Summary Tools

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

Summary Tools

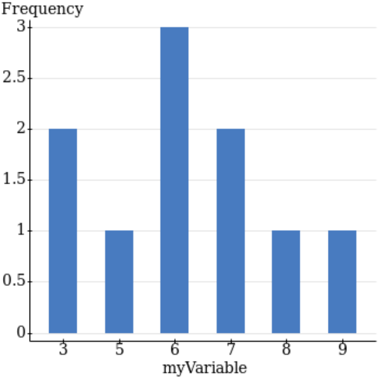
It may surprise you to learn that a professional Data Scientist or Statistican starts at the same point as you do, every time they open a new data set. They run basic tools on their data sets, many of which you've already seen in high school or even middle school. Pie charts, histograms, or maybe points plotted on an x and y axis - all are quite useful, not only for students, but for everyone!

Our goals today include review of some tools you have already seen, and some new tools that are still quite simple yet easy to use.

We will learn how to handle some of these tasks both by hand and also using StatCrunch. Some tasks, however, are really only feasable by computer.

Bar Plots

Bar plots and Histograms are really very similar. However, bar plots (or bar graphs) are used to represent categorical data, and they are usually drawn with bars separated.



Bar Plots

When drawing bar plots, be sure to label each category with its name. You can label the other axis with counts (how many are in each category) or with the overall percent. The graphic will look the same, except for the labels on one axis.

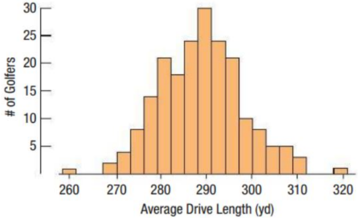
Histograms

Histograms usually represent continuous data, so to remind us of this fact, we usually draw them with their bars touching. The width of a bin is calculated as

$$\textit{bin width} = \textit{upper endpoint} - \textit{lower endpoint}.$$

Usually the upper endpoint is not included in the bin, while the lower endpoint is. However, you should not presume this is obvious to your readers. You may label each bin separately, or you can label the cut points between your bins. However, do not label a bin with a single value unless your bin truly contains just one single value. If this is the case, think carefully about whether your data are categorical or continuous.

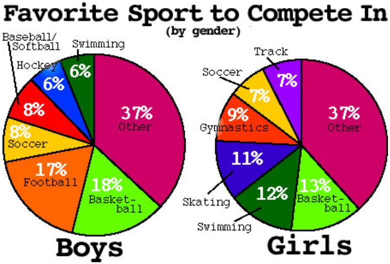
Histograms



Where do you think the top 10% of golfers fall on this graph?

Pie Charts

Bar graphs and Histograms can both be redrawn as pie charts. Here are some pie charts.



What attribute of this pair of graphs makes comparisons difficult?

Pie Charts

Note: Pie Charts are best drawn by computers! In StatCrunch, you use

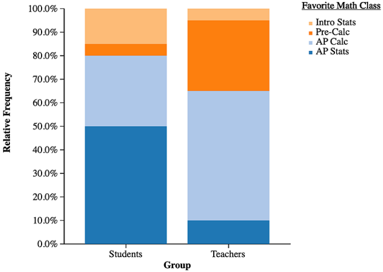
Graph > Pie Chart > With Data

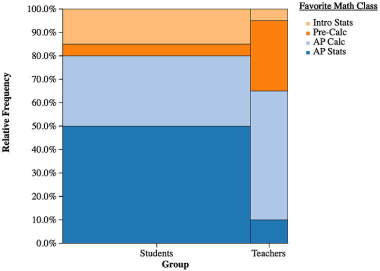
or Graph > Pie Chart > With Summary

depending upon whether you have observations or have already counted the number in each category.

Mosaic Plots

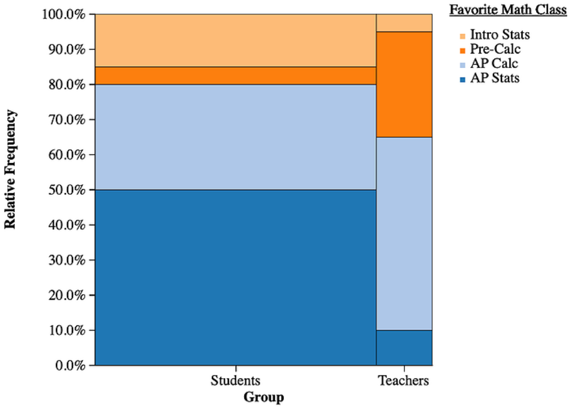
Mosaic plots are similar to bar graphs or histograms, but the widths of the bars are in proportion to their counts in the data.





This example from statsmedic.com shows a bar graph and mosaic plot that represent the same data.

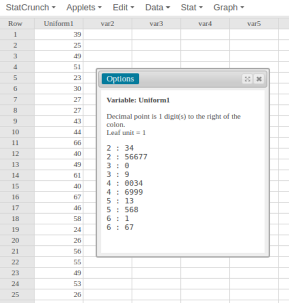
Mosaic Plots



If students and teachers voted with equal voice, which course would get the most votes? Which is larger, the total number of students who prefer AP Calc or the total number of teachers who do?

Stem and Leaf Plots

Stem and leaf plots are attributed to John Tukey, a famous mathematician who lived from 1915 - 2000. They make use of the fact that you can remove the last digit of an integer, thereby reducing your bins by a factor of 10. However, if you use the digit that you removed as a marker, you don't really lose anything. By doing this carefully, you end up with a horizontal histogram which summarizes your data graphically.



Row	Uniform1	var2	var3	var4	var5
1	39				
2	25				
3	49				
4	51				
5	23				
6	30				
7	27				
8	27				
9	43				
10	44				
11	66				
12	40				
13	49				
14	61				
15	40				
16	67				
17	46				
18	58				
19	24				
20	26				
21	56				
22	55				
23	49				
24	53				
25	26				

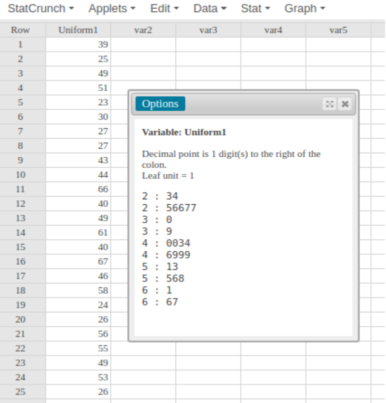
Variable: Uniform1

Decimal point is 1 digit(s) to the right of the colon.
Leaf unit = 1

```
2 : 34
2 : 56677
3 : 0
3 : 9
4 : 0034
4 : 6999
5 : 13
5 : 568
6 : 1
6 : 67
```

Stem and Leaf Plots

When you create these by hand, you first do it without regard to the order of the digits outside the stem. Then, you rewrite it with sorted leaves. Recreate this stem and leaf plot by hand.



Row	Uniform1	var2	var3	var4	var5
1	39				
2	25				
3	49				
4	51				
5	23				
6	30				
7	27				
8	27				
9	43				
10	44				
11	66				
12	40				
13	49				
14	61				
15	40				
16	67				
17	46				
18	58				
19	24				
20	26				
21	56				
22	55				
23	49				
24	53				
25	26				

Variable: Uniform1

Decimal point is 1 digit(s) to the right of the colon.
Leaf unit = 1

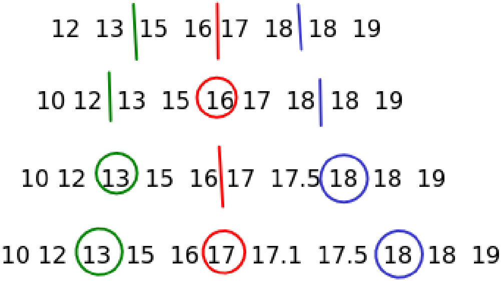
```
2 : 34
2 : 56677
3 : 0
3 : 9
4 : 0034
4 : 6999
5 : 13
5 : 568
6 : 1
6 : 67
```

Quartiles

Quartiles divide a data set into four mostly-equal parts.

- Sort the data
- Find the middle of the data (AKA “median” or Q2)
  - If you have an even number of items, take the average of the two in the middle.
- Find the median of all values less than the median (Q1)
- Find the median of all values greater than the median (Q3)

Four Quick Examples



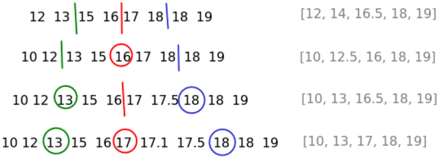
12 13 15 16 17 18 18 19

10 12 13 15 16 17 18 18 19

10 12 13 15 16 17 17.5 18 18 19

10 12 13 15 16 17 17.1 17.5 18 18 19

Five Number Summaries



12 13 15 16 17 18 18 19 [12, 14, 16.5, 18, 19]

10 12 13 15 16 17 18 18 19 [10, 12.5, 16, 18, 19]

10 12 13 15 16 17 17.5 18 18 19 [10, 13, 16.5, 18, 19]

10 12 13 15 16 17 17.1 17.5 18 18 19 [10, 13, 17, 18, 19]

Note: The five number summary includes the max and min as well as Q1, Q2, and Q3.

Q1, Q2, Q3

Find the quartiles (Q1, Q2, and Q3) for this data:

$$23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

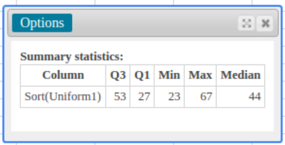
$$44, 46, 49, 49, 49, 51, 53, 55, 56, 58, 61, 66, 67$$

Q1, Q2, Q3

Find the quartiles (Q1, Q2, and Q3) for this data:

$$23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

$$44, 46, 49, 49, 49, 51, 53, 55, 56, 58, 61, 66, 67$$



Summary statistics:

Column	Q3	Q1	Min	Max	Median
Sort(Uniform1)	53	27	23	67	44

Q1, Q2, Q3

Do it again but add a 22 to the set:

$$22, 23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

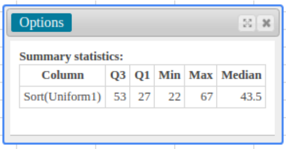
$$44, 46, 49, 49, 49, 51, 53, 55, 56, 58, 61, 66, 67$$

Q1, Q2, Q3

Find the quartiles (Q1, Q2, and Q3) for this data:

$$22, 23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

$$44, 46, 49, 49, 49, 51, 53, 55, 56, 58, 61, 66, 67$$



Summary statistics:

Column	Q3	Q1	Min	Max	Median
Sort(Uniform1)	53	27	22	67	43.5

Q1, Q2, Q3

Do it again but add a 62 to the set:

$$22, 23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

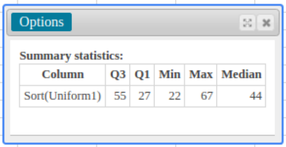
$$44, 46, 49, 49, 49, 51, 53, 55, 56, 58, 61, 62, 66, 67$$

Q1, Q2, Q3

Find the quartiles (Q1, Q2, and Q3) for this data:

$$22, 23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

$$44, 46, 49, 49, 49, 51, 53, 55, 56, 58, 61, 62, 66, 67$$



Summary statistics:

Column	Q3	Q1	Min	Max	Median
Sort(Uniform1)	55	27	22	67	44

Q1, Q2, Q3

Do it again but add a 50 to the set:

$$22, 23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

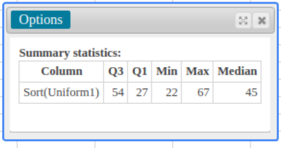
$$44, 46, 49, 49, 49, 50, 51, 53, 55, 56, 58, 61, 62, 66, 67$$

Q1, Q2, Q3

Find the quartiles (Q1, Q2, and Q3) for this data:

$$22, 23, 24, 25, 26, 26, 27, 27, 30, 39, 40, 40, 43$$

$$44, 46, 49, 49, 49, 50, 51, 53, 55, 56, 58, 61, 62, 66, 67$$



Summary statistics:

Column	Q3	Q1	Min	Max	Median
Sort(Uniform1)	54	27	22	67	45

Percentiles

We also have percentiles, which you have probably used. There are 99 percentiles. Note that the bottom 1% of the data are less than the first percentile, and the top 1% of the data are larger than the 99th percentile. The 99th percentile is the largest percentile.

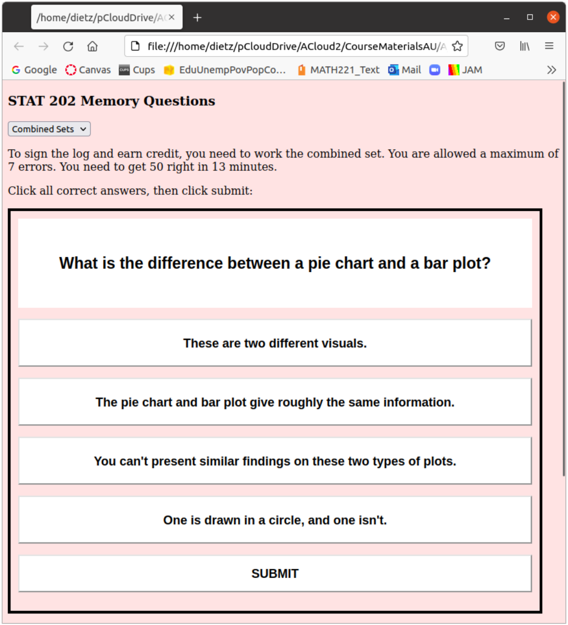
You can have a value “between” two percentiles or even “near” a percentile. However, you can’t have data “in” a percentile. Think of a percentile as a mile marker. A percentile is a value, not a range.

Other “tiles”

Many other “tiles” exist, including the nine deciles which cut data into 10 parts.

MEMORY QUESTIONS

Just three today!


/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

What is the difference between a pie chart and a bar plot?

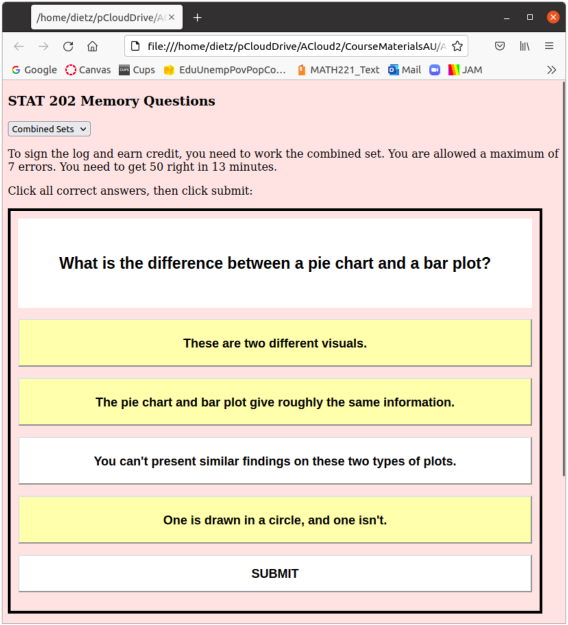
These are two different visuals.

The pie chart and bar plot give roughly the same information.

You can't present similar findings on these two types of plots.

One is drawn in a circle, and one isn't.

SUBMIT


/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM

STAT 202 Memory Questions

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

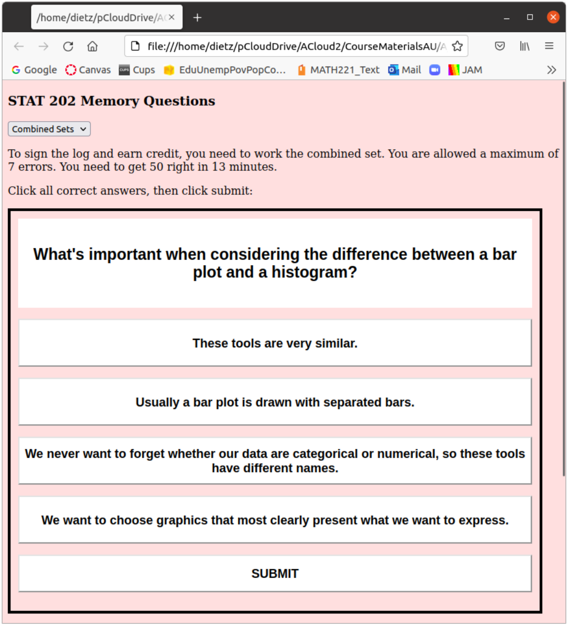
What is the difference between a pie chart and a bar plot?

These are two different visuals.

The pie chart and bar plot give roughly the same information.

You can't present similar findings on these two types of plots.

One is drawn in a circle, and one isn't.



/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
Combined Sets

STAT 202 Memory Questions

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

What's important when considering the difference between a bar plot and a histogram?

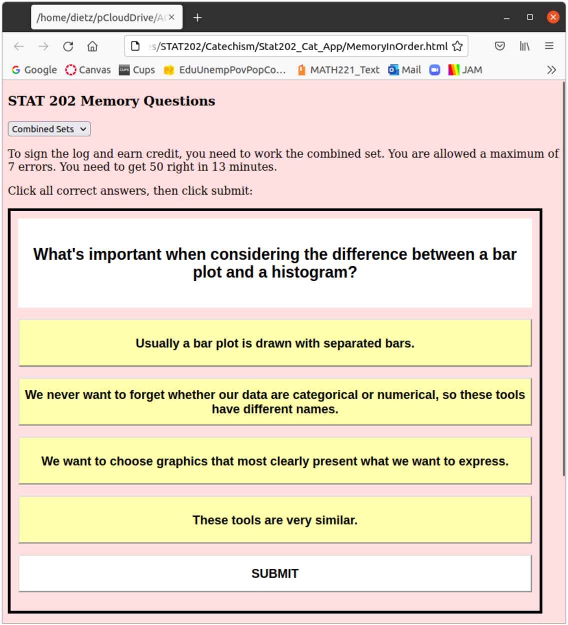
These tools are very similar.

Usually a bar plot is drawn with separated bars.

We never want to forget whether our data are categorical or numerical, so these tools have different names.

We want to choose graphics that most clearly present what we want to express.

SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM

STAT 202 Memory Questions

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

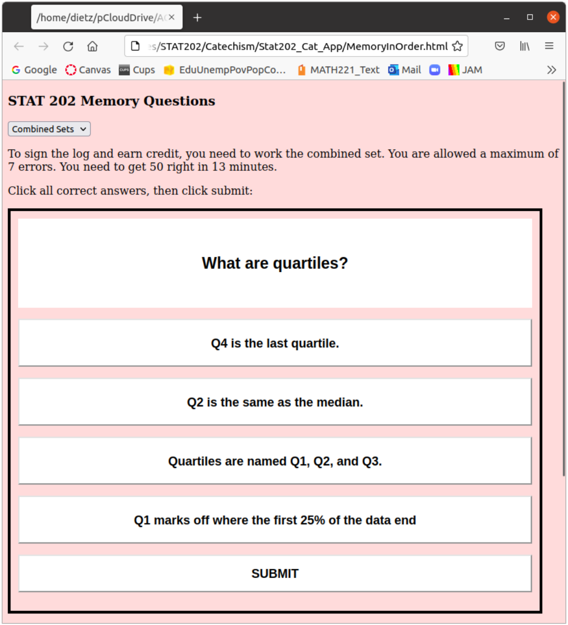
What's important when considering the difference between a bar plot and a histogram?

Usually a bar plot is drawn with separated bars.

We never want to forget whether our data are categorical or numerical, so these tools have different names.

We want to choose graphics that most clearly present what we want to express.

These tools are very similar.



/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

What are quartiles?

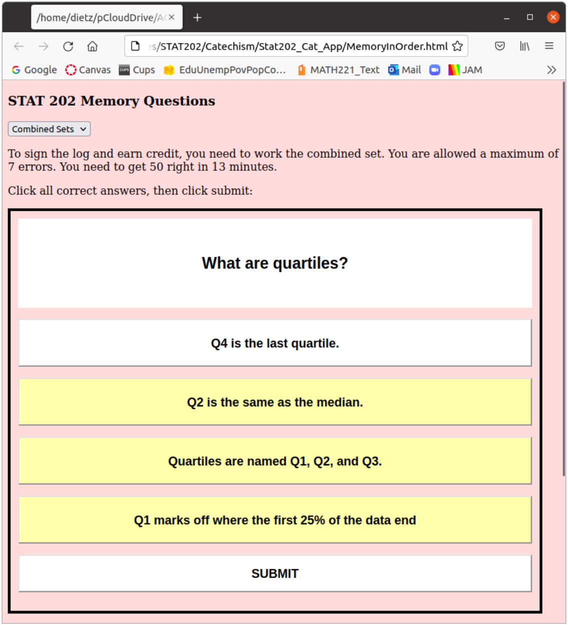
Q4 is the last quartile.

Q2 is the same as the median.

Quartiles are named Q1, Q2, and Q3.

Q1 marks off where the first 25% of the data end

SUBMIT



/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

What are quartiles?

Q4 is the last quartile.

Q2 is the same as the median.

Quartiles are named Q1, Q2, and Q3.

Q1 marks off where the first 25% of the data end

SUBMIT