Class Intervals

To work with continuous numeric data and to represent it in some sort of a graph or a chart, you have to separate the data into class intervals – that is, intervals of equal length. Let us imagine you measured the time it took you to get to Baruch every morning during the past month. The data is presented in the table:

Day	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Time	55	57	40	41	44	45	29	45	44	46	48	35	15	39	41
Day	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Time	44	45	46	48	44	47	28	47	42	73	42	45	44	41	40

If you look at the data, you will see that it ranges from 15 minutes (when you stayed at your friend's house who lives a walking distance from Baruch) to 73 minutes (when the train was stuck on Manhattan Bridge). However, it normally takes you around 45 to get to Baruch. These are the observations you can make by just looking at the data. However, for tables and colorful graphs that your professor wants you to submit, you have to work with the data some more.

First, the data has to be in order – normally, from smallest to largest observation. The result is something called an **ordered array**:

15	28	29	35	39	40	40	41	41	41	42	42	44	44	44
44	44	45	45	45	45	46	46	47	47	48	48	55	57	73

Now we have to establish class intervals. There is no rule set in stone on how to do it, but for best results, you should establish 5 to 7 intervals of equal length. If the data ranges from 15 to 75 minutes and we want 6 intervals, then each interval should be 10 minutes long. How did we get this number?

Interval length = Data range/Desired number of intervals = (73-15)/6 = 9.6 (approximately 10 minutes)

You can also use my favorite method – trial and error. You can say: 15 but less than 25, 25 but less than 35, 35 but less than 45, 45 but less than 55, 55 but less than 65, and 65 but less than 75 – this gives me 6 intervals, just what I need. Whichever method you choose, the resulting intervals have to be equal.

Let's now count how many observations fall into each interval. For example, only one observation – 15 minutes – falls within the first interval and two observations – 28 and 29 minutes – fall within the second interval. Fill in the rest of the table:

Interval	Number of Observations (Relative Frequency)
15 < 25	1
25 < 35	2
35 < 45	14
45 < 55	10
55 < 65	2
65 < 75	1

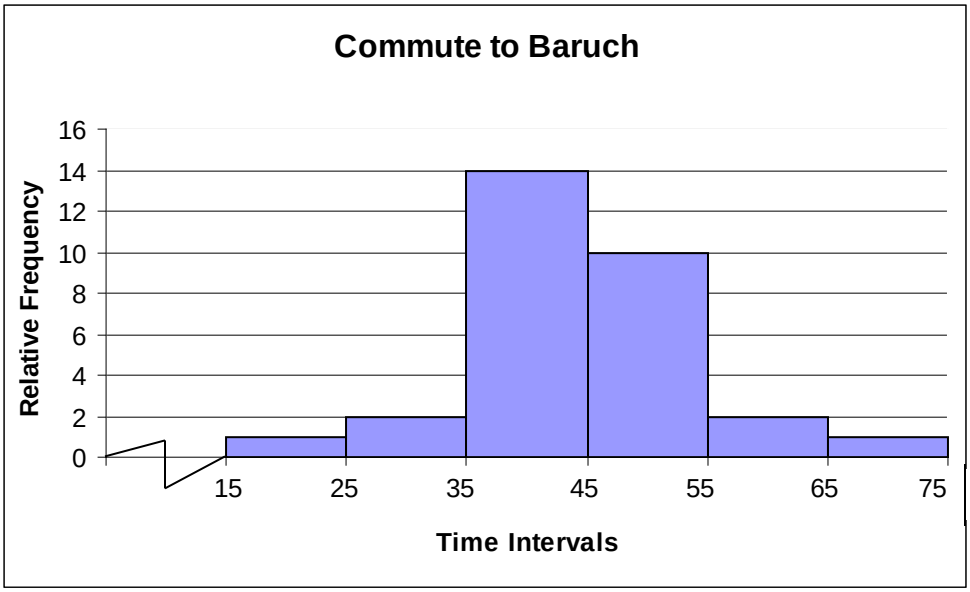
Congratulations, you have just prepared your first relative frequency distribution. With a relatively simple calculation, you can also obtain a percentage distribution. If 1 out of 30 observations falls within the first interval, then approximately 3% of the time it took you between 15 and 25 minutes to get to Baruch. Approximately 7% of the time, your commute took between 25 and 34 minutes. Fill out the rest of the table:

Interval	Relative Frequency	Percentage Frequency
15 < 25	1	3%
25 < 35	2	7%
35 < 45	14	47%
45 < 55	10	33%
55 < 65	2	7%
65 < 75	1	3%

The first and the third columns of this table represent the percentage frequency distribution. Based on this table you can build at least two graphs – a histogram and a polygon.

Histogram

From this point, drawing a histogram becomes easy: plot intervals on horizontal axis and relative frequencies on the vertical axis. Based on the commute data you collected, draw a histogram:



One distinctive feature of histogram is that the intervals we are using to draw it are adjacent, which means that there are no gaps between the vertical bars on the diagram.

How would you interpret this histogram? Practically the same way you interpreted raw data: most of the time it took you between 35 and 45 minutes to get to Baruch.

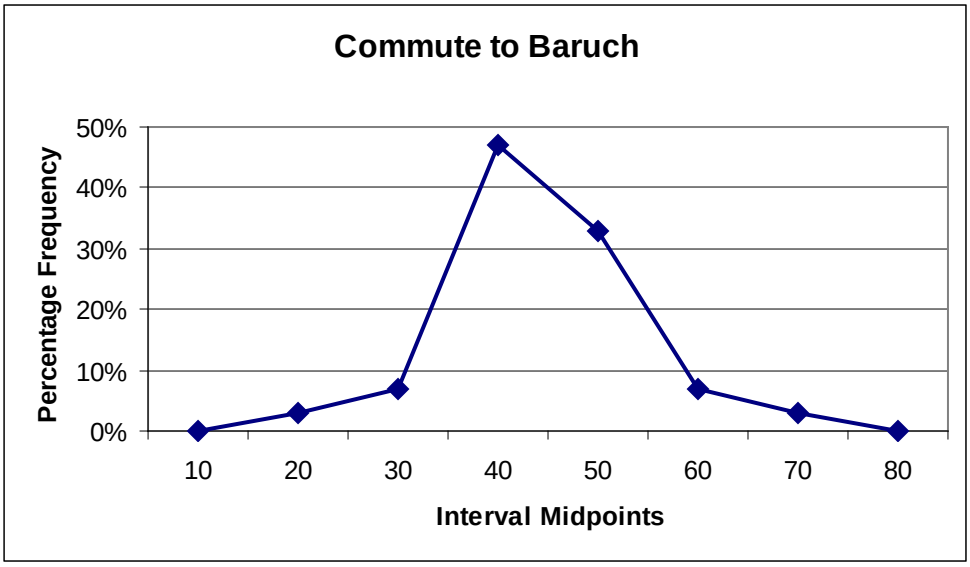
Polygon

Let's build a polygon based on percentage frequency distribution. To do this, you also have to know one trick. You have to establish two additional intervals: one immediately preceding your first interval and one immediately following your last interval. Remember: keep additional intervals the same length as your existing ones.

Notice that *no* observations fall within these new intervals: in the past month it has never taken you less than 14 minutes or more than 75 minutes to commute to Baruch. Creating these additional intervals allows us to maintain a distinctive feature of a polygon: it has to be enclosed, i.e. "touch" the horizontal axis, as you will see on the graph. Another important feature of the polygon is that it is built using the midpoint of each interval. The resulting data table is:

Interval	Midpoints	Percentage Frequency
5 < 15	10	0%
15 < 25	20	3%
25 < 35	30	7%
35 < 45	40	47%
45 < 55	50	33%
55 < 65	60	7%
65 < 75	70	3%
75 < 85	80	0%

Now build the polygon by plotting the midpoints on horizontal axis and percentage frequencies on vertical axis.



The interpretation of the polygon is very similar to that of the histogram. Observe that the commute time of approximately 40 minutes (between 35 and 44) is most frequent – most of the time it has taken you around 40 minutes to get to Baruch. You can also observe that since the graph crosses or "touches" the horizontal axis at 10 and 80 minutes, it has never taken you less than 10 or more than 80 minutes to commute to Baruch.

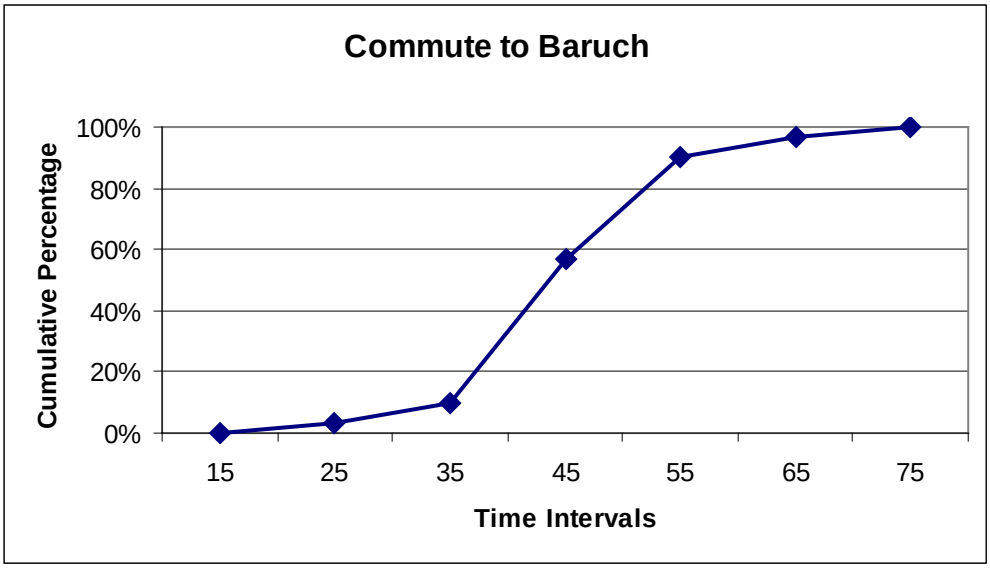
Cumulative Polygon (Ogive)

Now we got to the most interesting part of this topic. The cumulative polygon, or ogive, is called a "less-than" graph and results in a more interesting interpretation than the histogram or polygon. First, as usual, we have to make some preliminary modifications to the frequency distribution table: you have to form the cumulative percentage distribution. To do this, you have to add the percentage for each interval to the sum of percentages of all preceding intervals, as shown in the table:

Upper Boundary	Percentage Frequency	Cumulative Percentage
15	0%	
25	3%	3%
35	7%	10%
45	47%	57%
55	33%	90%
65	7%	97%
75	3%	100%

Observe that instead of using the entire interval, we use its upper boundary. Note also that we need one "imaginary" interval (or its upper boundary) in the lower end of the distribution, because the cumulative polygon has to cross or "touch" the horizontal axis at its lowest point. We do not need the upper interval because cumulative percentage frequency reaches 100% in the upper "real" interval from 65 to 75.

If you plot the interval's upper boundaries on the horizontal axis and the cumulative percentages on the vertical axis, you will get an ogive.



To interpret the ogive, you have to understand what cumulative percentage means. Let's interpret the meaning of the very last (rightmost) node: based on the data you collected, it has *always* taken you less than 75 minutes to get to Baruch. The word "always" means 100% of the time. How would you interpret the very first (leftmost) node on the graph? Based on the data that you collected, your commute to Baruch has *never* taken less than 15 minutes in the past month ("never" means 0% of the time). For a node corresponding to 55 minutes, for example, you can say that 90% of the time you have spent less than 55 minutes to get to Baruch. Now you see why a cumulative polygon is sometimes called a "less-than" graph.