Outlier Illustration

To illustrate the impact of an outlier on the mean, standard deviation, and variance, I generated normally distributed data with 10 cases, 50 cases, 100 cases, and 500 cases and then added an extreme value. The theoretical mean used was $\overline{Y}$ = 10 for all of the samples, and the extreme score added was 30 for each size data set.

n = 10 without outlier

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	10	7.37	12.39	10.1445	1.68141	2.827
Valid N (listwise)	10					

n = 10 plus outlier

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	11	7.37	30.00	11.9496	6.19552	38.384
Valid N (listwise)	11					

Difference in $\overline{Y}$ = 1.81

Difference in s = 4.51

Difference in s^2 = 35.56

n = 50 without outlier

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	50	8.28	14.26	10.7896	1.50846	2.275
Valid N (listwise)	50					

n = 50 plus outlier

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	51	8.28	30.00	11.1662	3.07669	9.466
Valid N (listwise)	51					

Difference in $\overline{Y}$ = .38

Difference in s = 1.57

Difference in s^2 = 7.19

n = 100 without outlier

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	100	3.87	14.72	9.9292	2.07397	4.301
Valid N (listwise)	100					

n = 100 plus outlier

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	101	3.87	30.00	10.1279	2.87173	8.247
Valid N (listwise)	101					

Difference in $\overline{Y}$ = .20

Difference in s = .80

Difference in s^2 = 3.95

n = 500 without outlier

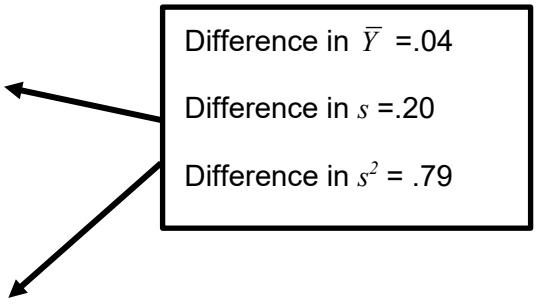
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	500	3.97	15.28	10.0099	1.88166	3.541
Valid N (listwise)	500					

n = 500 plus outlier

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
y	501	3.97	30.00	10.0498	2.08115	4.331
Valid N (listwise)	501					



The main point of this illustration is that the effect of a single outlier on the mean, standard deviation, and variance diminishes as the sample size increases. It is important to note that the outlier in my example is pretty extreme too, where the value of the outlier was three times the theoretical mean of the scores. Especially considering that I used a theoretical standard deviation of only 2 (and therefore a CV of only .2) when I generated the data, this outlier differs quite substantially from the other values in the sample. For most real data sets, it would probably be pretty rare to see such an extreme outlier, unless there was a coding error. This all suggests that, when we have a reasonably large sample size, the impact of outliers is much less troublesome that one might think. I am not saying that we should not be concerned about outliers—I just wish to make the case that our sample statistics do not perform that disastrously when we have a reasonable sample size.

Some caveats…One thing to keep in mind is that this is a good or best case scenario, in at least one respect--the distributions of scores in the samples I generated were all normal. For other distributions the effect of an outlier might be more (although it might also be less depending on the circumstances). I also examined just one outlier, not several. In the end, this is just one simple simulation and there are probably many other particular circumstances to consider when gauging the impact of outliers on the mean, standard deviation, or variance.