The Standard Deviation and the Distribution of Data Values:

The *Empirical Rule* and *Tchebysheff's Theorem*

Suppose that a data set has mean $\overline{X}$ and standard deviation s. We're used to working with and interpreting the mean $\overline{X}$, but what does the value of the standard deviation s tell us? It's a measure of dispersion or variability in the data set, but we can be more specific than that. Here are two useful rules.

I. The empirical rule. The empirical rule tells us that *if* the data follows a normal distribution, then:

> Approximately 68% of the data values can be expected to lie within a one standard deviation interval around the mean, i.e. in the interval $\overline{X} \pm s$.
>
> Approximately 95% of the data values can be expected to lie within a two standard deviation interval around the mean, i.e. in the interval $\overline{X} \pm 2s$.
>
> Virtually all (approximately 99.73%) of the data values can be expected to lie within a three standard deviation interval around the mean, i.e. in the interval $\overline{X} \pm 3s$.

So for example if $\overline{X} = 50$ and $s = 8$, we'd expect about 68% of the data in the interval 50 ± 1(8) or (42, 58); we'd expect about 95% of the data in the interval 50 ± 2(8) or (34, 66); and we'd expect virtually all of the data in the interval 50 ± 3(8) or (26, 74). We now have information regarding the dispersion of the data around the mean.

If you're interested in intervals having widths other than one, two, or three standard deviations, you can use normal curve tables to find the appropriate percentages.

II. Tchebysheff's Theorem. The empirical rule is limited in that it only applies to data that follows (at least approximately) a normal distribution. Here is a rule, called Tchebysheff's Theorem, that applies to *any* shape distribution:

> For any value of k that is ≥ 1, *at least* $100(1 - 1/k^2)\%$ of the data will lie within k standard deviations of the mean.

This is a general formula that can be used with any value of k of interest, as long as k is at least one. For example, if we'd like to build a one standard deviation interval around the mean, apply the formula with $k = 1$. It says that at least

$$100(1 - 1/1^2)\% = 0\%$$

of the data will lie within one standard deviation of the mean.

That's useless information of course…. we already knew that! The weakness of Tchebysheff's theorem is that, since it must apply to any shape distribution, it can't be very specific about the percentage of data in any interval. Also, it only gives a lower bound for that percentage. In this case the lower bound provides no useful information.

Let's try a two standard deviation interval. Plugging $k = 2$ into the formula, we learn that at least

$$100(1 - 1/(2^2))\% = 75\%$$

of the data will lie within two standard deviations of the mean. That's much more useful: it says that for *any* distribution we have assurance that at least ¾ of the data is within two standard deviations. Once again, keep in mind that this is a lower bound: the actual percentage could be as low as 75% or as high as 100%. (We already know that if the distribution has a normal shape then the actual percentage is closer to 95%.)

Let's suppose once again that we're working with a data set that has a mean of $\overline{X}$ = 50 and a standard deviation of s = 8. Here's a table showing what we can say about the distribution of the data, using both the empirical rule and Tchebysheff's Theorem. For practice with the formula, you should verify the results shown in the Tchebysheff column at k = 1.5, 2.5, 3, and 4.

Standard Deviations (*k*)	**Interval**	**Tchebysheff %**	**Empirical Rule %**
1	50 ± 1(8) or (42, 58)	At least 0%	Approx. 68%
1.5	50 ± 1.5(8) or (38, 62)	At least 56%	
2	50 ± 2(8) or (34, 66)	At least 75%	Approx. 95%
2.5	50 ± 2.5(8) or (30, 70)	At least 84%	
3	50 ± 3(8) or (26, 74)	At least 89%	Virtually all (99.73%)
4	50 ± 4(8) or (18, 82)	At least 94%	

If the data appears to follow a normal distribution, then the empirical rule is preferred as it is more specific. Otherwise, the empirical rule *does not apply* and Tchebysheff's Theorem should be used.