Working with one variable

When we want to say something about a single column in our data

This data for this exercise can be found at:

http://bit.ly/1P8ti6T

How do I know I am using one variable?

- Summarizing one variable
  - Average
  - Median
  - Sum
  - Maximum
  - Minimum
- Looking for an outlier in a distribution

How do I know which summary statistic(s) to use?

Well, it depends on the shape of the distribution.

But first some vocab:

Mean/Average - This is the sum of all values divided by the number of observations.

Median - If we rank everyone in the data by value, this is the value associated with the person (or people) in the middle.

**Mode** - This is the value that occurs most frequently in the data.

What is a distribution?

In general terms, it reflects how observations are spread out across the range of our data.

The range is all the values between the minimum and maximum values in our column.

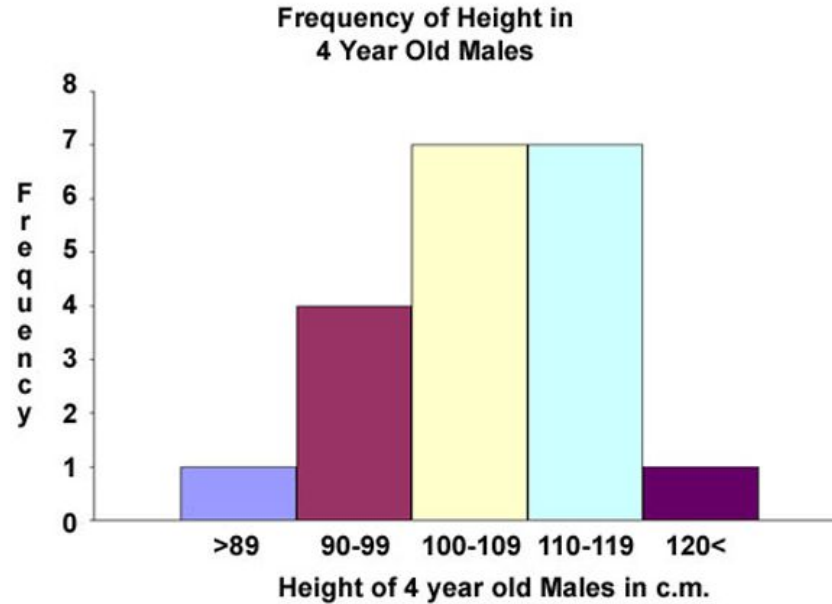A good way to see a distribution is to make a histogram.

What is a histogram?

Number of observations or people with this value

All possible values divided into "bins"
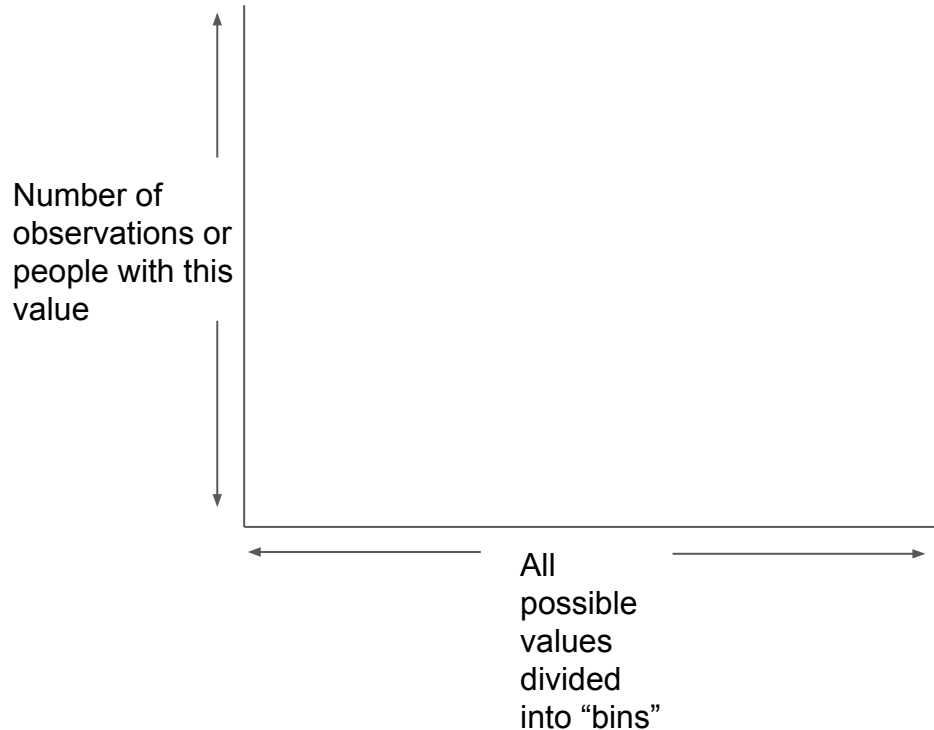
A histogram tells us how common each range in the data is. This is called the 'distribution' of the data.
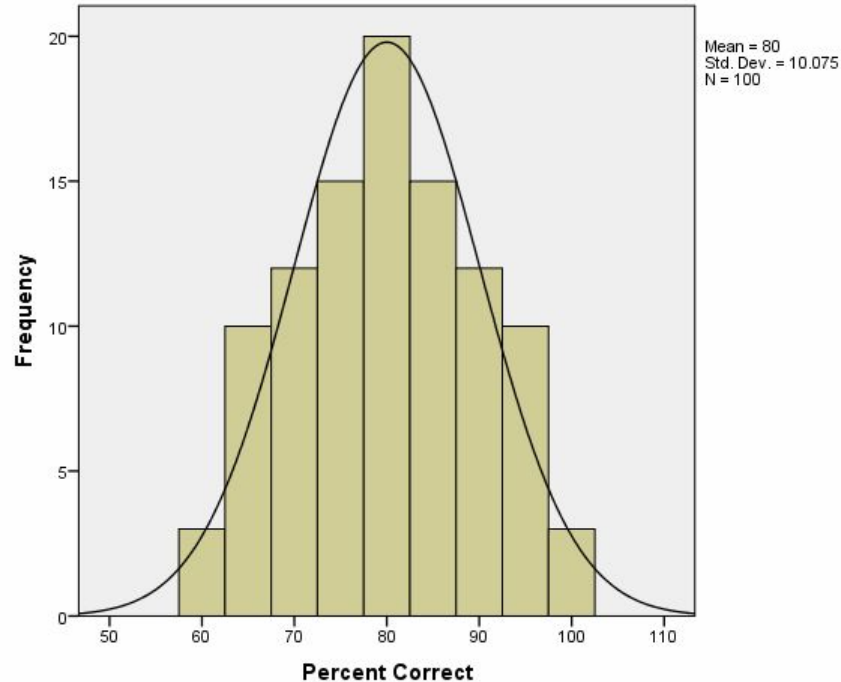
What is a histogram?

Here's an example histogram.

Frequency of Height in 4 Year Old Males

Here are some common distribution shapes

Normal

Average is an appropriate summary for this column.

Mean = 80
Std. Dev. = 10.075
N = 100

Here are some common distribution shapes

Normal with skew

The long right or left tail can move the mean to a value that isn't typical.
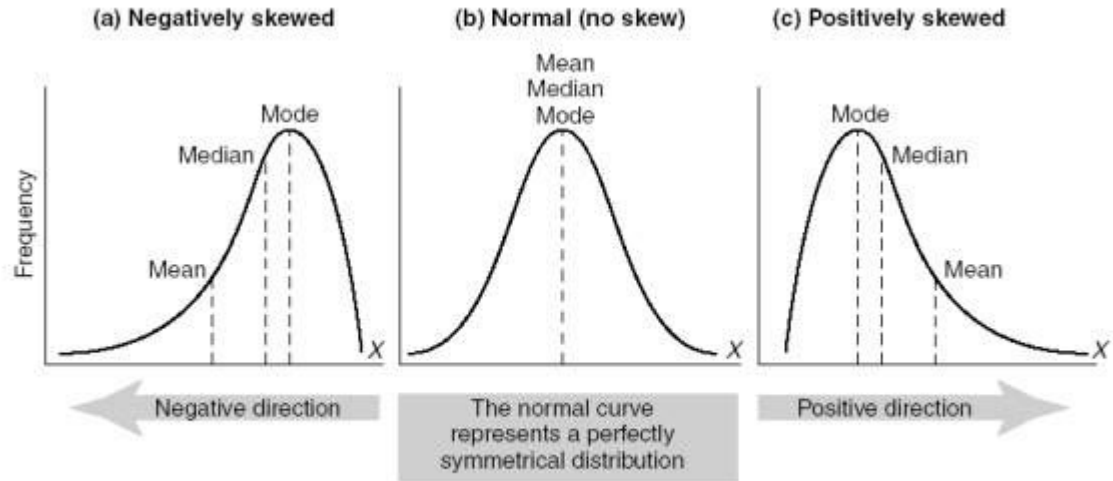
You should consider median or mode here.

| (a) Negatively skewed | (b) Normal (no skew) | (c) Positively skewed |

Negative direction

The normal curve represents a perfectly symmetrical distribution

Positive direction

FIGURE 15.6 Examples of normal and skewed distributions

Here are some common distribution shapes

Uniform

The mean here is 5.
Does that accurately
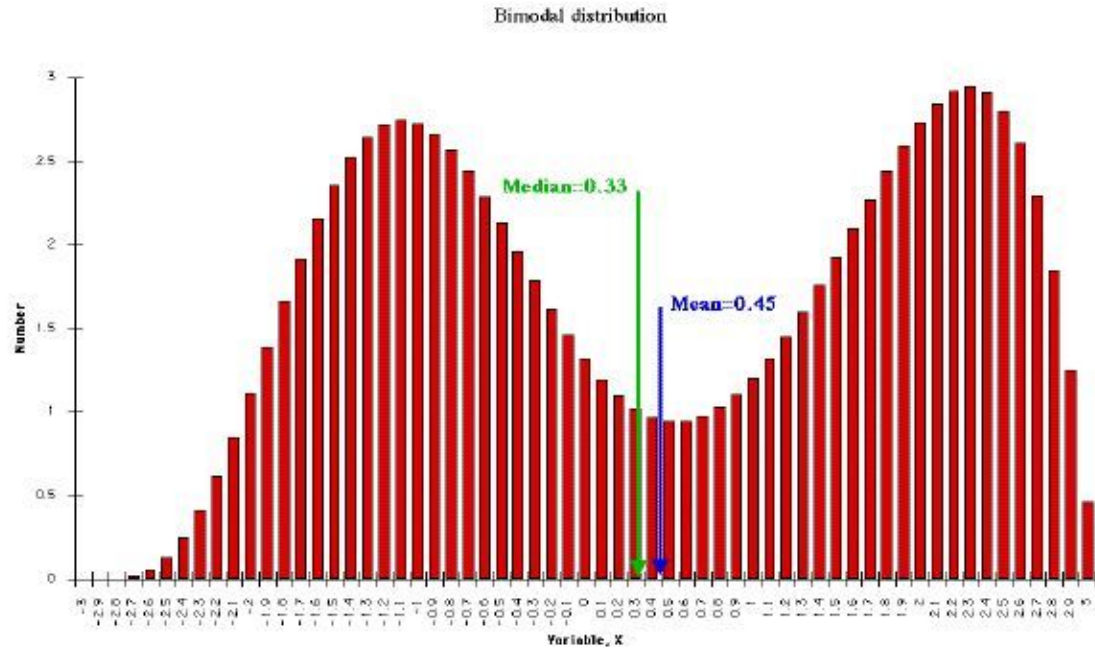reflect reality?

Symmetric, Uniform, Distribution
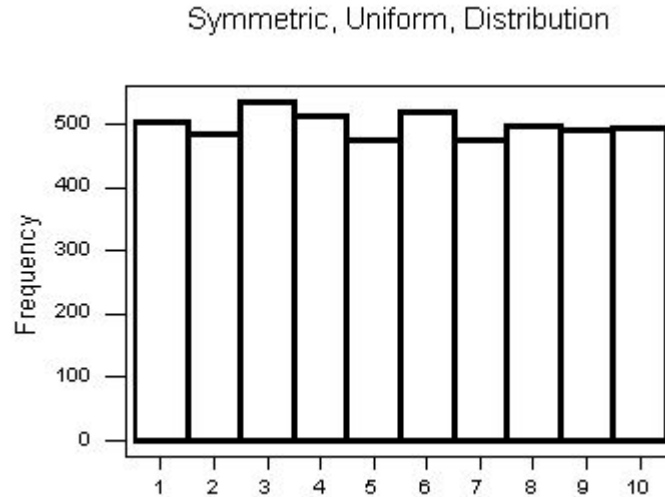
Here are some common distribution shapes

Bimodal

We see this often when we are talking about poverty or race.

Do the mean and median here accurately reflect reality?

Bimodal distribution

Let's go to the data

The data for this exercise is school-level. We have average test scores for each school, as well as some characteristics of the school.

Take a look at the data and make sure you understand what's going on.

Making a histogram

Fortunately google sheets makes it pretty easy to make a histogram.

- First let's click on 'Column O' to select the column.
- Then Insert -> Chart
- You should see something like this:

Making a histogram

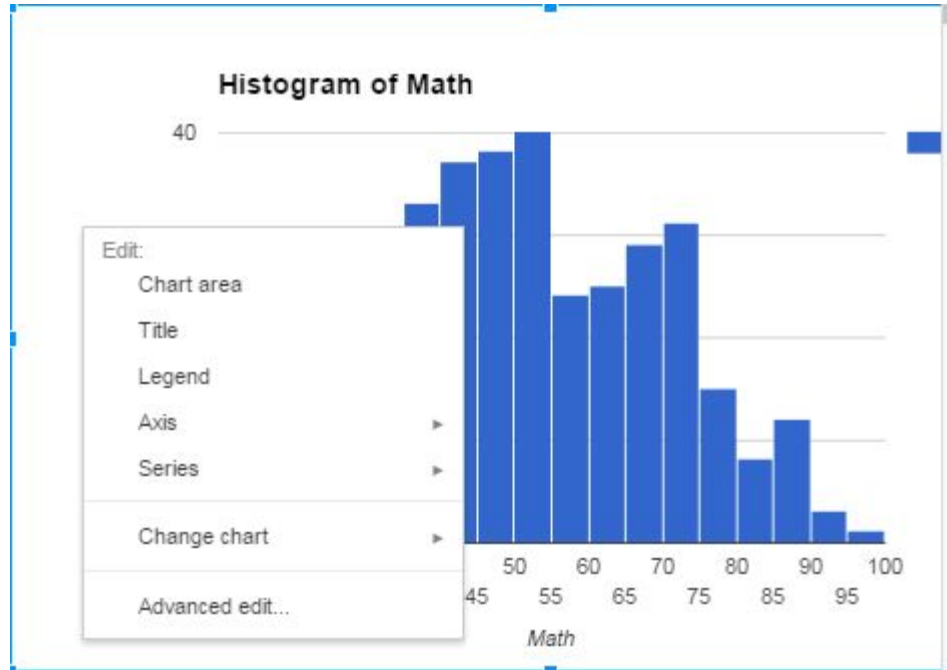Choose the histogram option, and you should get this:

Roughly, what distribution type does this approximate?
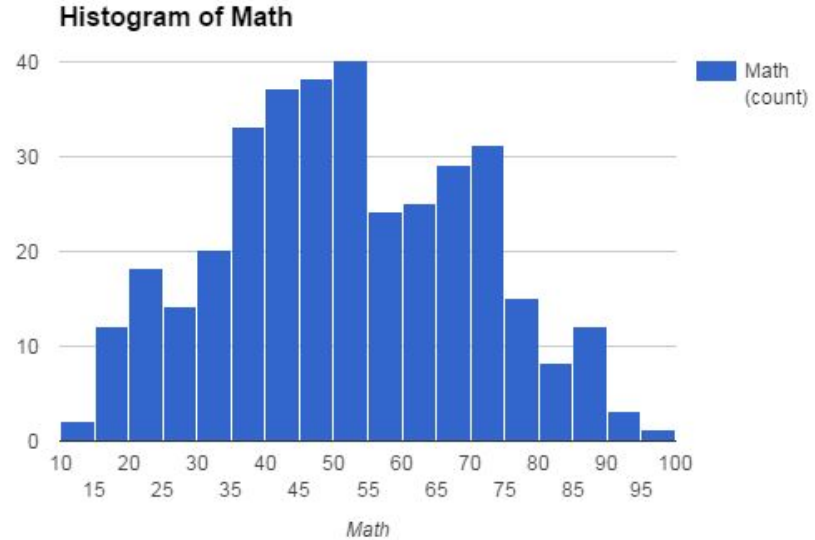
Making a histogram

This is looking okay, but there are a few things we can do to make it look a bit better.
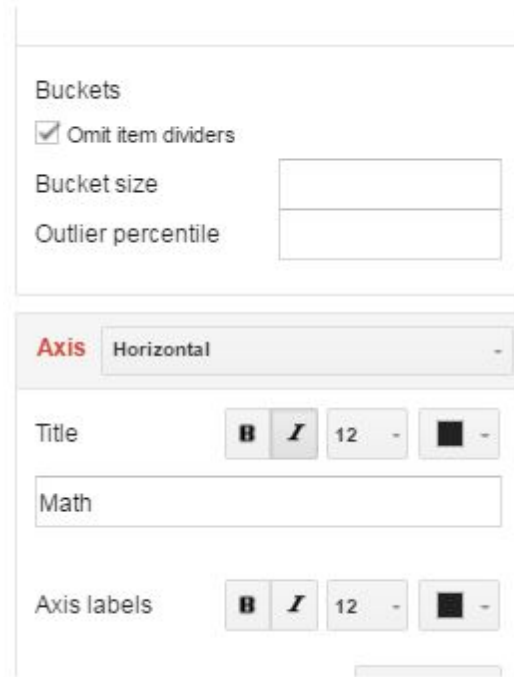
- Right click on the chart, and then choose Advanced Edit

Making a histogram

What happens if we change things like bucket size and the legend placement?

Play with this for a moment.
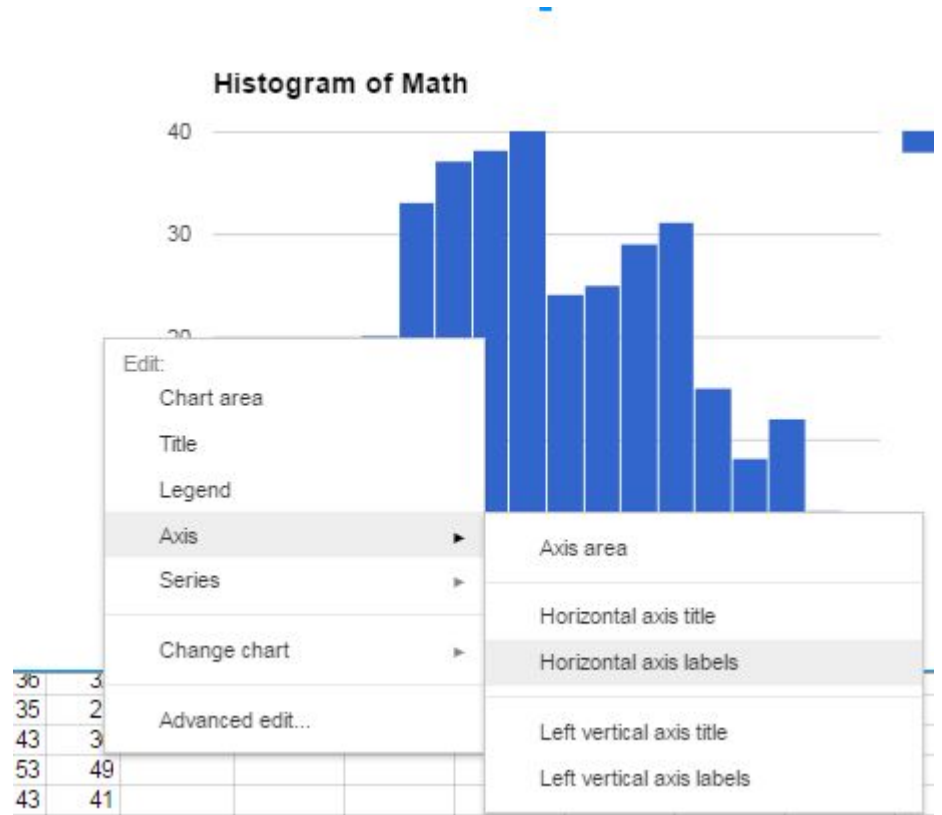
Making a histogram

We can also make the axis labels look a bit better/easier to read.

- Right click again and do Axis -> Horizontal axis labels
- Experiment with the options - in particular rotating the labels can help us here.

Making a histogram

Ta-dah!
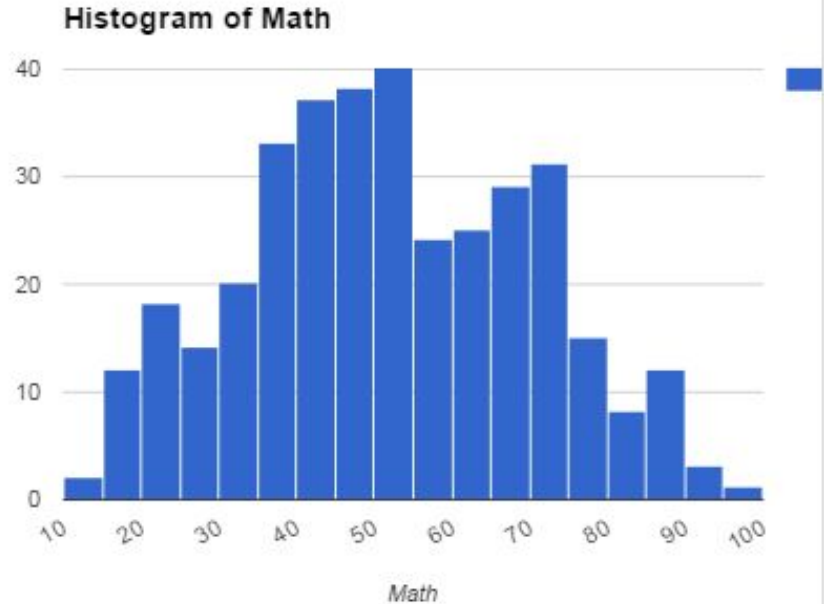
- Right click again and do Axis -> Horizontal axis labels
- Experiment with the options - in particular rotating the labels can help us here.

Histogram of Math

Making a histogram

As an exercise, let's make a histogram of the 'Poverty' variable. What do you see?