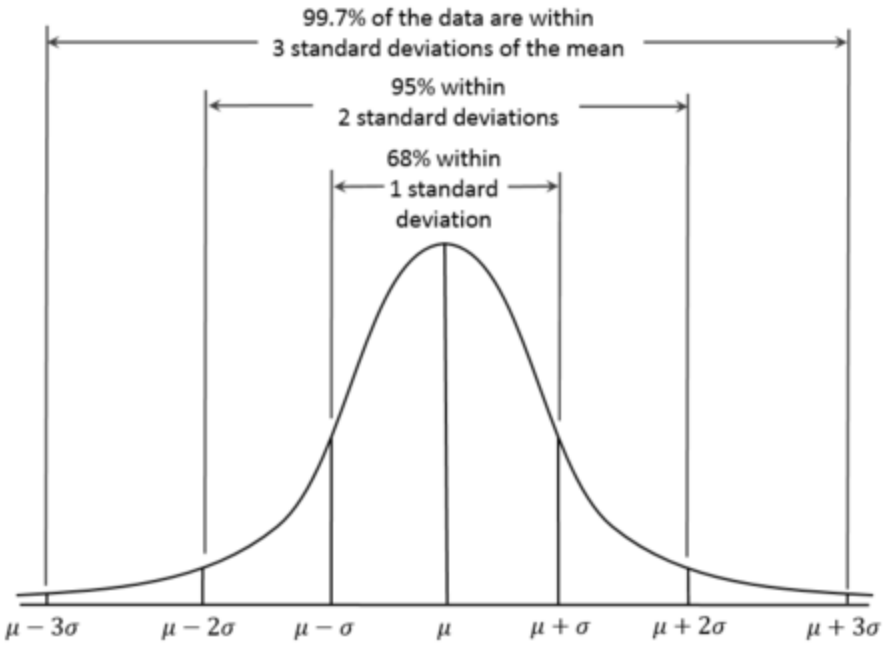
David Youngberg
BSAD 210—Montgomery College

LECTURE 08: THE NORMAL DISTRIBUTION AND THE CLT

I. Question: How tall can a woman be and still be considered "short?"
   a. This is obviously a judgement call but we can get some insight by understanding a normal distribution, which height (as well as many other variables) follow.
   b. A *normal distribution* has a bell-shaped distribution that is symmetric around the mean.
      i. By "symmetric" we mean each side of the mean has the same shape. This renders the mean equal to the median.
      ii. When we work with a normal distribution, a particular shape will be working in the math, but we have language to describe how actual data deviates from that particular shape.
   c. *Skew*—measurement of the symmetry of the distribution
      i. Symmetric means the tails are equally dispersed. Mean equals median: e.g. height. Normal distributions are symmetric.
      ii. A positive skew means there is a long tail (few extreme values) on the right. Mean is greater than median: e.g. salary
      iii. A negative skew means there is a long tail (few extreme values) on the left. Mean is less than median: e.g. test score
   d. *Kurtosis*—measurement of the "peakness" of the distribution
      i. If zero, the peak resembles that of a normal distribution.
      ii. If negative, the peak is flatter than the normal. More of the variance is due to observations near the mean.
      iii. If positive, the peak is sharper than the normal. More of the variance is due to extreme observations.

II. Empirical Rule
   a. This distribution follows the *empirical* rule:
      i. About 68% of all observations are within one standard deviation of the mean;
      ii. About 95% of all observations are within two standard deviations of the mean; and
      iii. About 99.7% of all observations are within three standard deviations of the mean.



iv. This graphical representation shows how each segment of the normal distribution breaks down.

v. Note anything beyond these three standard deviations is an outlier.

III. Central Limit Theorem

a. The *central limit theorem (CLT)* states that the sample means of large-sized will be normally distributed regardless of the shape of the distribution.

i. In other words, suppose you take 100 samples of a population, with each sample having many observations in it. If you take the average of each sample, you'll get 100 sample averages. Those 100 averages will form a normal distribution, with a few averages being very low or very high and many being right in the middle.

b. What's interesting is that the CLT works for ***any*** population distribution.

IV. Determine percent height

a. What percent of women are "average height?" What percent are "short?" What definition of "short" would result in 10 percent of women being "short?" We can answer these questions with Excel.

b. Excel has two main functions that explicitly deal with the normal distribution: "=NORM.DIST" and "=NORM.INV".

c. The first tells us the area under a normal distribution at a particular value. We will call this α (alpha) though Excel calls this probability.

d. In both cases, you must provide the standard deviation and average.

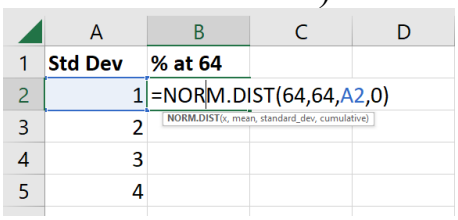
i. Height follows a normal distribution. Women are on average about 64 inches tall with a standard deviation of about 2 inches.

V. NORM.DIST

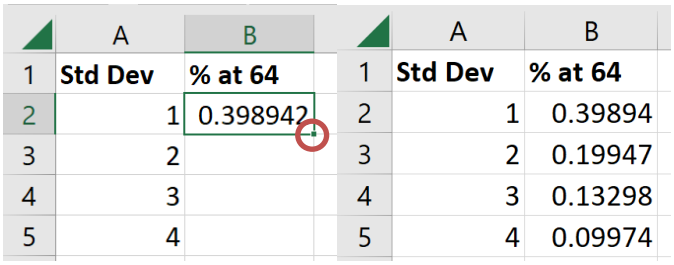
a. This function tells you the area under the curve tells you the portion of the population is that value. It has a cumulative option; right now, leave it off (0, or FALSE).

i. What percent of women are of average height? Type "=NORM.DIST(64,64,2,0)" and press ENTER. You should get about 19.95%.[1]

ii. Let's see how it changes as you change standard deviation. Set up the cells as you see below, with the standard deviation referencing another cell rather than being a number. (You can also click the appropriate cell while inputting information for the NORM.DIST command.)



iii. Once you press ENTER, you can copy/paste the cell (or double-click the black square in the lower right-hand corner) and the cell reference will update.

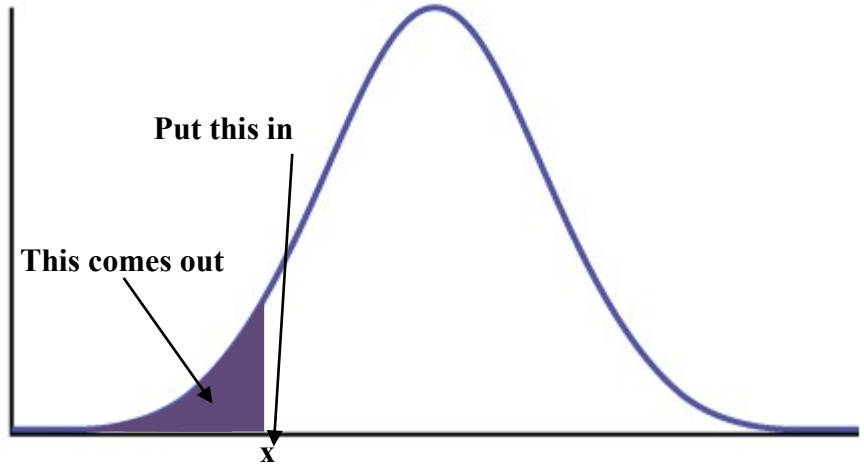


iv. Notice that as standard deviation increases, the portion of the population that's exactly the average decreases. That's because

[1] For the more mathematically inclined, this would seem like a mistake. To find this kind of information, you'd take the integral for the area under a curve (using the equation for a normal distribution). That's indeed exactly what happens when we determine the percent of observations for a particular range (e.g. percent of the population with a height from 60 to 65). But it can't work here because there's no range. It would be like finding the area of a rectangle when the width is zero. So what's going on? The answer is that Excel is programmed well. The programmers know that a precise estimate, when cumulative is set to FALSE or 0, doesn't apply here. So when you set cumulative as FALSE or 0, Excel uses something called a probability mass function instead of its regular cumulative distribution function. (If it didn't the result would always be zero, which is quite boring!) A probability mass function is built to handle discrete (read: non-range) values like what we're doing when we try to figure out the percent of the population who has exactly average height.

as standard deviation falls, observations are reallocated to the center to keep that bell-shaped curve. If you increase standard deviation, observations must leave the average value.

b. Cumulative Option
   i. If you select the cumulative option (1, or TRUE), it will tell you the α at that value or lower.



IMPORTANT: This function will ***always*** display for the value AND LOWER. If you want to find the area for higher than the x value, subtract the result from 1.

c. We can use this technique to answer our original question: what percent of women are "short?" To answer that question, we need a definition of "short." Suppose we consider any women whose 60 inches or shorter to be "short." What percent are 60 inches tall or less?
   i. "=NORM.DIST(60,64,2,1)" does this job. Note the "1" at the end; that's because we turned cumulative "on" and Excel will now look at what percent of the curve is at 60 inches or less. You should get about 2.275%, or a little over 2 percent of women are "short."
   ii. If you think more than just over 2 percent of women are short, then that means 60 inches is too low. Let's try different cutoffs:

	A	B
1	**Cutoff**	**%**
2	60	=NORM.DIST(A2,64,2,1)
3	61	=NORM.DIST(A3,64,2,1)
4	62	=NORM.DIST(A4,64,2,1)
5	63	=NORM.DIST(A5,64,2,1)

	A	B
1	**Cutoff**	**%**
2	60	0.0227501
3	61	0.0668072
4	62	0.1586553
5	63	0.3085375

iii. If you think something like 15 percent of women are short, then that means your definition of "short" is about 62 inches. If you think it's closer to 30 percent, your definition is about 63 inches.

d. We can also use this function to derive the Empirical Rule.

i. We begin with what's called a *standard normal distribution*: a **mean of zero** and a **standard deviation of one**. Here's a table of αs for various standard deviations:

-3	-2	-1	1	2	3
0.00135	0.02275	0.158655	0.841345	0.97725	0.99865

Remember: Excel always displays the α at the value and below.

ii. Here is the same thing with the formulas shown:

-3	-2	-1	1	2	3
=NORM.S.DIST(-3,1)	=NORM.S.DIST(-2,1)	=NORM.S.DIST(-1,1)	=NORM.S.DIST(1,1)	=NORM.S.DIST(2,1)	=NORM.S.DIST(3,1)

iii. Now subtract the left-side boundary from the right-side boundary (e.g. 0.841345 – 0.158655) and round:

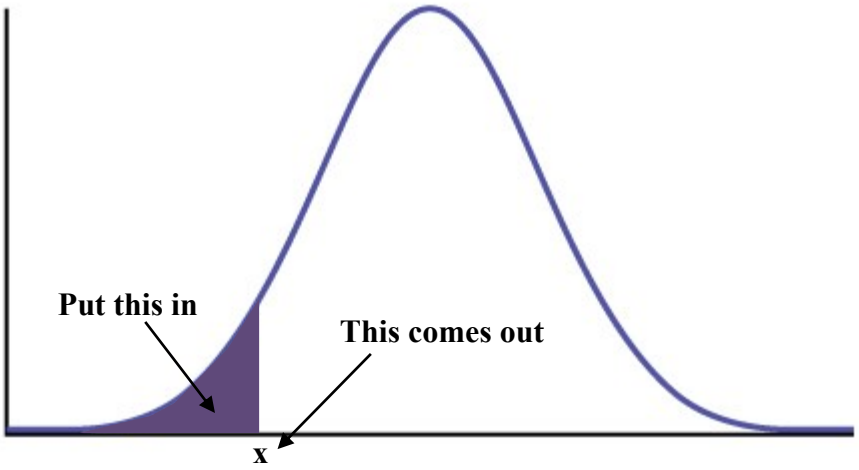
-3	-2	-1	1	2	3
0.00135	0.02275	0.158655	0.841345	0.97725	0.99865
		68%			
	95%				
99.7%					

VI. NORM.INV

a. The second function inverts the input and output; rather than putting in the x value and getting out the area under the curve, you put in the area under the curve and get out the x value.

i. You still have to provide the mean and standard deviation.

ii. As always, express your alpha (the area under the curve) as a decimal.



b. Suppose you think 25 percent of women are "short." What cutoff would you need to get 0.25? NORM.INV can do this:
    i. Type: =NORM.INV(0.25,64,2) You should get *about* 62.65 inches.
    ii. Note that putting in =NORM.DIST(62.65,64,2,1) gives you something very close to exactly 25%: 0.24984. If you reference the cell with the INV output, you'll get exactly 25%.