Getting Started with StatCrunch

Getting and Cleaning Data

Loading Data

Data can be found in a variety of locations. For this project, we will download a data file from the internet. To do so, follow these steps:

1. Go to the course website and click on the **StatCrunch** tab on the left-hand side of the screen. Then click on the **StatCrunch website** link.
2. You should now be at your StatCrunch homepage, and you can see the headings: **My Preferences**, **My Data**, **My Results**, etc. Under the **My Data** heading, click on **Enter the www address of a file** link. Then
    1. In the **WWW address** field enter: http://webspace.ship.edu/lebryant/MAT117Data/MAT117FirstDaySurvey.csv.
    2. Make sure the **Use first line as column names** box is checked.
    3. For **Delimiter**, select **comma**.
    4. Make sure the **Share with everyone** option is set to No.
    5. At the bottom of the screen click on **Load File**.
3. We should now see the data matrix of the first day survey that most of you completed. Notice that a few semesters have been included along with your responses this semester. Below is a picture of first six rows of the data matrix you should be seeing:

MAT117FirstDaySurvey.csv

StatCrunch | Applets | Edit | Data | Stat | Graph | Help

Row	Semester	gender	class yr	states	smoke	politics	height	newspaper	polictical pai	job	TV	cash	study time
1	Fall 2012	m	j	9	n	4	79	1	i	n	0	300	10
2	Fall 2012	m	so	7	n	4	66	1		n	150	0	2
3	Fall 2012	m	so	5	n	6	67	0	r	n	120	0	10
4	Fall 2012	m	so	20	n	6	70	0	r	n	60	10	5
5	Fall 2012	m	f	6	n	6	70	1	r	n	30	0	4
6	Fall 2012	f	sr	27	n	4	65	5	d	n	300	20	15

Now we summarize this data in various ways to see if we find anything interesting or potentially useful.

Recoding Data

When we collect data, we often get several responses that really should all be the same value. For example, when the survey asked for a respondent's gender, responses for male included "Male", "m" and "male" among others. If we make a frequency table using the variable *gender* right now, StatCrunch will consider all of these responses as a different option. To fix this we need to recode the variable. Follow these steps:

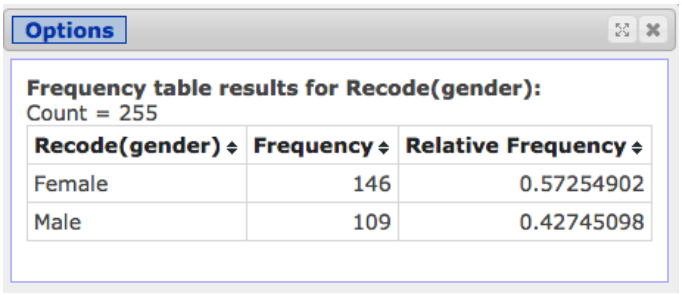
1. Click on the **Data** tab at the top of the screen and then select **Recode**.
2. A **Recode Columns** window should appear. Under **Select Column(s)** choose **gender**. Then at the bottom of the window click **Compute!**.
3. Now change every response indicating female to "Female", and every response indicating male to "Male". No one left this question blank, but if someone had, you might recode this to "No Response".
4. We should now see that a new column has been added to the data matrix named *Recode(gender)*.

Summarizing Data with Tables

Summarizing a Categorical Variable

Recall that to summarize a categorical variable we make a frequency table. Now that we have recoded the *gender* variable and created the new variable *Recode(gender)*, we can see how many respondents were male and how many were female. Follow these steps:

1. Click on the **Stat** tab at the top of the screen, mouse over **Tables**, and then select **Frequency**.
2. A **Frequency Table** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *Recode(gender)*.
    2. Under **Statistic(s)** choose *Frequency* and *Relative Frequency*. You can choose both using *shift+click*.
    3. At the bottom of the window click **Compute!**.
3. Now a new window appears displaying the frequency table. Recall that "relative frequency" is another term for "proportion".
4. To save this frequency table for later use, click on **Options** in the frequency table window and choose **Save**. Give it a name that will make sense to you later, and then click on **Save!**.
5. We can now view this table again by returning to the your StatCrunch homepage (where you saw the headings: **My Preferences**, **My Data**, **My Results**, etc.), and click on the **My Results** heading.
6. Below is the frequency table you should see.


Options

Frequency table results for Recode(gender):
Count = 255

Recode(gender)	Frequency	Relative Frequency
Female	146	0.57254902
Male	109	0.42745098

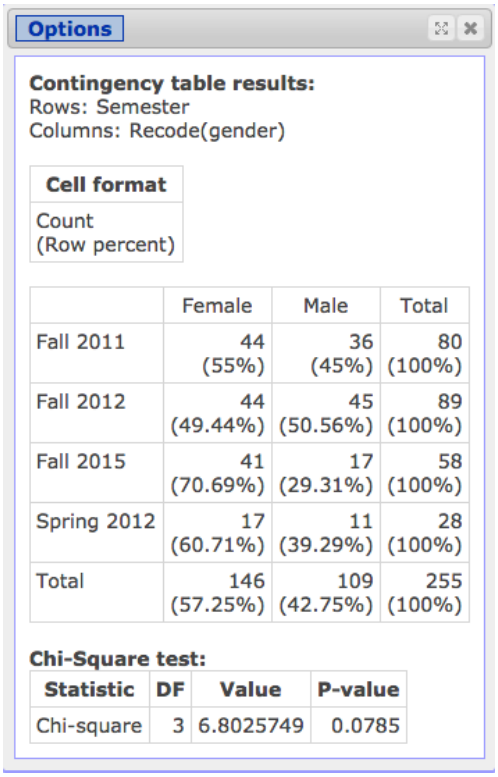
Summarizing Two Categorical Variables

Recall that to summarize two categorical variables we make a contingency table. This will let us see if there might be an association between the variables, or if they are most likely indepdendent. Consider the following question: Is the proportion of males and females in MAT117 the same every semester? We have some data that might help us answer the question. In particular, we collected gender and semester data. To investigate this question, follow these steps:

1. Since we recoded the variable *gender*, we know that *Recode(gender)* is ready to use. However, we should check to make sure that *Semester* is ready to use as well. Go through the steps of recoding a variable, and you should see that responses for the variable *Semester* have been entered consistently.
2. Next, click on the **Stat** tab at the top of the screen, mouse over **Tables**, mouse over **Contingency**, and then select **With Data**.
3. A **Contingency table (with data)** window should appear. Since it makes the most since to ask: Does the semester affect the proportion of males and females, we will make *Semester* the explanatory variable and *Recode(gender)* the response variable. It is common to make the rows of the contingency table correspond to the explanatory variable, and the columns correspond to the response variable. So, In this window, do the following:
    1. Under **Row Variable** choose *Semester*.

2. Under **Column Variable** choose *Recode(gender)*.
3. Under **Display** choose *Row percent*.
4. At the bottom of the window click **Compute!**.

4. Now a new window appears displaying the contingency table. For now, you can ignore the Chi-Square Test at the bottom of the window (we will cover this later in the course).
5. We can now save and use the table later just as we did for the frequency table.
6. Below is the frequency table we should see.



Options

Contingency table results:
Rows: Semester
Columns: Recode(gender)

Cell format
Count (Row percent)

	Female	Male	Total
Fall 2011	44 (55%)	36 (45%)	80 (100%)
Fall 2012	44 (49.44%)	45 (50.56%)	89 (100%)
Fall 2015	41 (70.69%)	17 (29.31%)	58 (100%)
Spring 2012	17 (60.71%)	11 (39.29%)	28 (100%)
Total	146 (57.25%)	109 (42.75%)	255 (100%)

Chi-Square test:

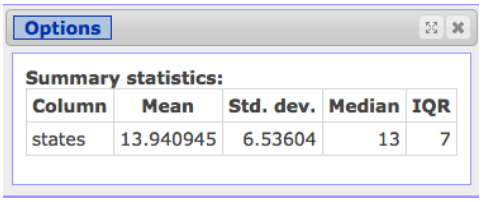
Statistic	DF	Value	P-value
Chi-square	3	6.8025749	0.0785

So, now that we see the contingency table, what do you think? Does the proportion of males and females vary from semester to semester?

Summarizing a Numerical Variable

We summarize a numerical variable by finding numerical summaries such as the mean, standard deviation, median, IQR, etc. Lets find some of these numerical summaries for the the variable *states*.

1. Click on the **Stat** tab at the top of the screen, mouse over **Summary Stats**, and then select **Columns**.
2. A **Summary Stats** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *states*.
    2. Under **Statistics** choose *Mean*, *Std. dev.*, *Median*, and *IQR*. We can choose multiple options using *shift+click*.
    3. At the bottom of the window click **Compute!**.
3. Now a new window appears displaying the summaray statistics.
4. We can now save and use the summary statistics later just as we did for the frequency table.
5. Below is the summary statistics table we should see.

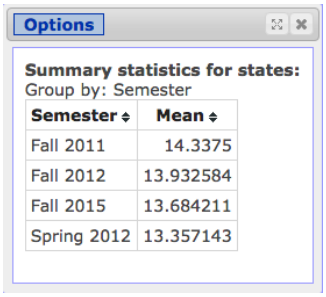


What might the fact that the mean and median are close indicate about the distribution of data for the variable *states*?

Summarizing a Categorical-Numerical Pair of Variables

Recall that to summarize a categorical-numerical pair of variables, we let the categorical variable be the explanatory variable and the numerical variable be the response variable. We look for difference between the groups determined by the categorical variable to see if there is an association between the variables. Consider the following question: On average, is the number of states visited by MAT117 the same every semester? We have some data that might help us answer the question. In particular, we collected states visited and semester data. To investigate this question, follow these steps (which are nearly identical to the steps for summarizing a numerical variable):

1. Click on the **Stat** tab at the top of the screen, mouse over **Summary Stats**, and then select **Columns**.
2. A **Summary Stats** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *states*.
    2. Under **Group by** choose *Semester*.
    3. Under **Statistics** choose *Mean*.
    4. At the bottom of the window click **Compute!**.
3. Now a new window appears displaying the mean for each semester.
4. We can now save and use the summary statistics later just as we did before.
5. Below is the table We should see.



So, now that we can compare the mean for the different semesters, what do you think? On average, is the number of states visited by MAT117 the same every semester?

Summarizing Data with Visualizations

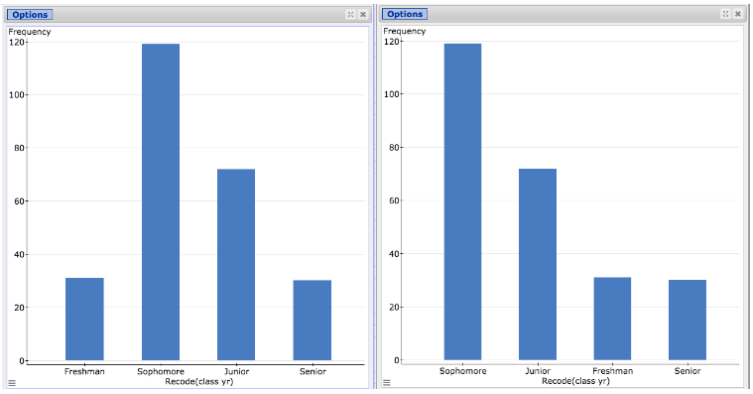
Visualizations for One Categorical Variable

We discussed two types of pictures used to summarize one categorical variable: bar plots and pie charts.

Bar plots We will make a bar plot for *class yr*. However, there are a few things to consider. First, we will want the order of the bars (from left to right) to be Freshman, Sophomore, Junior, and Senior. By default StatCrunch will order the bars alphabetically, so will need to fix that. Second, we do not want a bar corresponding to those respondents who did not answer the question. Follow these steps:

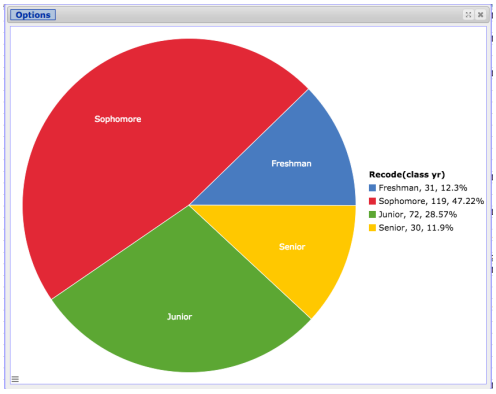
1. Recode the variable to fix how the data was entered. If someone did not give a response, change that value to "No Response". You should end up with a new variable *Recode(class yr)* with the possible values being "Freshman", "Sophomore", "Junior", "Senior", and "No Response".
2. Click on the **Edit** tab at the top of the screen, and click on Orderings. An **Orderings** window should appear. At the bottom of the window click on *Add new ordering*. Now, type "Freshman", "Sophomore", "Junior", "Senior" (in that order, one word per line). When you are done, click *Update*.
3. Click on the **Graph** tab at the top of the screen, mouse over **Bar Plot**, and then select **With Data**.
4. A **Bar Plot With Data** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *Recode(class yr)*.
    2. Under **Where** click in the field and type "Recode(class yr)" != "No Response". Be sure to include both sets of quotation marks. The != means not equal to, and this will ensure that there is not a bar corresponding to "No Response".
    3. Under **Type**, choose an option depending on what you want to display.
    4. Under **Order by**, choose *Value Ascending* for the bars to be ordered by actual class year and choose *Count Descending* for a Pareto chart.
    5. At the bottom of the window click **Compute!**.

Play around with the other properties including a title for your bar plot and labeling the axes to get a feel for what they do. Here are examples of what you should see.



Pie Chart For a pie chart, follow these steps:

1. Click on the **Graph** tab at the top of the screen, mouse over **Pie Chart**, and then select **With Data**.
2. A **Pie Chart With Data** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *Recode(class yr)*.
    2. Under **Where** click in the field and type "Recode(class yr)" != "No Response". Be sure to include both sets of quotation marks. The != means not equal to, and this will ensure that there is not a bar corresponding to "No Response".
    3. For the other options, play around with them to see what they do.
    4. At the bottom of the window click **Compute!**.

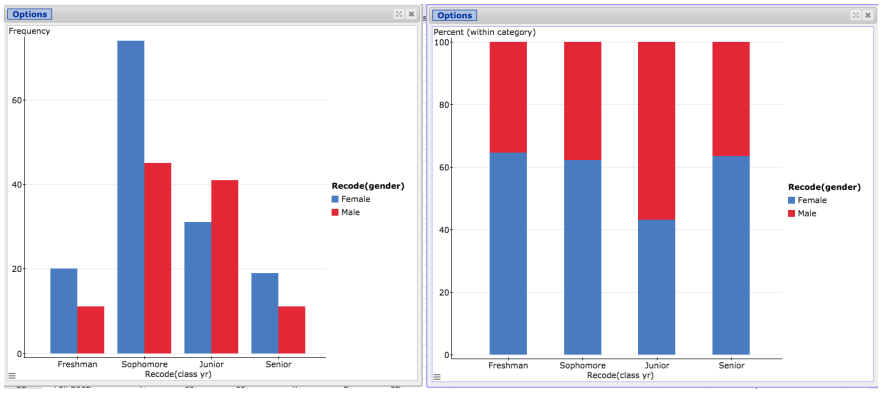


Visualizations for Two Categorical Variable

We discussed three types of pictures used to summarize two categorical variables: stacked bar plots, side-by-side bar plots, and mosaic plots. StatCrunch does not currently make mosaic plots, so we will focus on the other two. Let's consider the two variables *Recode(class yr)* and *Recode(gender)*. The steps are nearly identical to making a bar plot for one categorical variable. Follow these steps:

1. Click on the **Graph** tab at the top of the screen, mouse over **Bar Plot**, and then select **With Data**.
2. A **Bar Plot With Data** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *Recode(class yr)*.
    2. Under **Where** click in the field and type "Recode(class yr)" != "No Response" (or choose it from the dropdown menu). Be sure to include both sets of quotation marks. The != means not equal to, and this will ensure that there is not a bar corresponding to "No Response".
    3. Under **Group by**, choose *Recode(gender)*. Then under **Grouping options** choose "split bars" for a side-by-side bar plot and "stack bars" for a stacked bar plot.
    4. Under **Type**, choose an option depending on what you want to display.
    5. Under **Order by**, choose *Value Ascending* for the bars to be ordered by actual class year and choose *Count Descending* for a Pareto-like chart.
    6. At the bottom of the window click **Compute!**.

Here are some possible plots.



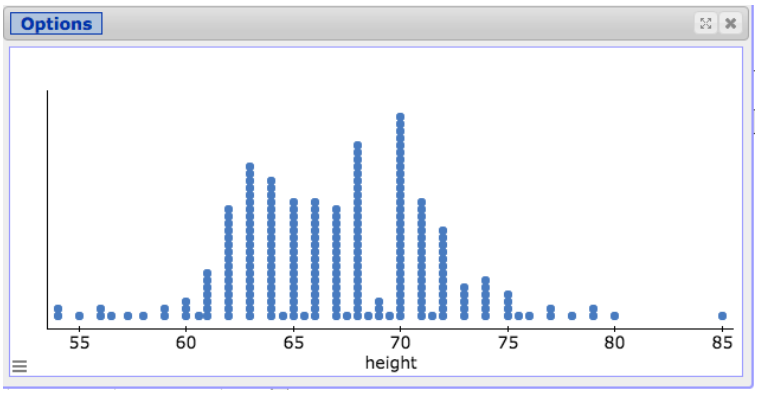
Visualizations for One Numerical Variable

We discussed three types of pictures used to summarize one numerical variable: dot plots, histograms, and box plots.

Dot plot We will make a dot plot for *height*. One thing we will find by examining the data is that some repsondents answered that they were only 5 inches tall and others that they were 115 inches tall (that over nine-and-a-half feet tall). These are certainly errors. To make our dot plot more reasonable (and accurate), let's restrict the value for the dot plot to be between 48 inches and 96 inches. Follow these steps:

1. Click on the **Graph** tab at the top of the screen, and select **Dotplot**.
2. A **Dotplot** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *height*.
    2. Under **Where** click in the field and type *height >= 48 and height <= 96*. This will ensure that only values between 48 and 96 are plotted.
    3. You can play around with the other options as you see fit.
    4. At the bottom of the window click **Compute!**.

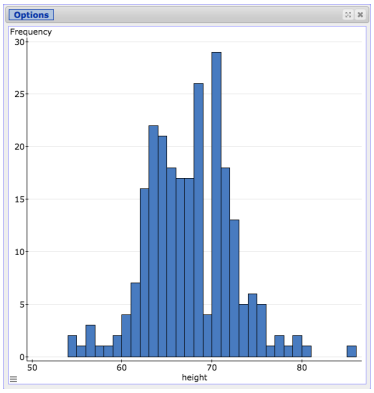
Here is what you should see.



Histogram We will also make a histogram for *height* with the same restriction between 48 and 96 inches. Follow these steps:

1. Click on the **Graph** tab at the top of the screen, and select **Histogram**.
2. A **Histogram** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *height*.
    2. Under **Where** click in the field and type *height >= 48 and height <= 96* (or choose it from the dropdown box). This will ensure that only values between 48 and 96 are plotted.
    3. Under **Bins** set the width to be 1 (or any other value you like)
    4. You can play around with the other options as you see fit.
    5. At the bottom of the window click **Compute!**.

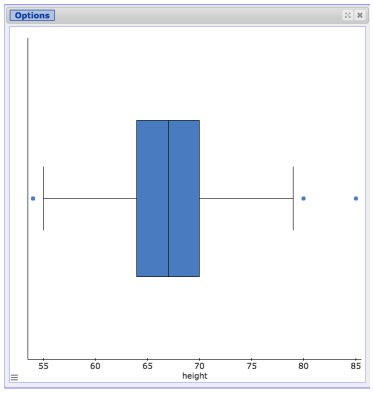
Here is what you should see.



Box plot We will also make a box plot for *height* with the same restriction between 48 and 96 inches. Follow these steps:

1. Click on the **Graph** tab at the top of the screen, and select **Boxplot**.
2. A **Boxplot** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *height*.
    2. Under **Where** click in the field and type *height >= 48 and height <= 96* (or choose it from the dropdown box). This will ensure that only values between 48 and 96 are plotted.
    3. Under **Other options** check both boxes (or don't depending on what you want to do).
    4. You can play around with the other options as you see fit.
    5. At the bottom of the window click **Compute!**.

Here is what you should see.



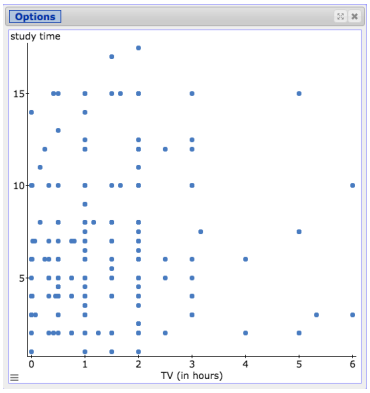
Visualizations for Two Numerical Variable

For two numerical variables we can make a scatterplot.

Scatterplot Let's make a scatterplot to compare study time with watching television. Since study time was recorded in hours and t.v. time in minutes, we will rescale the *TV* variable to be in hours as well. Also, there are some extreme values for study time including a respondent who plans to study 1,096 hours per week for MAT117! Since this drastically affects the scale of our scatter plot, let's only plot the values that are more reasonable. What is reasonalbe? If we compute Q3 + 1.5(IQR) we get 19 hours of study time (see if you can find Q3 and IQR of *study time* using StatCrunch), so lets limit study time to 19 hours or less.

1. Click on the **Data** tab at the top of the screen, mouse over Compute, and select **Expression**. Under **Expression** type *TV/60*. Under **Column label** type *TV (in hours)*. Now click on **Compute!**. This will create a new variable *TV (in hours)*.
2. Click on the **Graph** tab at the top of the screen, and select **Scatter Plot**.
3. A **Scatter Plot** window should appear. In this window, do the following:
    1. Under **X Variable** choose *TV (in hours)*.
    2. Under **Y Variable** choose *study time*.
    3. Under **Where** click in the field and type *"study time" <= 19*. This will ensure that only values less than or equal to 20 are plotted for *study time*.
    4. To add the best-fit line, under **Overlay polynomial order** choose 1. Remember, only do this if there appears to be a linear association.
    5. At the bottom of the window click **Compute!**.

Here is what you should see. Does there appear to be an association between studying and watching television?



Visualizations for One Categorical Variable and One Numerical Variable

When comparing one categorical variable and one numerical variable, we discussed makeing multiple box plots.

Multiple box plots Let's see who was carrying more cash when filling out the survey, those how have a job or those who don't. First, we must recode the variable *job* to ensure that responses such as *Yes* and *y* are treated as the same response. Also, we do not want a box plot corresponding to those who did not respond to the question about having a job. Again, we have some extreme outliers that, when plotted, make the box plots less readable. So we will only plot values of the *cash* variable that are less than or equal to $200.

1. Click on the **Data** tab at the top of the screen and select **Recode**. Proceed to recode the *job* variable so that the responses are *Yes*, *No*, and *No Response*.
2. Click on the **Graph** tab at the top of the screen, and select **Boxplot**.
3. A **Boxplot** window should appear. In this window, do the following:
    1. Under **Select Column(s)** choose *cash*.
    2. Under **Where** click in the field and type *"Recode(job)" != "No Response" and cash <= 200*.
    3. Under **Group by** choose *Recode(job)*.
    4. Under **Other options** check both boxes (or don't depending on what you want to do).
    5. You can play around with the other options as you see fit.
    6. At the bottom of the window click **Compute!**.

Here is what you should see.