Single variable distributions, histograms

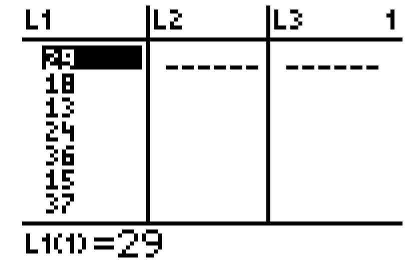
In this section we will be working with single lists of numbers that would usually represent a sample of values of a characteristic of a population. This list could be the measured heights of all the students in your class or in all RTT classes, for example. For this list we will find various statistical descriptive parameters such as the average (mean), quartiles, and measures of dispersion like variance or standard deviation. From these calculations made by the calculator (usually and thankfully), we can make inferences about a larger population (like all 11th graders in Arkansas or like all 11th graders in the world, or like all 11th graders in the universe). However, the only scientifically valid inferences would be about all 11th graders at ASMS. This goes back to the discussion on design research and common sense. If you measure and record the heights of all ASMS 11th graders, then you have the supreme sample of the entire population. For a given list/sample, we can also determine whether or not the data is normally distributed and make an inference about the distribution of the population from which the data was sampled. Most characteristics of populations are normally distributed.

To clarify, populations like all humans, American consumers, mosquito legs, and coke cans are quite large. However, a small number, even one, can be defined as a population. In those cases, of course, inferencing from samples would be useless.

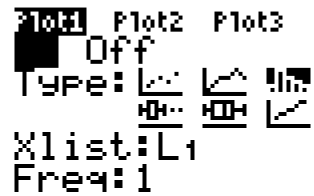
It is not likely that you will be analyzing a single list in your ISEF project, but we must be able to work with one list to understand the statistics of two or more lists.

Let’s begin by making a list of 100 random integers between 10 and 50 and placing these numbers in list 1, otherwise known as L1. Do **MODE**-*Float*-0-**Enter**. Do **MATH**-*PRB*-5 to get “randInt(“ to the screen. Key in ‘50,10,100)’ and then **STO-2nd-L1-Enter**. Wait until the first few integers pop to the home screen behind a brace.	randInt(50,10,10 0)→L_1 {48 47 16 31 26…

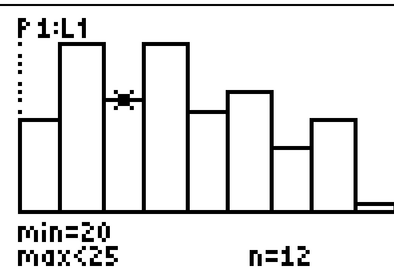
Now **STAT-**1 to get the screen shown with your integers. Your list will be different but should be numbers between 10 and 50



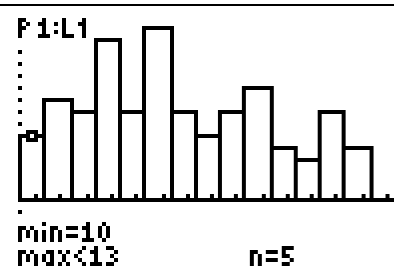
Now make a histogram of these numbers with **2nd-STATPLOT** (top row)-1. Turn the plot *On* with **ENTER**, cursor to the histogram as shown and press **ENTER** to enable the histogram, make sure *Xlist* is L1 and *Freq* is 1.



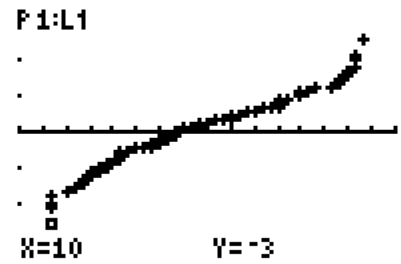
Now **ZOOM-**9-**TRACE** to get something like the screen shown. Cursor across the various intervals to see the number of occurrences in each. Note that you may remember the 9 better if you notice that it is labeled "ZoomStat" If a menu sequence number is off the screen cursor down to it at least once!



Do **WINDOW** (top row). You can change the interval size by changing *Xscl*. It is now 5 for the above graph (yours could be different). I'm changing it to 3. Since there will be fewer occurrences in smaller intervals, I will change *Ymax* from 21.06 to 15 as a guess. After your similar changes, don't use zoom again as the window variables will go back to the calculator's choice. Rather, do **TRACE** or **GRAPH-TRACE** if you don't like shortcuts, to get a screen similar to that shown. Exact is a subset of similar.



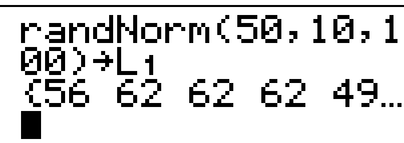
Do a “normal probability plot” of this list of random integers. **2nd-STATPLOT-**1 and select the last of the six plots shown. Make sure that *Data List* is L1. Use *Data Axis* of X. You could do as well with Y, but use X for now. Choose a *Mark* of “+”. **ZOOM-**9 and notice that the plot turns down and up on the ends. If this plot is a straight line, the data is indicated to be normally distributed. Since there is no associated probability calculated, you could not use this to test for normal distribution should that be your hypothesis for science research. (Unlikely for ISEF)



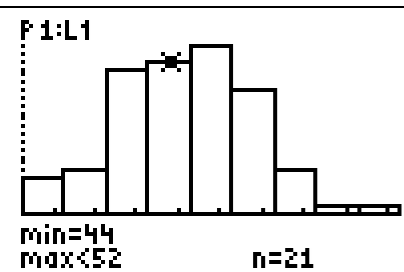
Your numbers in L1 can be sorted with **2nd-LIST-***OPS*-1 to bring “SortA(“ to the home screen. Key in “L1)” and **ENTER** to sort list 1 in ascending order. Now the list in the stat editor (**STAT-**1) can be more easily compared to the histogram. Had the sorting been done initially, there would have been of course no effect on the histogram.

Now let’s repeat the above exercises for a list of <u>normally</u> distributed integers, say 100 of them, with an average or mean of 50 and a “standard deviation” of 10. First we will write over list 1 (or you could use another list like L2).

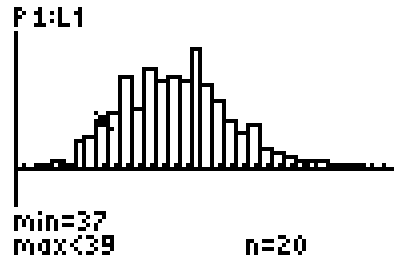
MATH-PRB-6 to get “randNorm(“ to the home screen. Finish it with “50,10,100)” –**STO-2nd-L1-Enter** to fill L1 with the random numbers then wait until the first several numbers appear with a brace. You can view the numbers in the stat editor and/or sort them if you wish.



Now make a histogram for L1 per the instructions above when random numbers were in L1. Notice that there are more occurrences clumped around the mean of 50 than further from the mean.

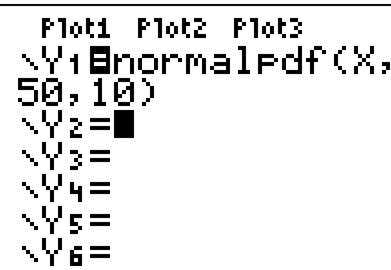


Now do a histogram of 500 normally distributed numbers by repeating the above exercise except for 500 instead of 100 generated numbers. The generation takes about a minute. In **WINDOW**, change *Xscl* to 2 and *Ymax* to 70. If you get something quite similar to this, there's a good chance you're going to make it in that you successfully used details from previous exercises.



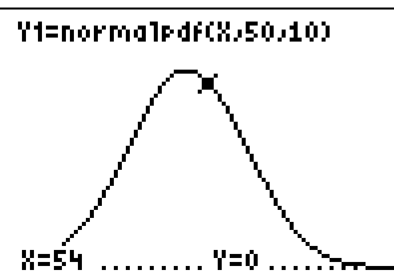
As the number of generated numbers increases to approach the size of most populations (large to just short of infinite, so you're limited on the calculator) and the window variable corresponding to Xscl is kept small, the histogram would approach a smooth "normal" curve. Let's see what this curve would look like. Discomfort warning-for those who must understand something completely before they start working with it. Maybe you could have used this warning earlier on this page. Anyway, just jump in. You may find a new way of learning which is very often the only way.

Y= (top row). Cursor to *Plot1* and turn it off with **ENTER**. Cursor to *Y1*=. Now **2nd-DIST**-1. Now finish the expression with "(**X,T,?,n**),50,10)" and **ENTER**. The key for X is in parenthesis in an attempt to avoid confusion. Did it work? 50 is the mean and 10 is a thing called the standard deviation, a measure of dispersion.

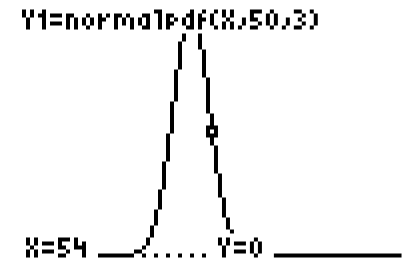


This Y= window will be often used to enter functions to be graphed and even more often for controlling which statistical plots and functions you want graphed together. Here we don't want Plot1 to be graphed with the normal probability density function.

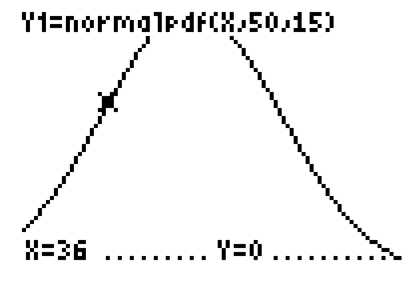
WINDOW. Set *Xmin* to 15, *Xmax* to 85, and *Ymax* to .05. Now Zoom-0 to get a plot of a the normal curve with mean of 50 and standard deviation of 10. Don't worry about the Y=0 indication for all values of X. This is just to give a picture for discussing dispersion and we will use "normalpdf(" as was intended later.



Now go back to **Y=** and change the 10 in Y1 to 3 and continue as before to see the following shape. You will keep the same Xmin and Ymin as in the previous graph. Again use **ZOOM**-0



One more and we can define and philosophize for a while. You could even take a break if you promise to come back sometime. Plot the normal curve as above except with a standard deviation of 15 instead of 3. As a minor note notice that you can use **2nd-INS** to avoid rewriting the right hand parenthesis. In this case, it wasn't very timesaving, but you can see how it could be.



We've just plotted the "normal" curve for three different standard deviations (10, 3, and 15) and the same mean or average of 50.

So what is a standard deviation? The standard deviation is the square root of the variance. See! That was easy. Oops, there's someone who is not intimately familiar with "variance"? Ok. For a population the variance is the average of the squares of the differences (or distances) between the numbers and the mean.

Population variance=$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2$ where N is the number of entities in the population.

The mean of a population is designated by the Greek letter µ. The standard deviation is designated by s.

Examine the three graphs of normally distributed populations. Notice that a smaller standard deviation causes the distribution to be clumped nearer around the mean.

Refer to the histograms we made with random numbers. Suppose lines connected the midpoints of the tops of the interval bars. These are the points marked when you cursor across with TRACE. These lines form a "frequency polygon". Lines connecting the corners form an "orgive". If you changed your mode to allow decimal numbers, generated more and more numbers, and kept decreasing the interval size *Xscl*, the top of your histogram (your frequency polygon or orgive) would begin to look like these curves that you just generated above. Hence the histogram of a finite number of values approaches continuous distribution functions. What about “kurtosis”? Is it a bad skin disease? Kurtosis refers to the shape of the three normal or “bell” curves you just made. It is relative. Relative to the first curve, the second is “leptokurtic”. Relative to the first curve and even more so to the second, the third curve is “platykurtic”.

Now concerning the above paragraph- you could probably do without it. Don’t you wish I had told you sooner? But hang on to the previous paragraph and we will continue.

For a very large typical population, what are the chances that an exact mean and standard deviation have been calculated? Not good. In fact, we wouldn’t need to take samples and make inferences from them if we had such data. However, in some problems you may work, a μ or s is assumed known or even assumed. But in the real world or ISEF environment, beware.

Before we do more calculator work with a single sample list and then finally infer some info on the population from which the sample was drawn and before we forget all about histograms, let’s take a short sidetrack to do something that could sidetrack *you* on a little test sometime. At some time in your RTT career, some cruel and deranged person may ask you to use the histogram feature on your TI83plus to make a bar chart. Suppose we have the following table of data:

Insipidity of Clandestine Tibetan Monks	408	300	718	214	101	1020
Year	1400	1500	1600	1700	1800	1900

STAT-1. Place the insipidity values in L1 and the corresponding years in L2. To clear L1 of previous values, cursor to “L1”, press **CLEAR**, and cursor down to the first entry. **2nd-STATPLOT**-1. Turn the plot On, highlight and select the histogram, make the *Xlist* “L1” (use 2nd-L1) and *Freq* “L2”. **WINDOW**. Make *Xmin*=1400, *Xmax*=2000, *Xscl*=100, *Ymin*=-300, *Ymax*=1200. If you want the label, use **2nd-DRAW**-0 and use the **ALPHA** key.