

SUMMARIZING QUANTITATIVE DATA

Histograms and Ogives (Cumulative Frequency Distributions)

Histograms

The first step in making a histogram is to make a frequency distribution or a relative frequency distribution. With qualitative data the classes are obvious. Not so with quantitative data. Consider the following (made up) data – test scores on a statistics exam:

82	52	73	48	70	82
71	79	71	69	91	78
96	88	71	87	65	80
68	81	84	88	82	99
80	97	90	88	78	75

Step 1. We must decide on the **number** of classes. This is not rocket science – there is no right or wrong number of classes. The goal is to generate a nice, easy-to-read summary of the data. Too few classes will result in a poor description. Too many classes will result in a poor description. A good rule of thumb is to take the square root of the total number of observations and round up or down. In our example, we have a total of 30 exam scores. $\sqrt{30} \approx 5.48$, so 5 or 6 classes will probably provide a good description.

Step 2. We must determine the **range** or **spread** of our data. Subtract the smallest exam score from the largest exam score – this is your range: $99 - 48 = 51$.

Step 3. Now comes the tricky part – **setting class width**. There are few rules to follow – just use a little common sense. We need to determine a starting point for our classes and a class width. With a range of 51 and 5 classes, our class width should be $51/5 = 10.2$ or about 10. With a range of 51 and 6 classes, our class width should be $51/6 = 8.5$ or about 8. Let's use 10 for a class width – a nice round number. My lowest exam score is 48 and my first class must include my lowest score. To make the classes easy to read, let's start with 40 – 50, 50 – 60, etc. But, classes must not overlap. What happens if we have an exam score of 60? Which class do we put it in – 50-60 or 60-70? Be creative. One way to avoid this problem and still have classes that are easy to read would be to change the class width to 9: 40 – 49, 50 – 59, 60 – 69....90 – 99. Now there is no overlap and it is obvious which class each score belongs in.

General rules for class width:

1. The lowest class must contain the lowest data value and the highest class must contain the highest data value.

2. The classes must have equal widths. **One exception:** Suppose we had one exam score of 8. To avoid starting the classes with 0-9, etc, ending up with 10 classes (too many), and having a bunch of empty classes in the middle (10-19, 20-29, 30-39), there is something called an open-ended class. I could do this:

less than 50	50-59	60-69	70-79	80-89	90-99
--------------	-------	-------	-------	-------	-------

Technically the lowest class has a width of 49 and the others have a width of 9. This little trick is permitted for the lowest and/or highest classes only!

3. There must be no overlap between the classes and there must be no ambiguity about which class a data value belongs to. If your data has decimals, your classes must also have decimals. If the classes are 40-49, 50-59, etc and one of the grades is 49.7, where does it go? To solve this problem, make the classes 40.0 – 49.9, 50.0 – 59.9, etc.

4. Check that the frequency distribution provides “**a good summary description**” of the data.

Example of too few classes: 40 – 69
70 – 99

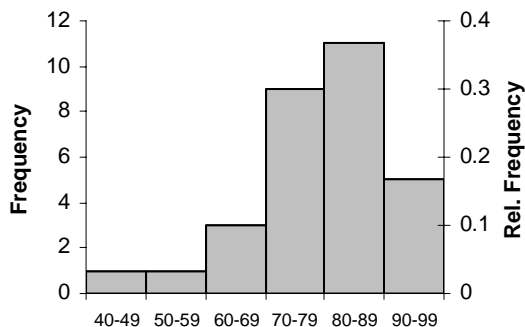
→ doesn't give enough information about the grade spread.

Example of too many classes: 47 – 49
50 – 52
53 – 55 etc.

→ You call this a summary? Why not just list all the grades?

Step 4. Now that we have a frequency distribution we can construct a **histogram**. We could also use a relative frequency distribution if you prefer (just like with qualitative data – divide each frequency by the total number of data values). A histogram is just a bar chart for quantitative data. The only difference is that we do not put spaces between the bars. The horizontal axis shows the class limits, and the vertical axis shows frequency.

Class	Freq.	Rel. Freq.
40-49	1	0.0333
50-59	1	0.0333
60-69	3	0.1000
70-79	9	0.3000
80-89	11	0.3667
90-99	5	0.1667



Ogives

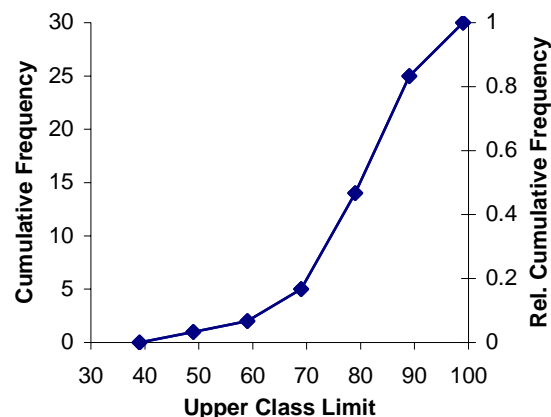
An ogive is a cumulative frequency distribution. The cumulative frequency for the 1st class is always the same as the frequency for that class. To get the cumulative frequency for the second class, **add** the frequencies for the 1st and 2nd classes. To get the cumulative frequency for the 3rd class, **add** the frequencies for the first 3 classes, etc. (i.e., we are “**accumulating**” frequencies like a snowball rolling down a hill). The cumulative frequency for the last class should be equal to the total number of data values or equal to 1 if using relative frequencies.

To draw the ogive:

1. Scale the vertical axis to reflect the cumulative frequencies (or relative cumulative frequencies). Scale the x-axis just as if you were making a histogram.

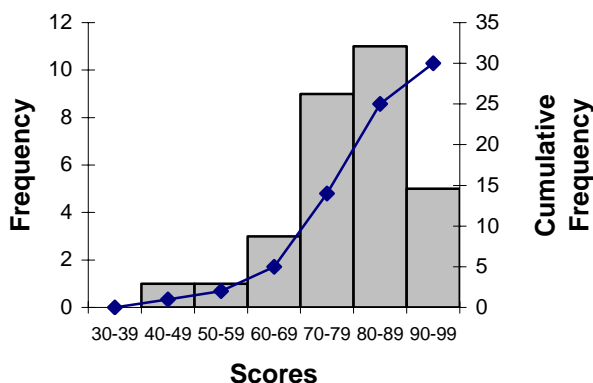
2. Start your ogive by plotting a point directly on the horizontal axis ($y = 0$) just to the left of the lower limit of the 1st class. For each class, plot a point with the x-coordinate equal to the upper class limit and the y-coordinate equal to the cumulative frequency for that class. Connect the points with straight lines. Cumulative frequencies keep getting bigger, so the ogive graph should always be “climbing”.

Class	Freq.	Cum. Freq.	Rel. Freq.	Rel. Cum. Freq.
40-49	1	1	0.03333	0.0333
50-59	1	2	0.03333	0.0666
60-69	3	5	0.1000	0.1666
70-79	9	14	0.3000	0.4666
80-89	11	25	0.3667	0.8333
90-99	5	30	0.1667	1.0000



Ogives are useful in situations where the people looking at your summary may only be interested in the percentage of data that falls above or below a certain cut-off point. For example, suppose you were only interested in how many students flunked this exam. You would locate 60 on the horizontal axis, go up and over and find the frequency of 2 (or 6.7%).

Ogives are also sometimes combined with regular frequency histograms, with the cumulative scale shown on a secondary y-axis:



Stem and Leaf Plot

Stem and leaf plots are very useful. They: 1) are easy to make; 2) are an easy way to order your data from smallest to largest (which you have to do to calculate a median, percentile, quartile, or hinge); and 3) provide the same information as a histogram. Let’s look at these scores again.

82	52	73	48	70	82
71	79	71	69	91	78
96	88	71	87	65	80
68	81	84	88	82	99
80	97	90	88	78	75

Step 1. Find the smallest and largest values and decide on the “stems”. The choice of stems depends on the numbers you’re working with. A “stem” is sort of like a “class” except it’s a single number.

4		Smallest value: 48. Largest value: 99.
5		My choice of stems are the <u>10’s digits</u> of all the numbers between 48
6		and 99.
7		
8		
9		

Step 2. Once you draw your stems like this, then you can go through your data one value at a time

4		8	and attach the <u>second</u> (“ones”) digit of each exam score to the
5		2	appropriate stem. The second digits are called the “leaves”.
6		895	
7		193110885	
8		20814788220	
9		67019	

4		8	Once you’ve attached all your “leaves”, redraw the stem and leaf plot
5		2	and put the “leaves” on each “stem” in order from smallest to largest.
6		589	
7		011135889	
8		00122247888	
9		01679	

This is not hard once you get the hang of it and you can now list the test scores in order from smallest to largest. If you turn the paper sideways, your stem-and-leaf becomes a kind-of histogram. The stem and leaf has one big advantage over the histogram. If you give someone a histogram, they cannot recreate the actual data that you built the histogram from. With the stem-and-leaf, they can have their cake and eat it too. They can look at it like a histogram if they just want a quick summary, or if they need to, they can actually regenerate all of the exam scores.

A **five number summary** is simply 5 numbers written horizontally and separated by commas, The five numbers are:

smallest data value, first quartile (Q1), median, third quartile (Q3), largest data value

A five number summary provides a measure of central location (median) and a “feel” for the spread or variability of the data. About 25% of the data values fall between adjacent numbers in the five number summary.

This is the 5 number summary for our exam scores: **48, 71, 80, 88, 99.**

The median score was 80. About 25% of the scores were between 48 and 71, about 25% between 71 and 80, about 25% between 80 and 88, and about 25% between 88 and 99.