The Sample Max (or Min)

What It Might Be

By Stephen N. Luko

Q How do we judge what the sample max (or min) might be when we don't know it — assuming a normal distribution?

A Suppose that we are given a sample mean and standard deviation along with the sample size that was used in calculating these statistics. Suppose further that the data are a random draw from a normal distribution. Nothing more. The question is: "How can we judge what the largest (or smallest) value in the sample was?" The concern might be that too large a value might have escaped to the next operation or might actually have been shipped to a customer. Again, we don't have the raw data, just the sample mean, standard deviation, and sample size.

We can judge the sample max (or min) with knowledge of the sample size, the sample mean, and sample standard deviation where the sample is assumed to be a random draw from a normal distribution. Let n be the sample size, $\bar{y}$ the sample average, and S_y the sample standard deviation. Note that subscripts in parentheses used below denote order statistics.

1. Generate a large number of samples, each of size n from a $N(0,1)$ distribution; recommend at least 100,000 samples. Call these variables v.
2. Calculate and save the sample mean, standard deviation, and max of v for each sample. Note that $v_{(n)}$ is the largest value in the sample.
3. Compute the standardized sample max, as shown below, for each sample of n. Note that this is similar to the traditional Grubbs statistic, used for testing a single outlier in a sample from a normal distribution. In the Grubbs scenario, tables are usually available for upper critical values such as 95 and 99 percent. Here we are using the Grubbs method as a bounding for the largest sample value in n. The Grubbs criteria for outlier detection are given in Reference 1, which includes a table of critical values for the Grubbs statistic g for varying sample size (see table on next page). Further detail in reference to the Grubbs method can be found in the earlier 1950 paper, Reference 2.

$$g = (v_{(n)} - \bar{v}) / S_v$$

Save g for each sample.

4. The values of g constitute the distribution of the standardized distance from a sample average that the maximum value in a sample of n might be. For a given sample size, this distribution is identical for any normal parent – as may be easily shown. Determine two bounding empirical percentiles from the distribution of g, say k_1 and k_2 such that $P(k_1 < g < k_2) = C$, where C is the desired probability, say 95 or 99 percent. Figure 1 shows the distribution of g for $n = 10$, obtained using Monte Carlo simulation and 250,000 trials.
5. Substituting as $g = (y_{(n)} - \bar{y}) / s_y$ we have:

$$P(k_1 < (y_{(n)} - \bar{y}) / s_y < k_2) = C$$

It follows that:

$$P(\bar{y} + k_1 S_y < y_{(n)} < \bar{y} + k_2 S_y) = C$$

6. When we substitute the actual sample average and standard deviation we get a confidence interval for the sample max.

$$\bar{y} + k_1 S_y < y_{(n)} < \bar{y} + k_2 S_y$$

The one-sided lower or upper bound may be similarly determined.

Figure 1 —

Distribution of g for $n = 10$[1]

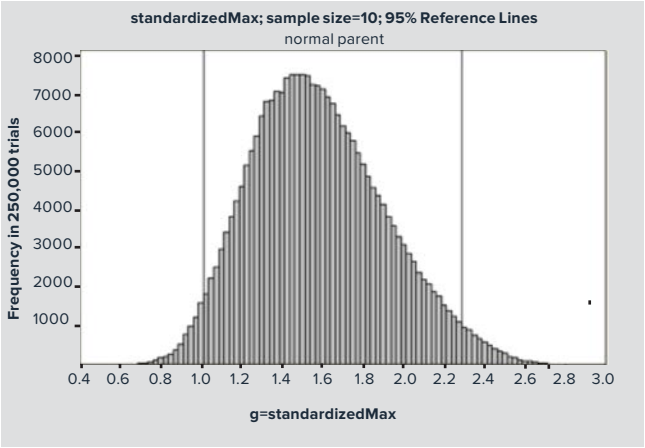

Table 1 —

Table of Critical Values for G (One-sided Test) When Standard Deviation is Calculated from the Same Sample[6]

Number of Observations n	5% Significance Level	2.5% Significance Level	1% Significance Level
3	1.15	1.15	1.15
4	1.46	1.48	1.49
5	1.67	1.71	1.75
6	1.82	1.89	1.94
7	1.94	2.02	2.10
8	2.03	2.13	2.22
9	2.11	2.21	2.32
10	2.18	2.29	2.41

EXAMPLE

Suppose for a sample of $n = 10$ the sample mean is 162 and standard deviation 12.4. Judge the sample largest value at 95 percent confidence.

Carrying out the routine summarized above, find that from the distribution of g that $k_1 = 1.011$ (the 2.5 percent point) and $k_2 = 2.288$ (the 97.5 percent point). The largest in $n = 10$ may then be bounded (using paragraph 5) with 95 percent confidence as:

$174.53 < \hat{y}_{(n)} < 190.37$ at 95 percent confidence.

A 90 or 99 percent interval may be similarly found. These may be shown to be:

$175.48 < \hat{y}_{(n)} < 188.98$; at 90 percent confidence

and $172.97 < \hat{y}_{(n)} < 192.80$; at 99 percent confidence.

Note that we are not predicting a future value since the data already exist, nor are we constructing an interval for a parameter, nor a tolerance interval for the whole distribution. This means that the interval we are constructing is not a classical prediction, confidence, or tolerance type interval. For details on these types of intervals that have appeared previously in this column, see References 3, 4, and 5. The sample maximum is an instance of a single unobserved random variable that one can make predictions about. We can use the Grubbs table for the upper bound of the sample max or lower bound for the sample min. The distribution of g is not symmetric. When Monte Carlo simulation is used, we need to use the actual distribution of g to figure the percentiles that are used for the chosen confidence level. It is interesting to note that very often people use the mean plus three standard deviations as an estimate for the maximum. In the example above that would be 199.2. Compare this to the actual intervals derived.

It is possible to do this type of analysis for other types of distributions, but it is worth remembering here that this article assumes that the distribution that was originally sampled was a normal distribution and that there were no outliers among the sample values.

REFERENCES

1. Grubbs, F.E., "Procedures for Detecting Outlying Observations in Samples, *Technometrics*, Vol. 11, No. 1, Feb. 1969, pp. 1-21.

2. Grubbs, F.E., "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics*, Vol. 21, 1950, pp. 27-28.

3. Luko, S.N., and Neubauer, D.V., Statistical Intervals: Nonparametric, Part 1, ASTM *Standardization News*, Vol. 41, No. 6, Nov./Dec. 2013, pp. 20-21.

4. Luko, S.N., and Neubauer, D.V., Statistical Intervals: Nonparametric, Part 2, ASTM *Standardization News*, Vol. 42, No. 1, Jan./Feb. 2014, pp. 12-13.

5. Luko, S.N., and Neubauer, D.V., Statistical Prediction Intervals, ASTM *Standardization News*, Vol. 42, No. 2, March/April 2014, 12-14.

6. Partial table from Grubbs, F.E., "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, Vol. 11, No. 1, Feb. 1969, reprinted by permission of the American Statistical Association (http://www.amstat.org).

Stephen N. Luko, UTC Aerospace Systems, Windsor Locks, Connecticut, is a past chairman of Committee E11 on Quality and Statistics, the current chair of Subcommittee E11.30 on Statistical Quality Control, and a fellow of ASTM International.

John Carson, Ph.D., of P&J Carson Consulting LLC, is the Data Points column coordinator. He is also vice chairman of E11.30 on Statistical Quality Control, part of E11 on Quality and Statistics, and a member of Committee E50 on Environmental Assessment, Risk Management, and Corrective Action.