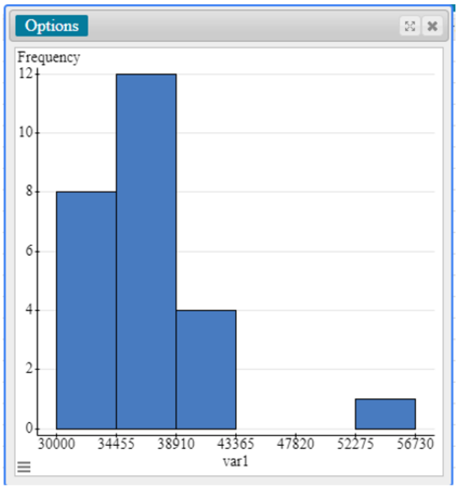
Histograms in StatCrunch

Hello! In this video I will be discussing two different ways of making histograms in StatCrunch. In the first example, we are given a table of raw data and asked to construct a histogram by finding the best values for our classes (also known as bins). In the second example, we are given a frequency table and asked to create a histogram from pre-established class sizes.

HISTOGRAM FROM RAW DATA

30,248	34,196	36,937	40,140
30,430	34,679	37,360	41,157
30,753	34,904	37,723	41,333
32,116	35,169	38,423	52,503
32,975	35,569	38,668	
33,673	35,743	38,960	



For our first example, let’s say we are given this table of 25 data values. Our problem wants us to find the optimal class sizes for this data set. StatCrunch’s histogram function will automatically format your histogram so that it looks “nice” by creating bins of automatic length. However, some homework problems require you to find your own class sizes using predefined rules, which means we can’t always use the histograms SC gives us automatically. Luckily, if we can find those class sizes by hand, StatCrunch will still create a histogram for us very quickly.

FIND CLASS WIDTH

$$\text{Number of Classes} = k,$$
$$\text{where } 2^k \geq \text{the number of observations}$$

$$25 = \text{number of observations}$$
$$2^4 < 25 \ \times$$
$$2^5 \geq 25 \ \checkmark \qquad \Longrightarrow \text{number of classes} = 5$$

$$\text{Class Width} \approx \frac{\text{Max Value} - \text{Min Value}}{\text{Number of Classes}}$$

$$\text{Class Width} \approx \frac{52440 - 30165}{5}$$

$$\text{Class Width} \approx 4455$$

Our first step is finding the number of classes that we need. There is a general rule that says that the number of classes, k, should be such that the value of 2^k is just slightly greater than or equal to the number of observations in our data set. So for example, we have 25 data values in our list. That means that we need to find a k such that 2^k is be greater than or equal to 25. So for example if we chose 4 for k, 2^k is less than 25. In other words, 2^4 is less than 25. But if we choose 5 for k, 2^5 is greater than 25. That means that instead of choosing 4 for our class size (or some other number) we would choose the value for k that makes 2^k just greater than our number of observations (25). In other words, our number of classes will be 5.

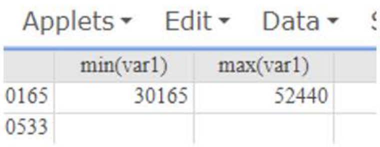
The next thing we need it to figure out our class width: how big are our bins going to be? The formula for this is found by taking the Max value in our data set subtracting the Minimum value in our data set and then dividing by the number of classes that we just found. To find our max and min values in stat crunch, there are two methods we can use.

First, we can just go to StatCrunch and order the data values by clicking on the tab next to the column name and clicking on “sort values ascending”. This puts our smallest data values at the top and our max at the bottom. From here we can easily see that the max value is 52440 and min value is 30165. From here we can plug those two numbers into our formula, divide by our number of class sizes, and find our class width.

The second method for finding the max and min is a little more labor intensive, but it is useful if you have a very large data set. First go to “Data,” then “Compute,” then “Expression.” Scroll down in the functions tab until you find the “max” function and then double click. Next double click on the variable name, here “var1.” This is just the name of the column you have your data in. Now hit “Okay.” Then compute. Notice that a new column has been created with the max value of our data set. The same process can be used to find the min. Go to “Data”, “Compute”, “Expression.” Again we go to “Build”, but this time instead of doing “max” we look for the “min” function. Again, double click on our column name, “var1.” Click “Compute.” And again, it will sort through and find our minimum value for us.

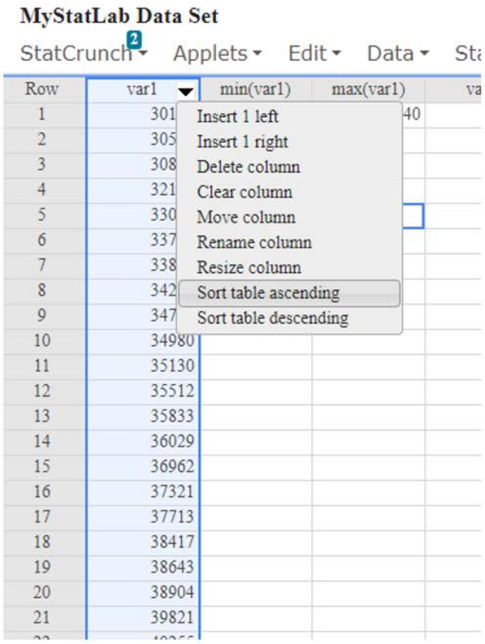
From here, (no matter what method you use) we can plug these values into our class width formula to find the size of our bins. We get 52440 minus the min of 30165. Divide this by our number of classes (which was 5), and this gives us a class width of 4455.

CHOOSE A STARTING POINT



	min(var1)	max(var1)
0165	30165	52440
0533		

$$\textit{Minimum Value} = 30165$$
$$\Downarrow$$
$$\textit{Starting Value} = 30000$$

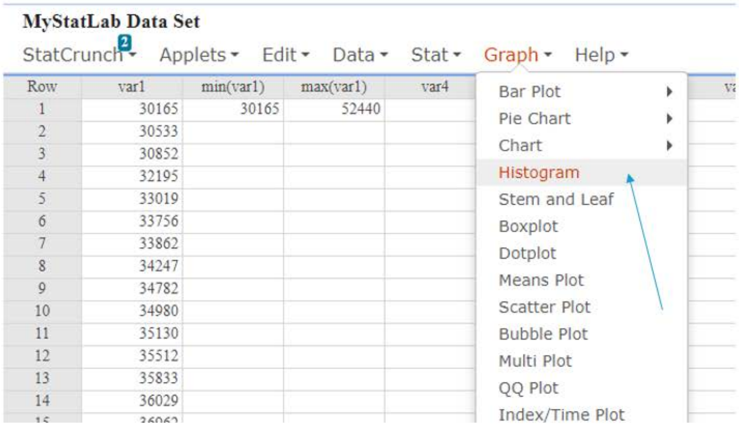


MyStatLab Data Set

Row	var1	min(var1)	max(var1)	va
1	301		40	
2	305			
3	308			
4	321			
5	330			
6	337			
7	338			
8	342			
9	347			
10	34980			
11	35130			
12	35512			
13	35833			
14	36029			
15	36962			
16	37321			
17	37713			
18	38417			
19	38643			
20	38904			
21	39821			

Next we need to choose a starting point for our histogram. To find this, we look at our minimum value (30165) and round down to the nearest “nice” number. It’s important to round it down, not up. This can vary depending on the numbers you are working with, but here our minimum value rounds down easily to 30000. The important thing when choosing a starting value is to check after you have made your histogram and make sure that all your data values are represented somewhere in your histogram.

HISTOGRAM

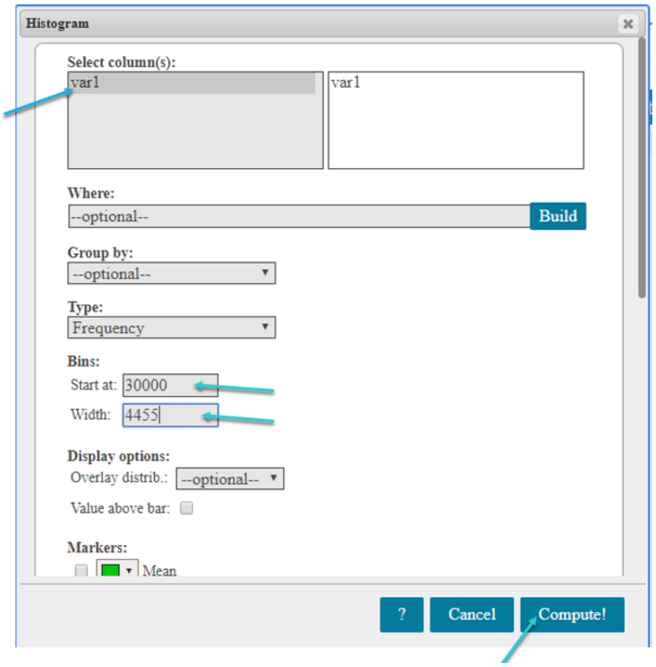


MyStatLab Data Set

Row	var1	min(var1)	max(var1)	var4
1	30165	30165	52440	
2	30533			
3	30852			
4	32195			
5	33019			
6	33756			
7	33862			
8	34247			
9	34782			
10	34980			
11	35130			
12	35512			
13	35833			
14	36029			

$$\textit{Class Width} \approx 4455$$

$$\textit{Starting Value} = 30000$$



Histogram

Select column(s): var1 | var1

Where: --optional-- Build

Group by: --optional--

Type: Frequency

Bins:

Start at: 30000

Width: 4455

Display options:

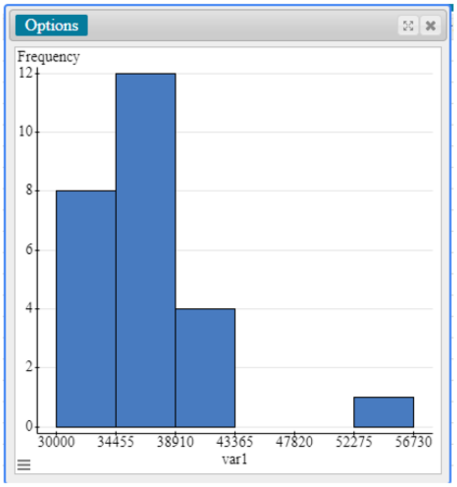
Overlay distrib.: --optional--

Value above bar: □

Markers:

□ 🟩 Mean

? Cancel Compute!



Now we are all ready to build our histogram. Go to “Graph”, then “Histogram”. Select the name of the column containing your data values. (Here “var1”) Next under “Bins”, enter your starting value in the “Start At:” box. So for our example we enter 30000.

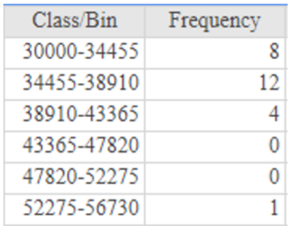
Now enter your class length in “Width:” So we will enter our width of 4455. Leave everything else alone, and then hit “Compute!”

This will create our histogram. Notice that our bins start at 30000 and all have a width of 4455. Also, the upper limit on the last data set, 56730 is bigger than the maximum value in our data set. Because of this, we know that all of our values are somewhere in our bins.

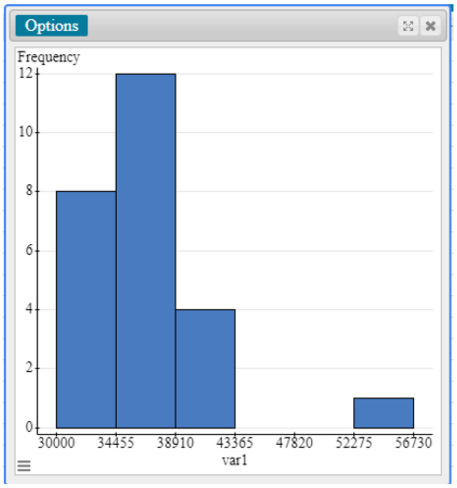
Also notice that you can hover your cursor over the bars and it will give you information about the frequency of cases in each bin and the length of the bin itself.

Also, be sure to check that our max value (52440) is in fact in one of these bins. We can do that by saying “Ok, our upper limit on the biggest bin (56730) is in fact bigger than our maximum, which means our max must be in here, which means all of our data values must be in here. In other words, we are just checking to make sure we chose our classes correctly.

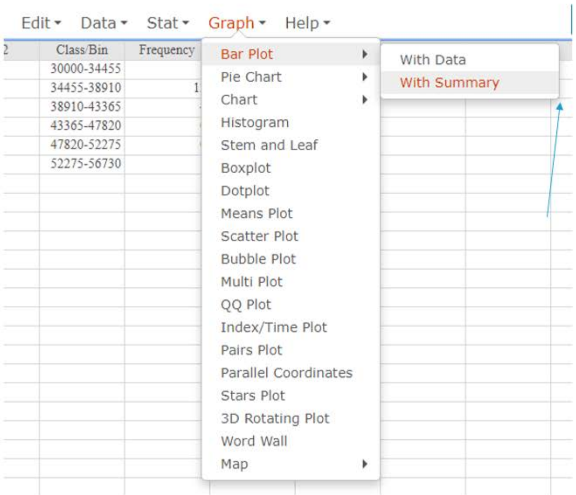
HISTOGRAM FROM FREQUENCY TABLE

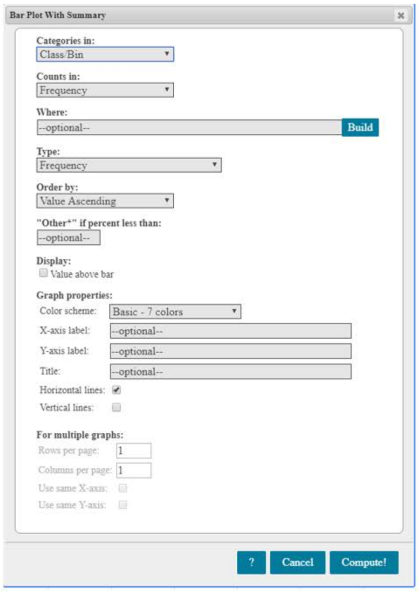


Class/Bin	Frequency
30000-34455	8
34455-38910	12
38910-43365	4
43365-47820	0
47820-52275	0
52275-56730	1



Next, for our second example, instead of being given raw data, we are instead given a frequency table. We are already given the bins, frequency, and class sizes, and they want us to create a histogram just using this. This one is a little bit simpler, we don't have to actually figure out any of the values by hand.





Bar Plot With Summary

Categories in: Class/Bin

Counts in: Frequency

Where: --optional--

Type: Frequency

Order by: Value Ascending

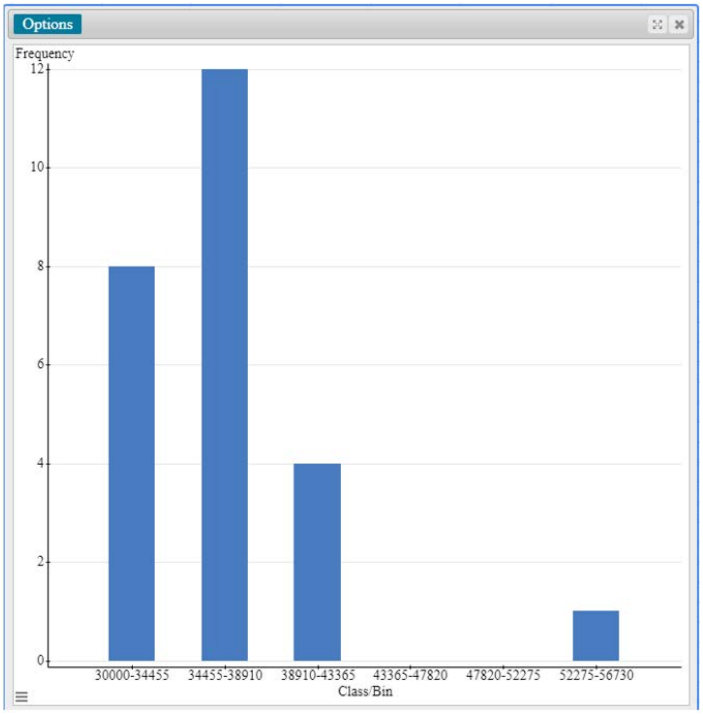
"Other*" if percent less than: --optional--

Display:
☐ Value above bar

Graph properties:
Color scheme: Basic - 7 colors
X-axis label: --optional--
Y-axis label: --optional--
Title: --optional--
Horizontal lines: ☑
Vertical lines: ☐

For multiple graphs:
Rows per page: 1
Columns per page: 1
Use same X-axis: ☐
Use same Y-axis: ☐

We can just go to straight to StatCrunch. Go to "Graph", then "Bar Plot", "With Summary". This will open up a new window. In the "Categories in:" section, select the name of the column containing your Classes/Bins. Next, in the "Counts in:" section, select the column name containing your frequencies (here it's called Frequency). Leave everything else alone, and select "Compute!"



StatCrunch will give us this bar plot. Notice that it doesn’t look the same as the histogram we got before. This isn’t a traditional histogram because the bars themselves don’t touch each other. But it still gives us all the information we need, and we can get a good sense of the distribution of the data. Again, if you hover your cursor it will give you the frequency values and the bin sizes.

And that is all there is to it. I hope you found this video helpful. If you are UNT student, you will find some links to other services and resources in the description of this video. Thanks for watching!