Paper SD-07

Maximizing Confidence and Coverage for a Nonparametric Upper Tolerance Limit for a Fixed Number of Samples

Dennis J. Beal, Science Applications International Corporation, Oak Ridge, Tennessee

ABSTRACT

A nonparametric upper tolerance limit (UTL) bounds a specified percentage of the population distribution with specified confidence. The most common UTL is based on the largest order statistic (the maximum) where the number of samples required for a given confidence and coverage is easily derived for an infinitely large population. This relationship can be used to determine the number of samples prior to sampling to achieve a given confidence and coverage. However, often statisticians are given a data set and asked to calculate a UTL for the number of samples provided. Since the number of samples usually cannot be increased to increase confidence or coverage for the UTL, the maximum confidence and coverage for the given number of samples is desired. This paper derives the maximum confidence and coverage for a fixed number of samples. This relationship is demonstrated both graphically and in tabular form. The maximum confidence and coverage are calculated for several sample sizes using results from the maximization. This paper is for intermediate SAS® users of Base SAS® who understand statistical intervals.

Key words: upper tolerance limit, order statistics, sample size, confidence, coverage, maximization

INTRODUCTION

A one-sided distribution-free (nonparametric) upper tolerance limit (UTL) is equivalent to a one-sided distribution-free confidence bound for a percentile of that population. No distributional assumptions are necessary such as normality, lognormality, gamma or any other continuous distribution. However, the nonparametric UTL does assume the data collected are randomly selected from an infinitely large population, are statistically independent samples, and are statistically representative of the population.

UTLs have both a confidence and coverage attribution. The *coverage* of a UTL is the percentage p ($0 < p < 1$) of the population distribution that is bounded by the order statistic from the sample. The *confidence* of a UTL is how confident one is that the specified order statistic bounds the percentile of the population distribution and is denoted $100\text{x}(1 - \alpha)\%$ where α is the Type I error rate ($0 < \alpha < 1$). A Type I error (α) is the probability of rejecting the null hypothesis when in fact the null hypothesis is true. Once the confidence, coverage and desired order statistic are specified, the minimum number of samples (n) necessary to achieve these parameters can be calculated (Beal 2012). The SAS code uses the SAS® System for personal computers version 9.3 running on Windows® 7.

THEORY OF ORDER STATISTICS

A one-sided nonparametric UTL assuming an infinitely large population that relates confidence ($1 - \alpha$), coverage (p), and the number of samples (n) is shown in Equation (1) (Hahn and Meeker, 1991).

$$p = \alpha^{1/n} \quad (1)$$

For a fixed sample size n, the objective function to maximize is the sum of confidence and coverage, as shown in Equation (2).

$$f(\alpha) = 1 - \alpha + p \quad (2)$$

Substituting Equation (1) into Equation (2) yields Equation (3).

$$f(\alpha) = 1 - \alpha + \alpha^{1/n} \quad (3)$$

To maximize Equation (3), we take the first derivative of $f(\alpha)$ and set it equal to 0, as shown in Equation (4).

$$f'(\alpha) = -1 + \frac{\alpha^{\frac{1}{n}-1}}{n} = 0 \tag{4}$$

Solving Equation (4) for α yields Equation (5) for *n* > 1.

$$\alpha = n^{\frac{n}{1-n}} \tag{5}$$

α from Equation (5) maximizes Equation (3) as shown by Equation (6) for *n* > 1.

$$f''(\alpha) = \frac{(1-n)\alpha^{\frac{1-2n}{n}}}{n^2} < 0 \tag{6}$$

Therefore, the maximum confidence term from Equation (3) is shown in Equation (7).

$$1-\alpha = 1 - n^{\frac{n}{1-n}} \tag{7}$$

The maximum coverage term from Equation (3) is shown in Equation (8).

$$p = \alpha^{1/n} = n^{\frac{1}{1-n}} \tag{8}$$

GRAPHS OF THE OBJECTIVE FUNCTION

Figure 1 shows line plots of the function from Equation (3) to be maximized on the vertical axis with α on the horizontal axis. The function of confidence plus coverage is shown for selected number of samples (*n*) from 5 to 100 where the function is maximized. The top figure of Figure 1 shows the complete function for all α (0 < α < 1). The bottom figure shows the same curves, but magnifies the plot for only smaller values of α (0 < α < 0.20) since the functions are maximized within this range. The vertical lines show where the function is maximized for each *n*. Figure 1 shows as *n* increases the optimal α decreases. As α decreases, confidence increases and coverage decreases for any *n*.

Any combination of confidence and coverage along each line plot may be selected for each *n*. For example, for *n* = 10 one could choose 99% confidence (α = 0.01) with approximately 63% coverage. This would result in only 99% confidence + 63% coverage = 162% combined confidence and coverage. Selecting the optimal α = 0.0774 yields approximately 92.26% confidence and 77.43% coverage for a total of 169.7%. Table 1 shows the maximized confidence, maximized coverage and optimal α for various *n* using Equations (5), (7) and (8).

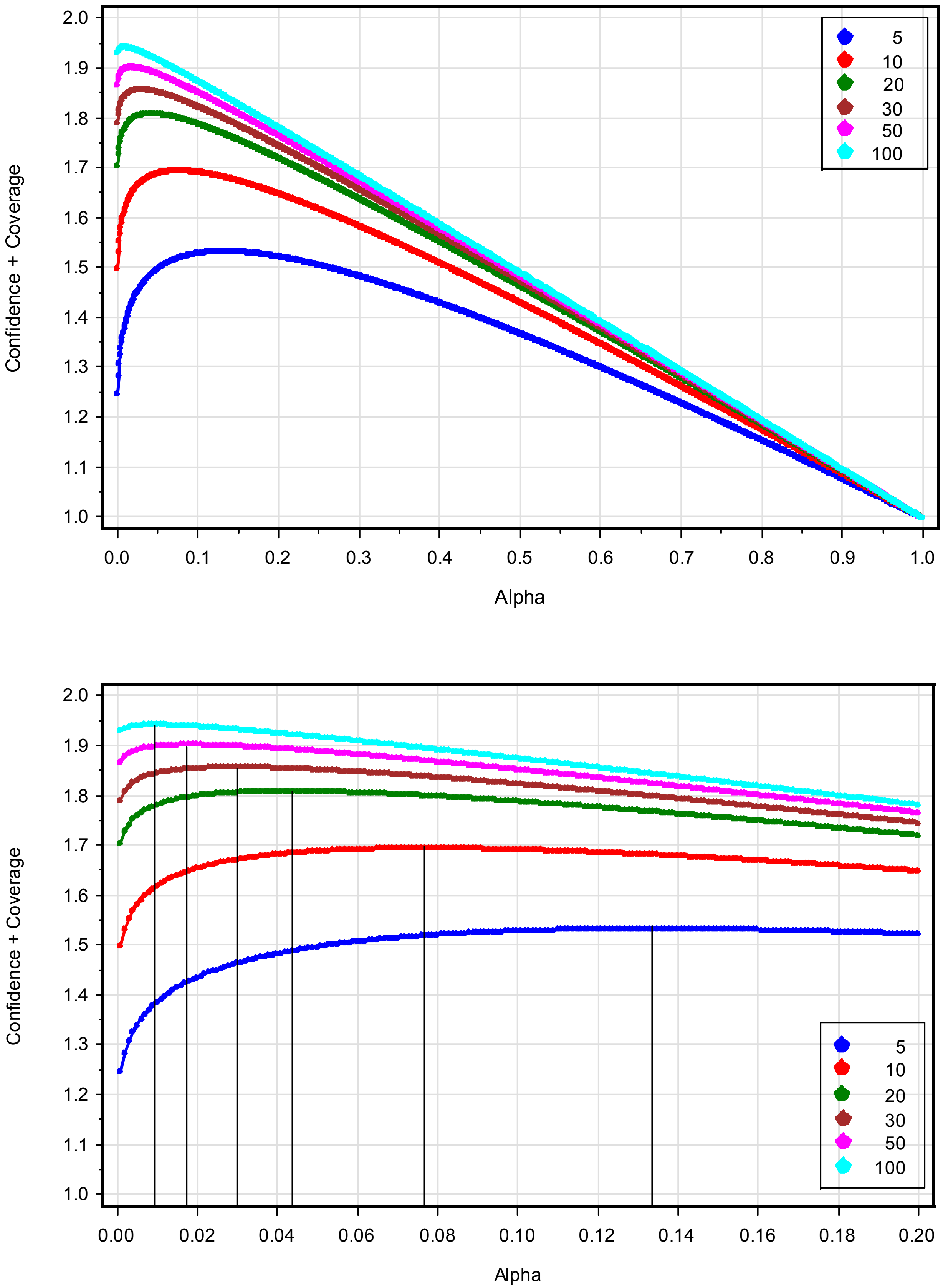

Figure 1. Line plots of confidence plus coverage objective function for *n* = 5, 10, 20, 30, 50, 100

Table 1. Optimal confidence and coverage for selected sample sizes

Sample Size (*n*)	Optimal α	Optimal Confidence (%)	Optimal Coverage (%)	Confidence + Coverage (%)
2	0.250	75.000	50.000	125.00
3	0.192	80.755	57.735	138.49
4	0.157	84.251	62.996	147.25
5	0.134	86.625	66.874	153.50
6	0.116	88.353	69.883	158.24
7	0.103	89.671	72.302	161.97
8	0.093	90.713	74.300	165.01
9	0.084	91.557	75.984	167.54
10	0.077	92.257	77.426	169.68
11	0.072	92.847	78.679	171.53
12	0.066	93.352	79.780	173.13
13	0.062	93.788	80.755	174.54
14	0.058	94.169	81.627	175.80
15	0.055	94.506	82.413	176.92
16	0.052	94.805	83.124	177.93
17	0.049	95.072	83.772	178.84
18	0.047	95.313	84.365	179.68
19	0.045	95.531	84.910	180.44
20	0.043	95.729	85.413	181.14
21	0.041	95.911	85.879	181.79
22	0.039	96.077	86.313	182.39
23	0.038	96.230	86.717	182.95
24	0.036	96.371	87.095	183.47
25	0.035	96.502	87.449	183.95
26	0.034	96.624	87.781	184.40
27	0.033	96.737	88.094	184.83
28	0.032	96.843	88.390	185.23
29	0.031	96.942	88.669	185.61
30	0.030	97.036	88.933	185.97
40	0.023	97.726	90.975	188.70
50	0.018	98.153	92.327	190.48
75	0.013	98.742	94.332	193.07
100	0.010	99.045	95.455	194.50

SAS CODE FOR GENERATING GRAPHS

The SAS code that calculates the curves for the confidence plus coverage objective function is shown below.

```
data a;
  do n = 5, 10, 20, 30, 50, 100;
    alpha_opt = n**(n/(1-n));  ** equation 5;
    conf_opt = 1 - alpha_opt;  ** equation 7;
    p_opt = n**(1/(1-n));      ** equation 8;
    do alpha = 0.001 to 0.999 by 0.001;
      fn = 1 - alpha + alpha**(1/n);  ** equation 3;
       output;
       end;
    end;
  output;
  run;
```

CONCLUSION

For a given data set with fixed number of samples *n*, the confidence and coverage can be selected for a nonparametric UTL on the maximum result, assuming an infinitely large population from which the representative samples are drawn. However, for small samples there is insufficient data to achieve both high confidence and high coverage. An increase in confidence will cause a decrease in coverage, while an increase in coverage will cause a decrease in confidence. This paper derives the equations to calculate the maximum confidence and coverage for any $n > 1$. This relationship is demonstrated both graphically and in tabular form for various values of *n*. These results allow the data analyst to obtain the maximum confidence and coverage for a nonparametric UTL from any data set.

REFERENCES

Beal, Dennis J. 2012. "Sample Size Determination for a Nonparametric Upper Tolerance Limit for any Order Statistic," *Proceedings of the 20th Annual Conference of the SouthEast SAS Users Group*.

Hahn, G. and W. Meeker. 1991. *Statistical Intervals: A Guide for Practitioners*. 91-92. New York, New York: John Wiley & Sons, Inc.

CONTACT INFORMATION

The author welcomes and encourages any questions, corrections, feedback, and remarks. Contact the author at:

Dennis J. Beal, Ph.D.
Senior Statistician / Risk Scientist
Science Applications International Corporation
301 Laboratory Road
Oak Ridge, Tennessee 37831
phone: 865-481-8736
e-mail: beald@saic.com