

## One variable samples and what to do with them.

I don't need to hear from you what to do with them.

This summer one of our RTT students visited the planet Slingoff in the fifth quadrant, randomly selected 15 Cardassian males, measured the widths of their necks in inches, and recorded this random sample in list 1 of a TI83plus. Make a list called "NECKS" by **STAT-1**, scroll to *L6* (or the last list name), place the cursor on the list name, and cursor to the right. You are in ALPHA locked mode (you can do this anytime with **2<sup>nd</sup>-ALPHA** and get out of it by simply pressing the **ALPHA** key). Type in "NECKS" and **ENTER**. Place the following neck measurements in your list: 17.06, 18.83, 13.15, 17.30, 14.29, 12.26, 10.48, 15.06, 18.94, 15.74, 13.88, 16.14, 12.31, 13.17, and 21.14. Now let's generate some one variable stat parameters for this list. Don't worry if your numbers occasionally differ in the 2<sup>nd</sup> or 3<sup>rd</sup> decimal place. The numbers shown are rounded from the numbers that actually generated these graphics:

<p><b>Stat-Calc-1-ENTER.</b> "1-Var Stats" appears on the home screen. Finish by <b>2<sup>nd</sup>-LIST</b>-(whatever number corresponds to "NECKS"). Now "1-Var Stats LNECKS" is on the home page. Press <b>ENTER</b> to get the following screen if your mode was set for 3 decimals under Float.</p>	<pre>1-Var Stats x̄=15.317 Σx=229.752 Σx²=3640.575 Sx=2.946 σx=2.846 ↓n=15.000</pre>
---	--

<p>The rest of that screen is as shown: If some of the 15 entries occurred more than once, the number of occurrences of each could be placed in another corresponding list, say L2. Then "L2" would be keyed in after "NECKS" on the home screen. This second list is called the "Freq" or frequency list.</p>	<pre>1-Var Stats ↑n=15.000 minX=10.475 Q1=13.153 Med=15.061 Q3=17.301 maxX=21.143</pre>
--	---

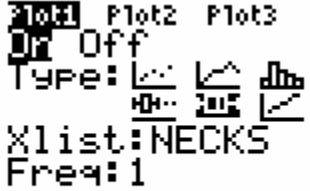
So what about these numbers? The first one, called "X-bar", is simply the mean or arithmetic average of the numbers of the list. This is followed by the sum of all the list numbers and the sum of the squares of all the list numbers. Next,  $\sigma x$  is obtained with the formula:

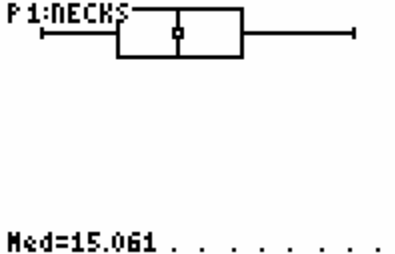
$$(\sigma x)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad \text{Substitute Xbar for } \mu.$$

Whereas the variance of a population is found by dividing the sum of the differences squared by  $n$ , the population size, the variance of a sample is found by dividing by  $n-1$ ,  $n$  being the number of numbers in the sample. So  $Sx$  differs from  $\sigma x$  by a factor of the square root of  $(n/n-1)$ . Try it. Multiply 2.846 by the square

root of  $(15/14)$  to get an  $S_x$  of 2.946. This is an estimate of the population standard deviation. Why  $n-1$ ? Because that is the number of “degrees of freedom” left after the first number in the sample is selected. You can explain it that way if anyone should ask. Really understand it? Don’t worry about it. Just remember for now how to get  $S_x$  and that it is the best estimate of the population’s (from which this random sample was taken) standard deviation. It is indeed called the “sample standard deviation”. The other is assuming that your list is a population itself. The  $\min X$  and  $\max X$  are the min and max numbers in the sample. The difference between them is called the “range”.  $Q_1$ , called the “first quartile”, is the number such that  $\frac{1}{4}$  of all the numbers in the sample are less than  $Q_1$ . Sometimes this set of numbers is considered the “first quartile” but for now consider it  $Q_1$ . The Med is actually  $Q_2$ , the number such that  $\frac{1}{2}$  the numbers are less than  $Q_2$  and  $\frac{1}{2}$  are greater than  $Q_2$ . Note that for an odd number of numbers, this definition would need to be replaced with “the middle number in a set of numbers”.  $Q_3$  is such that  $\frac{3}{4}$  of the numbers are greater than  $Q_3$ . These are the first, second, and third “quartiles” of a list of numbers or sample. Run 1 variable stats on lists  $\{1,2,3,4\}$  and  $\{1,2,3,4,5\}$  for clarification.

You can obtain the min, max, and quartile points in picture form with a box plot or modified box plot. Two or three box plots can be drawn on the same screen to visually compare two or more lists. Make a regular box plot for “NECKS”.

<p><b>Y=.</b> Turn off all plots except <i>Plot1</i> and all functions with the cursor and <b>ENTER</b> keys. 2<sup>nd</sup>-<b>STATPLOT</b>-1. Turn on the plot, select “NECKS” and the normal box plot as shown.</p>	
--	--

<p><b>ZOOM-9-TRACE.</b> Do the quartile points including <math>Q_2</math> (the Median) match the stats you ran previously? Cursor over horizontally to the other points to verify.</p>	 <p>Med=15.061 . . . . .</p>
--	--

The modified box plot is for plotting “outliers” (points more than 1.5 times the distance between  $Q_3$  and  $Q_2$  that are beyond  $Q_3$  or short of  $Q_1$ . These outliers are plotted as marks and not as part of the whisker as in the normal box plot. You can, of course, plot more than one box at a time. Simply enable the plots under **Y=.**

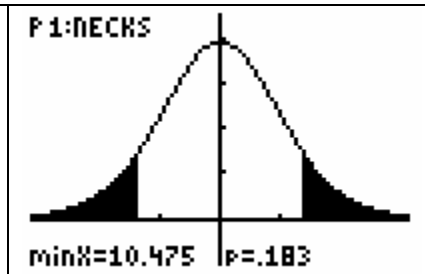
So now what do we do with all this information, including the sample list itself? If this is a **proper** sample of the population (in this case the neck widths of all Cardassians) proper meaning **random** and from the **right population**, we make inferences about the statistical parameters of that population, like its mean and standard deviation (dispersion). There are indeed other defined parameters of a population, but these two are the most important, widely used, and enough for us.

There are two cited ways of connection between samples and their populations. One is to develop a hypothesis about the population parameter and then to test it with sample data. The other is to take sample data and make inferences about the associated population parameters. These two philosophies are shown in the form of the different statistical tests in the TI83plus, like the T and Z tests (hypothesis) vs. the T and Z interval tests (inference). Actually, it's all just about samples containing information about the populations from which they are drawn.

First we will use the T test and T interval test. One problem with Z tests is that the standard deviation of the population must be known. If you're hypothesizing or inferencing about the mean of your population, what are the chances of knowing the standard deviation? When you attempt to use a Z test, you'll be faced with inputting the population  $\sigma$ . At this point, the user will say, "I think I need another test."

Let's make a hypothesis about the mean of all Cardassian neck widths. The Xbar of your sample is 15.317. Suppose though for now that you have previous information or observations which compelled you to say 14.25 inches is the mean for true Cardassians and that these measurements from the fifth quadrant are of some other species. So your hypothesis is that the mean of the sample population is not equal to the mean of the true Cardassian population.

**Y=.** (Top left) Turn off all functions and plots.  
**STAT-TESTS-2.** Select *Data*. Enter a  $\mu_0$  of 14.25.  
 In *List:* paste "NECKS" by **2<sup>nd</sup>-LIST**-(number of the NECKS list). Select the hypothesis of " $\neq\mu_0$ "  
 Select *Draw* and try *Calculate* later for comparison.



The  $p = .183$  is the area under the shaded portions of the distribution curve. The total area under the curve is 1. This 0.183 is a probability. It is **not** the probability associated with " $\mu \neq \mu_0$ ". Don't let a statistics professor hear you say this, but the easiest way to interpret p is that it is the probability associated with the null hypothesis, in this case, that " $\mu = \mu_0$ ". Since this probability is not less than 0.05 or 5%, the null cannot be rejected. Your hypothesis is rejected under the 95% rule

burden of proof. Even the loosest rule of 90% won't help you. Even though the probability appears in your favor, it's not favorable enough to accept your hypothesis. That's science and stats biz. This shaded area under the curve is sometimes termed the "rejection region". If the region is small enough like less than 0.05, you can reject the null and accept your hypothesis. This is a usage dependent term and may be defined to be just the opposite of this by some. Use your judgment if it's on a test.

Before leaving the T test, change your hypothesis to the statement that the population of the sample has necks of mean width less than 14.25 and use Calculate instead of Draw to get a p of 0.909 that the measured mean is greater than 14.25 (the null). Now repeat selecting a hypothesis of greater than 14.25 to get a null probability (remember to not talk carelessly like this around a real statistics professor) of 0.091 that the measured population has a mean neck of less than 14.25. Do you see a relationship among the three p's? How about that 2 times .091(+) is the .183? Or that 1-.909 is .091? Can you select an "=" hypothesis with this test? No, but you could deviously use the null and think of it as your hypothesis.

Now let's take purportedly a different approach and work with the TInterval test. This is a lot easier than the previous exercise.

<p><b>STAT-TESTS-8.</b> Select <i>Data</i>. Paste NECKS for <i>List</i> with <b>2<sup>nd</sup>-List-NECKS</b> as before. Enter a <i>C-Level</i> of 0.95. Cursor to <i>Calculate</i> and <b>ENTER</b>.</p>	<pre>TInterval (13.685, 16.948) x̄=15.317 Sx=2.946 n=15.000</pre>
---	---

Here you have a "95% confidence interval" of the mean of the measured population being between 13.685 and 16.948.

Before leaving this line of discussion, we'll (if you read on) see how a Z test could be used and discuss something exotic called "The Central Limit Theorem" and use terms such as "sampling distribution" and "distribution of the means". Here's the theorem:

Regardless of the distribution of the parent population with mean  $\mu$  and variance  $\sigma^2$ , the distribution of the means (also called the sampling distribution) of random samples will approach a normal distribution with mean  $\mu$  and variance  $\sigma^2/n$  as the sample size  $n$  goes to infinity.

You may use this theorem without really realizing it or understanding it. Notice how the sampling distribution is more sharply defined (clumped around the mean) than is the parent population, by a factor of the square root of  $n$ .

Here's a problem. A group of eminent students from another math and science school claim that the average number of hours studied per week by math and science students is 15.3 with a standard deviation of 4.1. You take a random sample survey of 36 students at ASMS and get an  $\bar{X}$  of 14. What is the probability that a sample of 36 on a population with mean 15.3 and  $\sigma$  of 4.1 will give an average of no more than 14? We use the Z-Test in a seemingly convoluted and arcane way.

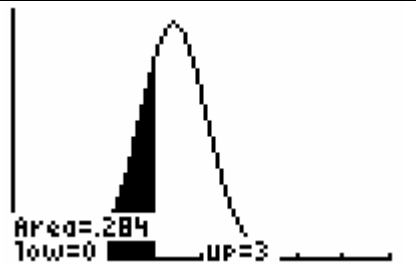
<p><b>STAT-TESTS-1.</b> Select <i>Stats</i>. Place 14 in <math>\mu_0</math> and 4.1 in <math>\sigma</math>. Place 15.3 in <math>\bar{X}</math> and 36 in <math>n</math>. Select <math>&gt;</math> for the hypothesis so you will get a probability for <math>&lt;</math> (the null, remember?). Select and Enter <i>Calculate</i> to get a probability <math>p</math> of 0.029.</p>	<pre>Z-Test Inpt:Data  [STAT] μ₀:14 σ:4.1 x̄:15.3 n:36 μ:≠μ₀ &lt;μ₀ &gt;μ₀ [D0] [ ] Draw</pre>
---	--

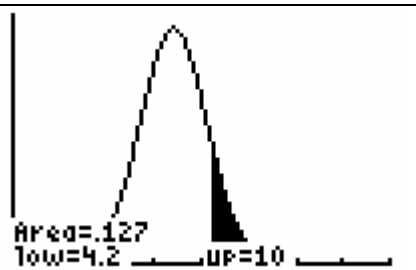
The use of the Central Limit Theorem is taken care of for you, including dividing the variance by 36. You can use the Z test because of the sampling distribution is normal enough for an  $n$  as large as 36. This last one is pretty confusing. Let's stop this discussion while you're still in a daze so you can rest better. You probably won't ever need to do this particular sequence. You may think it's all wrong anyway.

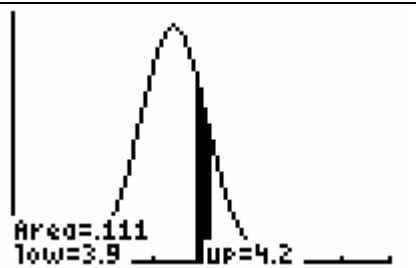
Now that you're rested, let's finish this page with some more "normal" exercise. Someone received a large government grant to study the lengths of index fingers of college professors. The mean was found to be 3.4 inches with a standard deviation of 0.7 inches. What percentage of professors are expected to have a finger length of 3 inches or less? Of 4.2 inches or more? Between 3.9 and 4.2 inches? What length is such that 90% of all professors have shorter index fingers? What middle range contains 80% of all professors? Why can't you get a grant to answer such hard questions?

<p><b>WINDOW.</b> <math>Xmin=0, Xmax=8, Xscl=1, Ymin=0, Ymax=0.6, Yscl=1, Xres=1</math>. <b>2<sup>nd</sup>-DISTR-DRAW-</b> 1. Finish with "0,3,3.4,.7)" for the lower bound, upper bound, mean, and <math>\sigma</math> as shown</p>	<pre>ShadeNorm(0,3,3.4,.7)</pre>
--	----------------------------------

ShadeNorm( is a drawing form of normalcdf( from **2<sup>nd</sup>-DISTR-2**.

<p><b>ENTER</b> to get the result shown. The percentage is 28.4%</p>	
--	--

<p>Repeat with a lower bound of 4.2 and an upper of 8 to get the following. At first you also get the previous shading? So <b>CLEAR</b> back to a clear home screen. <b>2<sup>nd</sup>-DRAW-1-ENTER</b>. Now repeat the above with the new bounds to answer the second question. The answer is 12.7%.</p>	
---	---

<p>Now handle the between 3.9 and 4.2 question. It's a good sign if you can do this without more details here- and good practice.</p>	
---	--

Now for the last two questions. Remember that the normal distribution is symmetrical about the mean.

<p>Use <b>2<sup>nd</sup>-DISTR</b> to get the home screen shown.</p>	<pre>invNorm(.9,3.4,. 7) 4.297 normalcdf(2.5,4. 3,3.4,.7) .801</pre>
--	--

Note that the first answer is 4.3 inches.  $4.3-3.4=0.9$ .  $3.4-0.9=2.5$ . Therefore by symmetry 2.5 is the lower bound for normalcdf(. Area  $1.0-0.9=0.1$ . Therefore

again by symmetry since all values above 4.3 have an area of 0.1 (because we already know that all values below 4.3 have an area of 0.9), the area below 2.5 also has an area of 0.1 and the middle part has an area of 0.8, verified by `normalcdf(` using 2.5 and 4.3 as the test bounds. The answer to the last question is of course 2.5,4.3.