

Chapter 4 – The Standard Deviation as a Ruler and the Normal Model

The standard deviation is the most common measure of variation; it plays a crucial role in how we look at data. Z-scores measure standard deviations above or below the mean as a pure number (no units) and are useful as measures of relative standing. Normal models are very useful as many random variables (at least approximately) follow their unimodal, symmetric shape.

Z-SCORES

The z-score for an observation is $z = \frac{(obs - \bar{y})}{s}$ or $\frac{obs - \mu}{\sigma}$ for a population, where *obs* is the value of interest.

Positive values indicate the observation is above the mean; negatives mean the value is below the mean. Calculating them is easy as long as one keeps in mind that the subtraction in the numerator must be done before the division. Calculators follow the arithmetic hierarchy of operations.

For example, in the 2004 Olympic women's heptathlon, Austra Skujyte of Lithuania put the shot 16.4 meters; the mean distance for all shotputters in the contest was 13.29 m with standard deviation 1.24 m. Carolina Klufft won the long jump with a 6.78 m jump; the average for all contestants was 6.16 m with standard deviation 0.23 m. Who actually did "better" relative to the other contestants? Scaling makes a direct comparison of the performances difficult, if not impossible. Z-scores can answer the question.

Two examples of the calculation for Skujyte are at right. The first (incorrect!) indicates she was 5.7 standard deviations *above* the mean – unreasonable given the standard deviation. The problem is failing to perform the subtraction first or enclosing the numerator in parentheses. The second (correct!) calculation indicates Skujyte's jump was 2.51 standard deviations above the mean. The last calculation shown indicates that Klufft's jump was 2.70 standard deviations above the mean. Relative to the field, her jump was the better performance.

```
16.4-13.29/1.24
5.682258065
(16.4-13.29)/1.24
2.508064516
(6.78-6.16)/.23
2.695652174
```

WORKING WITH NORMAL CURVES

What proportion of SAT scores are between 450 and 600? SAT scores for each of the three tests (writing, verbal and math) are approximately normal with mean 500 and standard deviation 100, or $N(500, 100)$. There are two ways to answer this question with a TI calculator. One will draw the curve; the other just answers the question. Both start at the same place: the Distributions menu. On an 83/84, press $\text{2nd}[\text{VARS}]$ (DISTR). If you are using an 89, the Distr Menu is F5 in the Statistics/List Editor application. On a TI-83/84, the screen at right will appear. Notice the arrow pointing down at the bottom left. There are more distributions which can be used; more will be said about some of them later. At the top of the screen, there are two choices DISTR (the default) which merely gives distribution values, and DRAW which will draw the curves and shade the appropriate areas.

```
0:QUIT DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:tpdf(
5:tcdf(
6:x²pdf(
7:x²cdf(
```

First, let's answer the question. On an 83/84 the menu option to select is 2: normalcdf(; Normal Cdf is menu option 4 on 89s. Either press the down arrow and then ENTER or just press 2 . The command will be transferred to the home screen. It normally requires two parameters to be entered: the z-score for the low end of the area of interest and the z-score for high end. Separate the entries by commas, and finish by closing the parentheses.

A score of 450 is 0.5 standard deviations below the mean, so its z-score is -0.5. A score of 600 is 1 standard deviation above the mean; its z-score is 1.

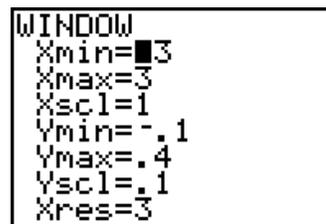
```
normalcdf(-.5,1)
.5328072082
```

The command has been entered in the screen at right, and **ENTER** was pressed to execute the command. We see that about 53.3% of all scores on the SAT will be between 450 and 600. TI-83/84 calculators default to a mean of 0 and standard deviation 1. Since we are working with standardized values (z -scores), there is no need to input these values.

At right is the TI-89 series input screen. With these calculators, you need to explicitly specify the mean and standard deviation for the distribution of interest.



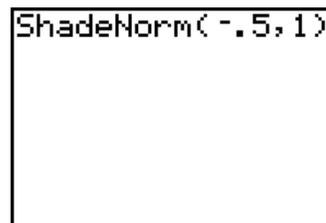
With TI-83/84 calculators, to find the area and have it shaded, one needs to first set the window (**ZOOM** 9 does not work here). For any normal model, most values will range from about -3 to 3 (three standard deviations either side of the mean) by the 68-95-99.7 Rule. Since the whole area under the curve is 1, the height of the curve will be a small number; we have set the Y_{min} to -0.1 and Y_{max} to 0.4 as in the screen at right.



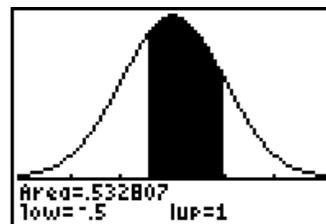
Now press **2nd** **VAR** (**DISTR**) and arrow to **DRAW**. We want choice 1, so press **ENTER**.



The command has been transferred to the home screen. Enter the z -scores for the low end of interest and the high end as at right, then press **ENTER**.



The graph should look like the one at right. Again, we see the area is about 53.3%; we also see what portion of the normal curve it represents.

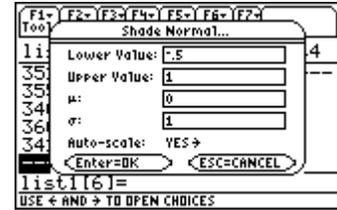


When working with the **DRAW** option, the graph must be cleared between successive commands or the shaded area will accumulate until the whole curve is shaded. To clear the drawing, press **2nd** **PRGM** (**DRAW**) then press **ENTER** to select option 1: **C1rDraw**.

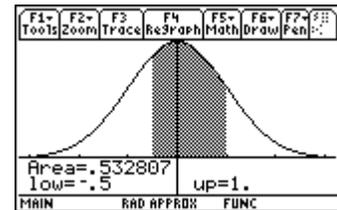
If you are using a TI-89 series calculator, this is a little more straightforward. From the **F5** **Distr** menu, press the right arrow key to expand the **Shade** menu. **Shade Normal** will be highlighted. Press **ENTER** to select that option.



The input screen is similar to the one seen above that simply computes the area. There is one major difference at the bottom. You have the option to have the calculator **Auto-scale** the graph. Use the right arrow key to expand the option box, and the down arrow key to move the selection to YES, if needed.



Displaying the graph puts you into the **Graph** application. To return to the Statistics app, press **2nd|APPS**.

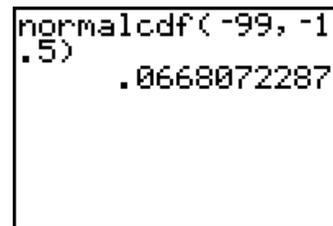


Another Example

A cereal manufacturer makes boxes labeled as 16 ounces; but the boxes are actually filled according to a normal model with mean 16.3 ounces and standard deviation 0.2 ounces. We want to know what fraction of all boxes will be “underweight,” that is, contain less than the advertised 16 ounces.

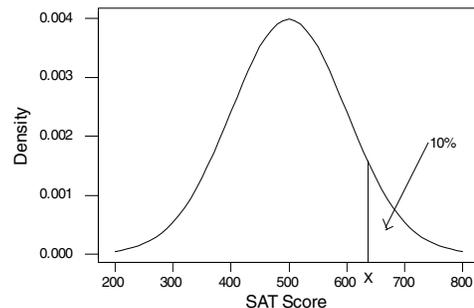
Strictly speaking, Normal models extend from $-\infty$ to ∞ (negative infinity to infinity). On the calculator, ∞ is represented as $1e99$ (10^{99}). To enter this, one presses **1|2nd|.99|9|9|**, but practically, any “very large” negative number (say, -99) will work for $-\infty$ and any large positive number (say 99) for ∞ since we know almost all of the area is between -3 and 3 standard deviations away from the mean.

We want to know what fraction of all boxes are less than 16 ounces, so the low end of interest is $-\infty$ (we entered -99 as the stand-in); the upper end of interest is 16 which corresponds to a z-score of -1.5 . The command and the result are at right. We see that about 6.7% of all boxes of this cereal should be underweight.

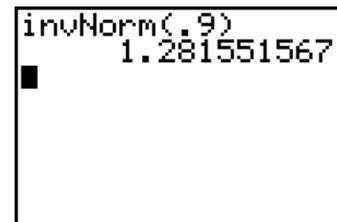


WORKING WITH NORMAL PERCENTILES

Sometimes the area under the curve is given and the corresponding value of the variable is of interest. For example, in the SAT model used before, how high must a student score to place in the top 10%? In a sketch of the normal curve, the unknown value, we’ll call it X , separates the top 10% from the lower 90%. We first have to find a corresponding z-score.



This is the opposite, or inverse, situation of that we’ve just explored. On the TI-83/84 DISTR menu, the command is **3:invNorm(.9)**. Press **2nd|VARS|3|** to transfer the command to the home screen. The parameter for this command is area to the *left* of the point of interest (.90 or 90%). Press **ENTER** to execute the command. The z-score of interest is 1.28. To be in the top 10%, your score must be 1.28 standard deviations above the mean. We have to solve the equation

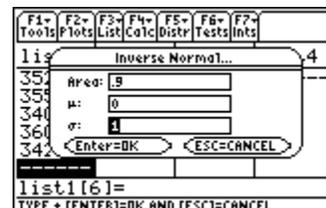


$z = (x - \mu) / \sigma = 1.28$ or $(x - 500) / 100 = 1.28$ After doing the algebra, we see that a score of 628 will put a person in the top 10% of all SAT scores; practically since scores are reported rounded to multiples of 10, a score of 630 is needed.

If you are using a TI-89, use **Inverse Normal** from the **[F5]** Distr menu, found by expanding option **2: Inverse**.



To do this computation on TI-89 calculators, you must again specify the mean and standard deviation (0 and 1) as I have done here. You will find the same z-score as above. Complete the algebra to find the SAT score of interest.



Another Example

The cereal company's lawyers are not happy with 6.7% of boxes being underweight. They want at most 4% to be underweight. What mean must the company reset its machines to in order to achieve this target? We'll use **invNorm** for a standard normal model to find the z-score corresponding to 4% of the area below this value, then use some algebra to solve for the unknown mean.

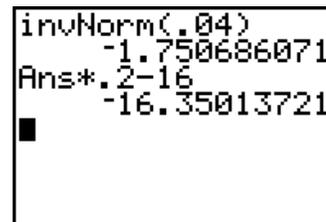
The z-score of interest is -1.75 . We need to solve the equations

$$z = (x - \mu) / \sigma = -1.75$$

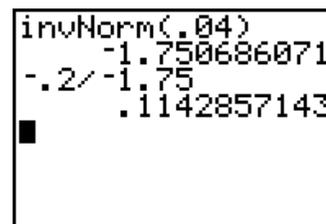
$$(16 - \mu) / 0.2 = -1.75$$

Now multiplying both sides by 0.2, and subtracting 16 from both sides gives

$-\mu = -16.35\sigma_x^2$ or $\mu = 16.35$. In order to achieve the target of no more than 4% of boxes being underweight, the machine will have to be set for an average of 16.35 ounces per box.



Now, suppose the president of the cereal company wants the mean to be no more than 16.2 oz (she doesn't want to give away too much cereal.) To meet that target along with the 4% target for the proportion of boxes that are underweight, the company must change the standard deviation. We already know that a z-score of -1.75 corresponds to the 4th percentile. What changes here is the algebra: we know the desired μ and want to find σ . $z = (x - \mu) / \sigma = -1.75$ becomes $(16 - 16.2) / \sigma = -.2 / \sigma = -1.75$, and finally we find $\sigma = 0.114$ oz.

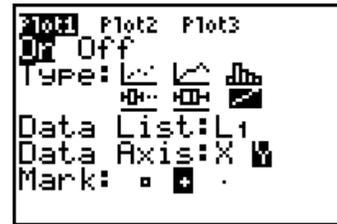


IS MY DATA NORMAL?

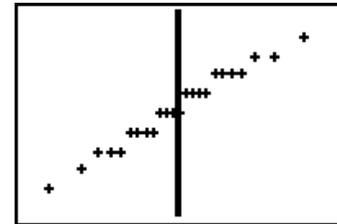
It is one thing to assume data follows a normal model. When one actually has data the model should be verified. One method is to look at a histogram: is it unimodal, symmetric and bell-shaped? Another method is to ask whether the data (roughly) follow the 68-95-99.7 rule. Both of these methods might work well with a fairly large data set; however, there is a specialized tool called a normal probability (or quantile) plot that will work with any size data set. This plots the data on one axis against the z-score one would expect if the data were exactly normal on the other. If the data are from a normal distribution this plot will look like a diagonal straight line.

TI-83/84 Procedure

Recall the data on pulse rates. In one histogram, they appeared symmetric and unimodal. In another, they appeared somewhat uniform. Might we consider these data as having come from a normal model? The data are in list L1. Press $\text{2nd}[\text{Y=}]$ to get to the first Stat Plot screen. Once here, you should always check that all plots are off except the one you will use. Select the plot to use and press ENTER . The normal probability plot is the last plot type. Use the right arrow to move there and press ENTER to move the highlight. Notice you have a choice of having the data on either the x or y axis. It doesn't really matter which you choose; many statistical packages put the data on the x -axis; many texts (including DeVeaux, Velleman, and Bock) put the data on the y -axis. As we have seen before, you have a choice of three marks for each data point. Select the one you prefer.



Pressing $\text{ZOOM}[\text{9}]$ displays the graph. This graph is very straight. These data could indeed be considered to have come from a normal distribution. We also see an indication of *granularity* in the plot. That is because we have several data values which occurred several times.

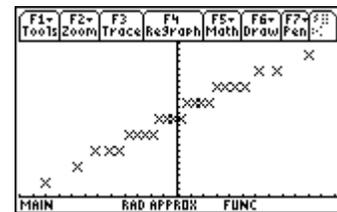


TI-89 Procedure

With the pulse rate data in list1, press F2 (Plots) and select choice 2: Norm Prob Plot. The plot number defaults to one more than the last plot defined. Enter the list name to use (press $\text{2nd}[\text{=}]$ to get to the VAR-LINK screen). Notice you have a choice of having the data on either the x - or y -axis. It doesn't really matter which you choose; many statistical packages put the data on the x -axis; many texts (including DeVeaux, Velleman and Bock) put the data on the y -axis. As we have seen before, you have a choice of marks for each data point. Select the one you prefer. The calculator will store the z -scores in a list. Just take the default here. Pressing ENTER calculates the z -score for each data value. To display the plot, press F2 again, check that all other plots are "turned off" (uncheck them by moving the cursor and pressing F4) then press F5 to display the plot.



The graph is very close to a straight line. These data may be considered as having come from a normal distribution. We also see an indication of *granularity* in the plot. That is because we have several data values which occurred several times.

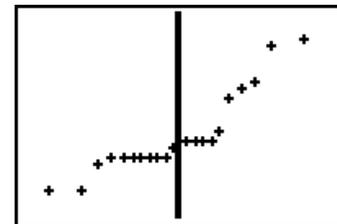


Skewed distributions often show a curved shape. Data on the cost per minute of phone calls as advertised by Net2Phone in *USA Today* (July 9 2001) to 22 countries were as follows:

7.9	17	3.9	9.9	15	9.9	7.9	7.9	7.9	7.9	8.9
21	6.9	11	9.9	9.9	7.9	3.9	22	9.9	7.9	16

We have entered in a list and have defined the normal probability plot as above.

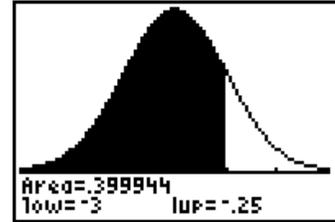
The plot obtained is at right. Not only does it show a general upward curve, it also displays granularity again. This occurs when a particular data value occurs several times (as with 7.9 cents per minute which was in the list seven times.)



WHAT CAN GO WRONG?

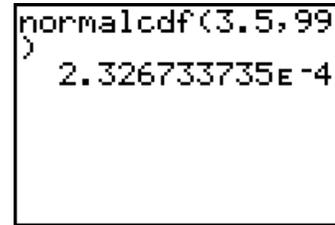
Why is my curve all black?

For the SAT scores curve, the graph indicates more than half of the area is of interest between 200 and 475 (z-scores of -3 and $-.25$); the message at the bottom says the area is 40%. This is a result of having failed to clear the drawing between commands. Press 2nd|PRGM then ENTER to clear the drawing, then reexecute the command.



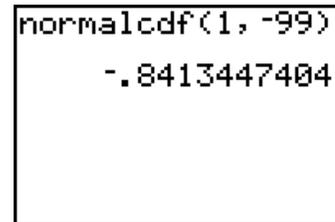
How can the probability be more than 1?

It can't. If the results look like the probability is more than one, check the right side of the result for an exponent. Here it is -4 . That means the leading 2 is really in the fourth decimal place, so the probability is 0.0002. The chance a variable is more than 3.5 standard deviations above the mean (this would be a box of the cereal more than 17 ounces) is about 0.02%.



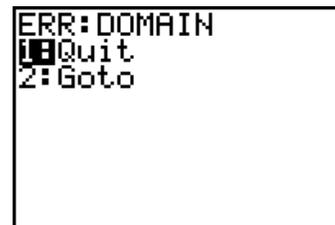
How can the probability be negative?

It can't. The low and high ends of the area of interest have been entered in the wrong order. As the calculator does a numerical integration to find the answer, it doesn't care. You should.



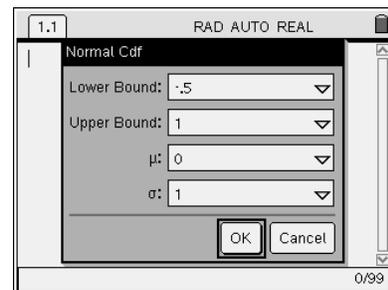
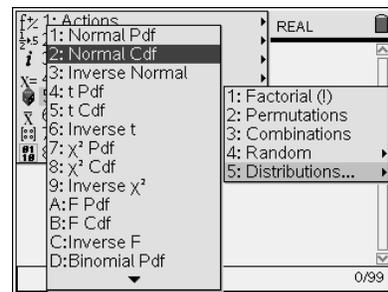
What's Err: Domain?

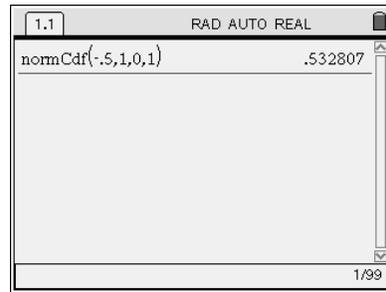
This message comes as a result of having entered the `invNorm` command with parameter 90. (You wanted to find the value that puts you into the top 10% of SAT scores, so 90% of the area is to the left of the desired value.) The percentage must be entered as a decimal number. Re-enter the command with parameter .90.



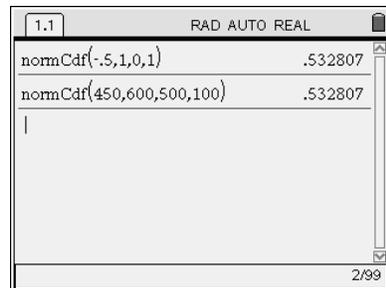
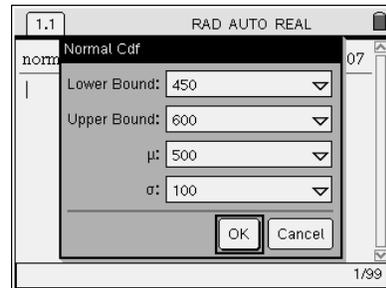
Commands for the TI-Nspire™ Handheld Calculator

To compute the area under a standard normal curve, start on a Calculator screen. Press menu , and then select Probability, Distributions, and Normal Cdf. This opens an input box. Enter the bounds, zero for the mean, and 1 for the standard deviation. tab to OK and enter . For the SAT example use $-.5$ and 1 as the bounds.

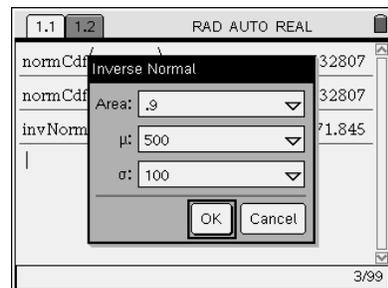
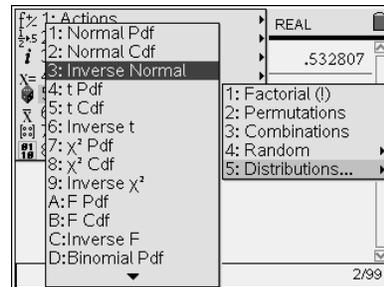


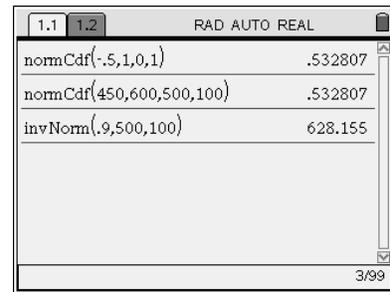


To compute the area under any other normal curve, enter the bounds, the mean, and the standard deviation. We can compute the same result for the SAT problem by using 450 and 600 for the bounds, 500 for the mean, and 100 for the standard deviation.



To compute the percentile, press MENU , and then select Probability, Distributions, and Inverse Normal. Type the area, mean, and standard deviation, TAB to OK and ENTER .





To create a normal probability plot, first type the values into a list. For example, consider the list of pulse rates we named *rate*. Press $\left[\text{2nd}\right]\left[\text{STAT}\right]$ and then select Data & Statistics. At first you will see a plot of dots. Use the arrows to move to the bottom of the display until “Click to add variable” appears. Press $\left[\text{2nd}\right]\left[\text{STAT}\right]$, highlight the variable, in this case *rate*, and $\left[\text{enter}\right]$ again. You will see the dotplot. Now press $\left[\text{menu}\right]$, select Plot Type, and then Normal Probability Plot.

