probability). So if we are calculating a greater than question, we must think about the opposite (less than or equal to). In this problem, we want to find 10 or more. The opposite of this would be 9 or less. Therefore, we will calculate the percentage for 9 or less, and then subtract the answer from 100%. This is sometimes called a "complement" proportion. In Statcato, put in the following. Under "Number of trials", put in the sample size 72. Under "constant" put in the number of events 9. Under "Event probability", put in 0.155. Now push the "Cumulative Probability" button and push "compute".



Binomial Probability Distribution

Distribution

Distribution Parameters:

Number of trials: 72

Event probability: 0.155

Compute:

○ Probability density

● Cumulative probability

○ Inverse cumulative probability

Inputs and Outputs

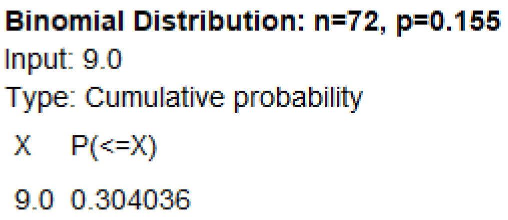
Input(s):

○ Column:

● Constant: 9

Store Results in: (optional)

(e.g. C1 for column label, or variable name)



```
Binomial Distribution: n=72, p=0.155
Input: 9.0
Type: Cumulative probability
 X    P(<=X)
 9.0  0.304036
```

Notice the probability of getting 9 or less is 0.304 or 30.4%. This is the complement percentage to what we are looking for. So the probability of getting 10 or more people that smoke should be 100% − 30.4% = 69.6%. This may not be a high enough percentage to assure us that we will get at least 10 people that smoke. I would recommend collecting more data (increase the sample size).

Example

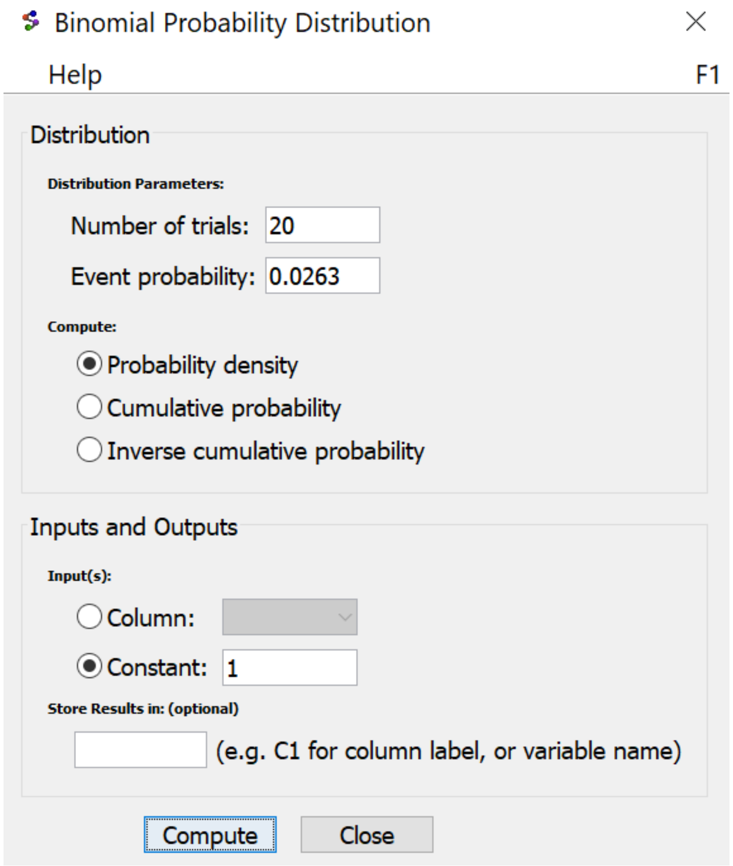
Suppose a person is playing a game of roulette that has a 1/38 or 2.63% chance of winning. The gambler plans to play the game 20 times. What is the probability that he or she wins just once?

Open Statcato and click on “Calculate” menu. Then click on “probability distributions” and “binomial”. Remember to calculate equal, you need to click on the “probability density button”.

Number of Trials = 20

Event Probability = 0.0263

Number of Events = 1 *(Put this in the “constant” box.)*



Distribution

Distribution Parameters:

Number of trials: 20

Event probability: 0.0263

Compute:

● Probability density
○ Cumulative probability
○ Inverse cumulative probability

Inputs and Outputs

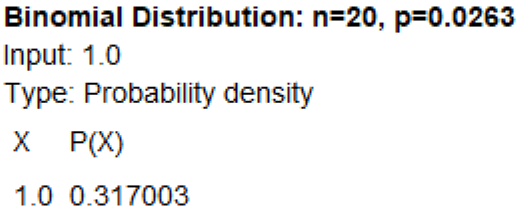
Input(s):

○ Column:

● Constant: 1

Store Results in: (optional)

(e.g. C1 for column label, or variable name)



```
Binomial Distribution: n=20, p=0.0263
Input: 1.0
Type: Probability density
 X    P(X)
 1.0  0.317003
```

Notice the answer can be found under "P(X)". So the gambler has a 0.317 (31.7%) chance of winning the game once.
