Empirical rule in Excel: demo notes

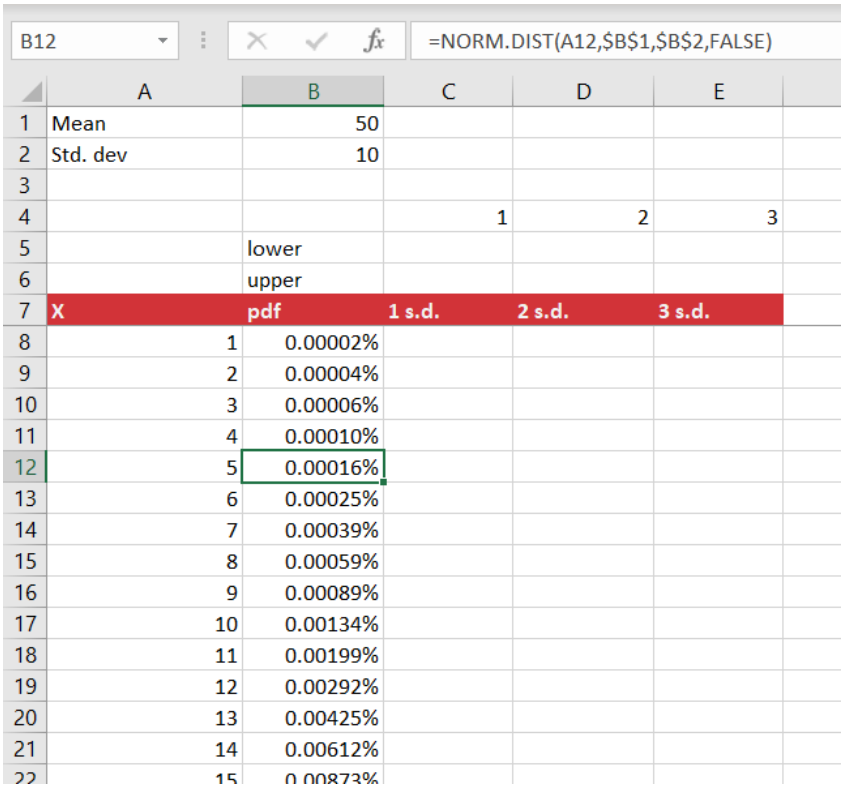
Download the exercise file: `empirical-rule.xlsx`

Also known as the 68-95-99.7 rule, the empirical rule is a statistical principle which states that, for a normal distribution:

- 68% of all datapoints lie within one standard deviation of the mean
- 95% of all datapoints lie within two standard deviations of the mean
- 99.7% of all datapoints lie within three standard deviations of the mean

Let's visualize this in Excel.

1. Our starting point lists a mean of 50, standard deviation of 10 and the numbers 1-100 ranging down rows 8-107.
   a. We want to find how likely it is for a datapoint to fall in that range. This is called a *probability density function*.

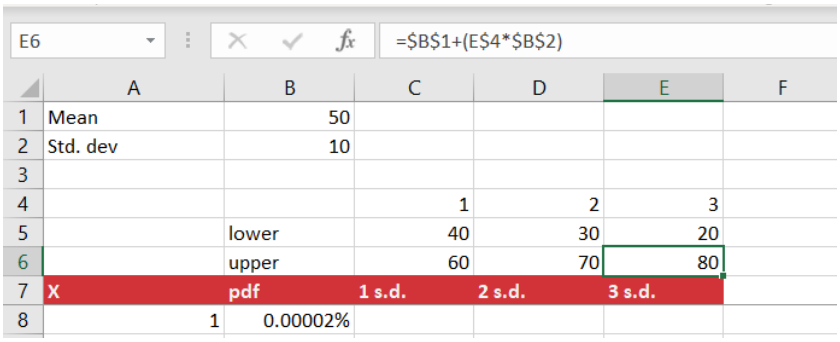


B12 | =NORM.DIST(A12,B1,B2,FALSE)

	A	B	C	D	E
1	Mean	50			
2	Std. dev	10			
3					
4			1	2	3
5		lower			
6		upper			
7	**X**	**pdf**	**1 s.d.**	**2 s.d.**	**3 s.d.**
8	1	0.00002%			
9	2	0.00004%			
10	3	0.00006%			
11	4	0.00010%			
12	5	0.00016%			
13	6	0.00025%			
14	7	0.00039%			
15	8	0.00059%			
16	9	0.00089%			
17	10	0.00134%			
18	11	0.00199%			
19	12	0.00292%			
20	13	0.00425%			
21	14	0.00612%			
22	15	0.00873%			

   a. For example, we see that we'd expect a value from a normal distribution with a mean of 50 and standard deviation of 10 to be 5 0.00016% of the time.

2. Now we want to calculate 1, 2 and 3 standard deviations from the mean so that we can visualize what percentage of values fall within those ranges due to the empirical rule.
   a. Take the upper and lower bounds of these ranges in cells C5:E6.



E6 | =B1+(E$4*$B$2)

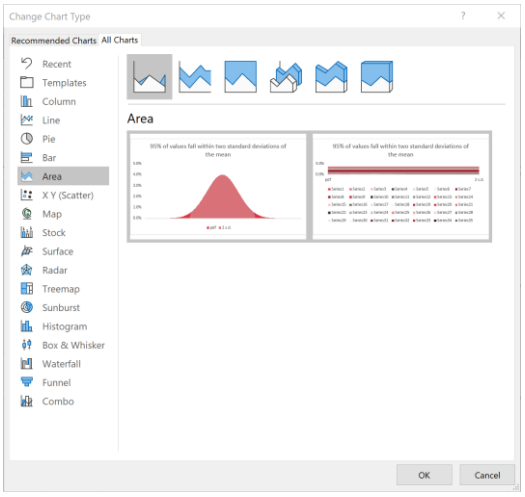
	A	B	C	D	E	F
1	Mean	50				
2	Std. dev	10				
3						
4			1	2	3	
5		lower	40	30	20	
6		upper	60	70	80	
7	X	pdf	1 s.d.	2 s.d.	3 s.d.	
8	1	0.00002%				

3. Now we will use conditional logic to pick up the parts of the pdf that fall within 1, 2 or 3 standard deviations of the mean. The formula for cell C8 will be `=IF(AND($A9>C$5,$A9<C$6),$B9,0)`. With these mixed references applied, you can fill out the rest of the range through column `E`.
   a. You can see that more of the distribution is included in your range as you increase the standard deviations.

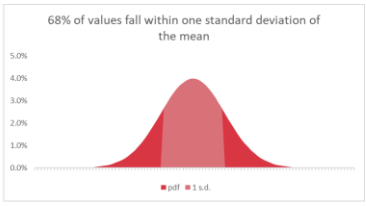
	A	B	C	D	E	F
1	Mean	50				
2	Std. dev	10				
3						
4			1	2	3	
5		lower	40	30	20	
6		upper	60	70	80	
7	X	pdf	1 s.d.	2 s.d.	3 s.d.	
23	16	0.01232%	0	0	0	
24	17	0.01723%	0	0	0	
25	18	0.02384%	0	0	0	
26	19	0.03267%	0	0	0	
27	20	0.04432%	0	0	0.00044318	
28	21	0.05953%	0	0	0.00059525	
29	22	0.07915%	0	0	0.00079155	
30	23	0.10421%	0	0	0.00104209	
31	24	0.13583%	0	0	0.0013583	
32	25	0.17528%	0	0	0.00175283	
33	26	0.22395%	0	0	0.00223945	
34	27	0.28327%	0	0	0.0028327	
35	28	0.35475%	0	0	0.00354746	
36	29	0.43984%	0	0	0.00439836	
37	30	0.53991%	0	0.005399097	0.0053991	
38	31	0.65616%	0	0.006561581	0.00656158	
39	32	0.78950%	0	0.007895016	0.00789502	
40	33	0.94049%	0	0.009404908	0.00940491	
41	34	1.10921%	0	0.011092083	0.01109208	
42	35	1.29518%	0	0.01295176	0.01295176	
43	36	1.49727%	0	0.014972747	0.01497275	
44	37	1.71369%	0	0.017136859	0.01713686	
45	38	1.94186%	0	0.019418605	0.01941861	
46	39	2.17852%	0	0.021785218	0.02178522	
47	40	2.41971%	0.02419707	0.024197072	0.02419707	
48	41	2.66085%	0.02660852	0.026608525	0.02660852	
49	42	2.89692%	0.02896916	0.028969155	0.02896916	
50	43	3.12254%	0.03122539	0.031225393	0.03122539	

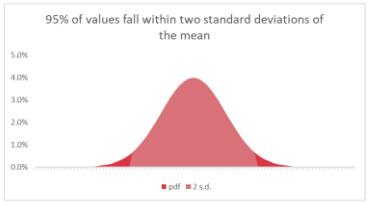
   b. What percentage of all values do you find in each column?

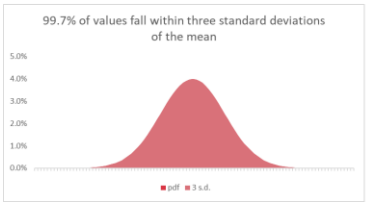
4. The charts to the right of the table show you what data falls within one, two and three standard deviations of the mean. We are using area charts to do so.



5. We can now see the results of the empirical rule on the normal distribution.
   a. So much of our data lies within three standard deviations of the mean that it's nearly impossible to detect any outliers.







Back to the slides for a wrap-up…