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.

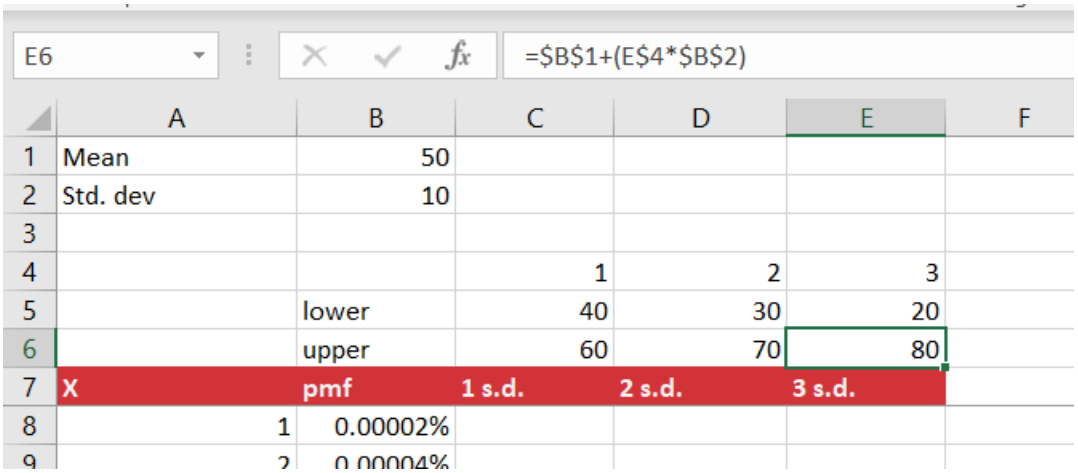
1. Our starting point lists a mean of 50, standard deviation of 10 and the numbers 1-100 ranging down rows 8-107.
   a. We want to find how likely it is for a datapoint to take on each value in that that range. This is called a *probability mass function* (pmf).

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

	A	B	C	D	E	F
1	Mean	50				
2	Std. dev	10				
3						
4			1	2	3	
5		lower				
6		upper				
7	X	pmf	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%				

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.
b. Sum the values in column `B`. Approximately what number do you get?

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.



	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	pmf	1 s.d.	2 s.d.	3 s.d.	
8	1	0.00002%				
9	2	0.00004%				

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($A8>C$5,$A8<C$6),$B8,"")`. With these mixed references applied, you can fill out the rest of the range through column `E`.

C36 =IF(AND($A36>C$5,$A36<C$6),$B36,"")

	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**	**pmf**	**1 s.d.**	**2 s.d.**	**3 s.d.**	
32	25	0.17528%			0.17528%	
33	26	0.22395%			0.22395%	
34	27	0.28327%			0.28327%	
35	28	0.35475%			0.35475%	
36	29	0.43984%			0.43984%	
37	30	0.53991%			0.53991%	
38	31	0.65616%		0.65616%	0.65616%	
39	32	0.78950%		0.78950%	0.78950%	
40	33	0.94049%		0.94049%	0.94049%	
41	34	1.10921%		1.10921%	1.10921%	
42	35	1.29518%		1.29518%	1.29518%	
43	36	1.49727%		1.49727%	1.49727%	
44	37	1.71369%		1.71369%	1.71369%	
45	38	1.94186%		1.94186%	1.94186%	
46	39	2.17852%		2.17852%	2.17852%	
47	40	2.41971%		2.41971%	2.41971%	
48	41	2.66085%	2.66085%	2.66085%	2.66085%	
49	42	2.89692%	2.89692%	2.89692%	2.89692%	
50	43	3.12254%	3.12254%	3.12254%	3.12254%	
51	44	3.33225%	3.33225%	3.33225%	3.33225%	
52	45	3.52065%	3.52065%	3.52065%	3.52065%	

a. Also sum the probabilities in columns `C:E`. What values do you get?
b. What happens when you increase/decrease the standard deviation and mean?

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 an area chart.

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.