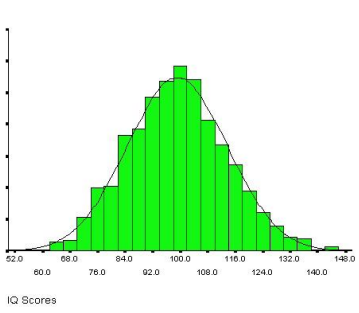
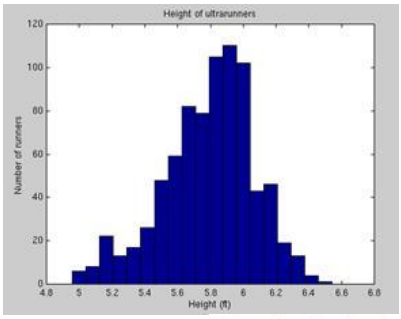
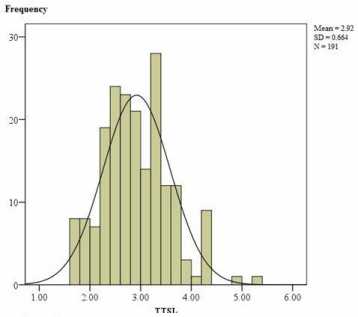
David Youngberg
BSAD 210—Montgomery College

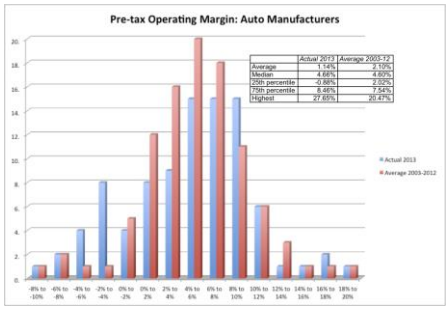
LECTURE 08: THE NORMAL DISTRIBUTION AND THE CLT

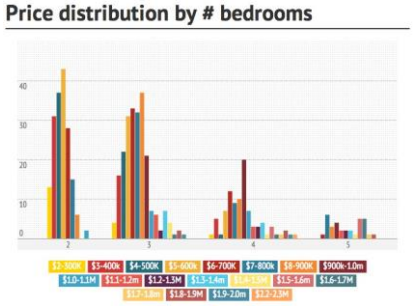
I. Normal Distribution[1]











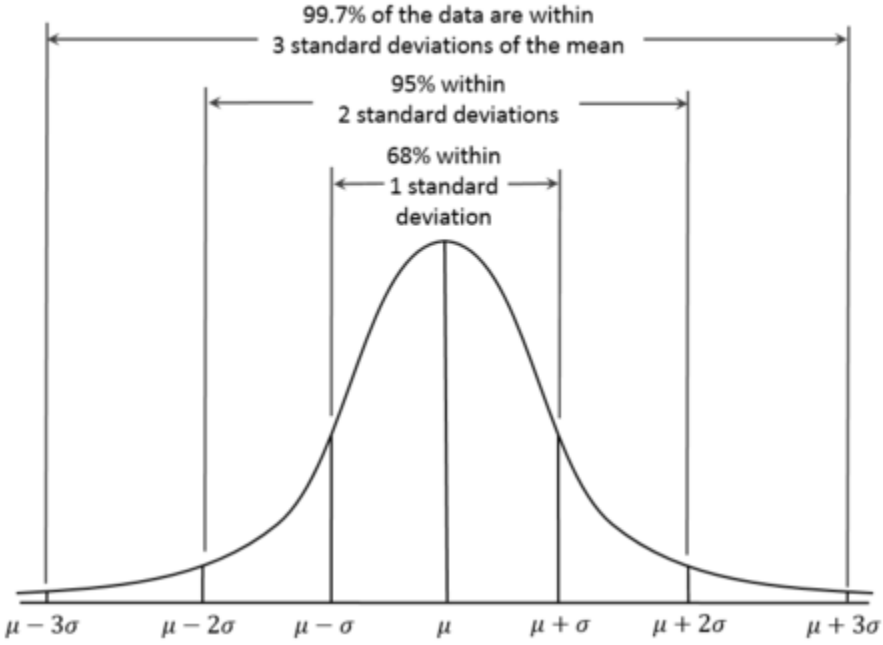
a. There are many other examples: crop yields; customer traffic; sales data; product quality; and so on.
b. All of these examples have a distribution we call a *normal distribution*—a bell-shaped distribution that is symmetric around the mean.
    i. By "symmetric" we mean each side of the mean has the same shape. This renders the mean equal to both the median and the mode.
    ii. While none of these empirical examples have *exactly* the perfect bell-curved shape, many of them approximate it. Thus we analyze data, we assume an ideal bell-curve.

---

[1] IQ: http://www.psychology.emory.edu/clinical/bliwise/Tutorials/SOM/smmod/scalemea/print2.htm
Height: http://ib.berkeley.edu/courses/ib162/Week1.htm
Time-to-stop line (time taken to determine to stop for a yellow light):
http://www.fhwa.dot.gov/publications/research/safety/09049/
Pre-tax Net Profit Margin (Operating Margin) of Auto Makers:
http://aswathdamodaran.blogspot.com/2013/09/valuation-of-week-1-tesla-test.html
Price distribution of homes sold in Berkeley by number of bedrooms:
http://www.berkeleyside.com/2013/02/15/berkeley-house-prices-tick-up-after-years-of-slump/

c. While many variables have a normal distribution, there are some that don't. Examples:
    i. Proportion of Americans who voted for each candidate in the last Presidential election.
    ii. The number of people along a beach.
    iii. Income.

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. Qualities of a Normal Distribution
a. *Skew*—measurement of distribution symmetry
    i. Symmetric means the tails are equally disperse. 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

b. *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.

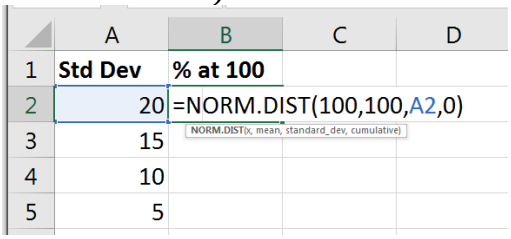
IV. Central Limit Theorem
   a. The *central limit theorem (CLT)* states that the sample means of large-sized will be normally distributed regardless of the shape of the distribution.
      i. In other words, suppose you take 100 samples of a population, with each sample having many observations in it. 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.

V. Overview
   a. Excel has two main functions that explicitly deal with the normal distribution: "=NORM.DIST" and "=NORM.INV".
   b. 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.
   c. The second tells us the value that would result at a particular α.
   d. In both cases, you must provide the standard deviation and average.

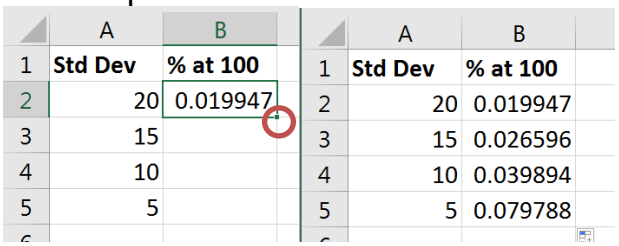
VI. "=NORM.DIST"
   a. This function tells you the area under the curve tells you the portion of the population is that value. It has a cumulative option; right now, leave it off (0, or FALSE).
      i. IQ follows a normal distribution with an average of 100 and a standard deviation of 15. What percent of the population has an exactly average IQ? Type "=NORM.DIST(100,100,15,0)" and press ENTER. You should get about 2.66%.[2]

[2] 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 an IQ from 80 to 95). 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

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.

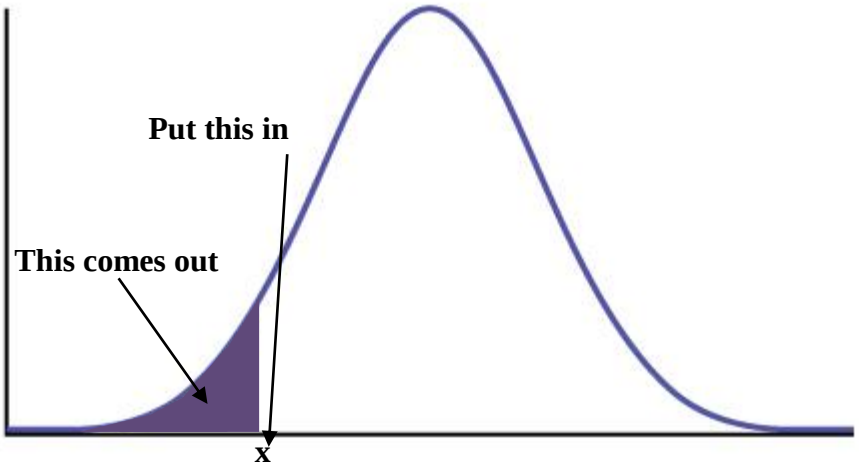


iv. Notice that as standard deviation falls, the portion of the population that's exactly the average increases. That's because as standard deviation falls, observations are reallocated to the center to keep that bell-shaped curve. If you increase standard deviation, observations must leave the average 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.

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 100 IQ. Interpreting the results as percents in this way is completely valid.



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.

ii. How will the cumulative portion change as we change standard deviation? Let's find out. Begin as before but now the cumulative option is "1" or TRUE.

	A	B	C	D
1	**Std Dev**	**% at 100**	**% 100 or less**	
2	20	0.019947	=NORM.DIST(100,100,A2,1)	
3	15	0.026596		
4	10	0.039894		
5	5	0.079788		

NORM.DIST(x, mean, standard_dev, cumulative)

iii. As before, double-click the box in the lower right-hand corner.

	A	B	C
1	**Std Dev**	**% at 100**	**% 100 or less**
2	20	0.019947	0.5000
3	15	0.026596	
4	10	0.039894	
5	5	0.079788	

	A	B	C
1	**Std Dev**	**% at 100**	**% 100 or less**
2	20	0.019947	0.5000
3	15	0.026596	0.5000
4	10	0.039894	0.5000
5	5	0.079788	0.5000

iv. Why is it always 50%? Because while the standard deviation changes kurtosis, it doesn't change skew. The distribution

remains symmetric so when the x value equals the mean, it will always be 50%.[3]

v. But if you set the value to something else, well:

	A	B	C	D	E	F	G	H	I	J	K	L
1		% At X or Less										
2	**Std Dev**	**120**	**115**	**110**	**105**	**101**	**100**	**99**	**95**	**90**	**85**	**80**
3	20	0.8413	0.7734	0.6915	0.5987	0.5199	0.5000	0.4801	0.4013	0.3085	0.2266	0.1587
4	15	0.9088	0.8413	0.7475	0.6306	0.5266	0.5000	0.4734	0.3694	0.2525	0.1587	0.0912
5	10	0.9772	0.9332	0.8413	0.6915	0.5398	0.5000	0.4602	0.3085	0.1587	0.0668	0.0228
6	5	1.0000	0.9987	0.9772	0.8413	0.5793	0.5000	0.4207	0.1587	0.0228	0.0013	0.0000

c. 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:

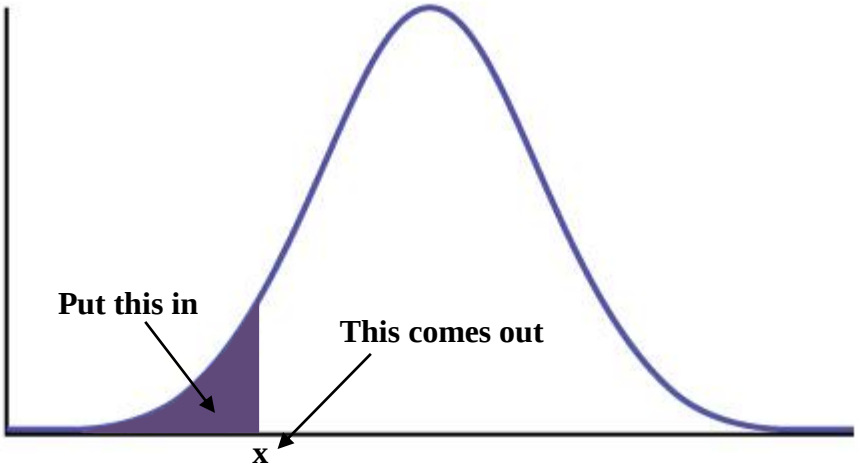
-3	-2	-1	1	2	3
0.00135	0.02275	0.158655	0.841345	0.97725	0.99865
		68%			
	95%				
99.7%					

VII. "=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.

[3] Astute observers would note that 50% stands in seeming contradiction to our earlier estimate of 2.66% at exactly 100. If 50% is below average (and 50% is above average), how is there any room for exactly average? Again, this is the natural result of using two different equations for when cumulative is set to FALSE (or 0) versus when it is set to TRUE (or 1). Again, if Excel used the cumulative distribution function here the result would be zero, not 0.0266.

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



b. As always, express your alpha (the area under the curve) as a decimal.
c. Imagine you work at company that makes kitchen appliances. Like all firms in this industry, you offer a warranty with your product. If it breaks within the warranty time frame, the company will replace it. But it's tricky determining how long the warranty should be:
    i. If you make the warranty too short, few people will buy the appliance. Many will think the appliance is of terrible quality.
    ii. If you make the warranty too long, you'll have to replace too many appliances and will make the entire line unprofitable. Lifetime warranties are only good for products that last a really long time.
    iii. Imagine a dishwasher lasts an average of 10 years with a standard deviation of 1.5 years. The distribution of the dishwasher lifespan is normal. The accounting department says the firm can afford to replace 4% of dishwashers sold. How long should the warranty last?
    iv. We are looking for the area under a normal distribution such that the α is 0.04. Where should the x be? (Note that if a dishwasher lasts less than x, we replace it; thus we are concerned with the left-handed side of the distribution.)
    v. Type "=NORM.INV(0.04,10,1.5)" and press ENTER. You should get about 7.37. You might round that down to a 7 year warranty. (But don't round it up! Then you'll be replacing more than 4% of dishwashers!)
d. Suppose you manage a sales team and you want to offer a bonus to the best salespeople. If you set the threshold for a bonus too high, no one will get the bonus and will resent you for setting a goal no one could

make. If you set it too low, it will be too easy and people won't work any harder. Indeed, the naturally hard workers will resent you for giving their lazier colleagues a bonus as well.

i. Imagine the average salesperson makes $40,000 in sales per month with a standard deviation of $5,000. Sales follow a normal distribution.
ii. Imagine you want to give 20% of your workforce a bonus. What's the sales target to get the bonus?
iii. Now you want to award the TOP salespeople; you're curious about the area under the curve but on the right-handed tail. If you want to reward the top 20%, then you want to not reward the bottom 80%. Remember, Excel always outputs the bottom of the distribution, even if you define the bottom to be large.
iv. Type "=NORM.INV(0.8,40000,5000)" and press ENTER. You should get about 44,208.11; best to round this up to $45,000; if you round down, you'll give bonuses to more than 20%.
v. Note that in this case, you might still give bonuses to more than 20%, even after rounding up. That's because the promise of a bonus will encourage people to work harder. Maybe you'd want make the cut-off $46,000 (though if there are more sales, you can afford more bonuses), but this gives you a starting point.