Paper SD-93

Maximizing Confidence and Coverage for a Nonparametric Upper Tolerance Limit on the Second Largest Order Statistic for a Fixed Number of Samples

Dennis J. Beal, Leidos, Oak Ridge, Tennessee

ABSTRACT

A nonparametric upper tolerance limit (UTL) bounds a specified percentage of the population distribution with specified confidence. The confidence and coverage of a UTL based on the second to largest order statistic is evaluated 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 second largest order statistic 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 the second largest order statistic 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, objective function

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 so statistical inferences can be applied to 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 $100(1 - \alpha)\%$ 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 9.4 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) on the largest order statistic (maximum) is shown in Equation (1) (Hahn and Meeker, 1991).

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

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

$$f(\alpha, p) = 1 - \alpha + p \tag{2}$$

LARGEST ORDER STATISTIC

For the largest order statistic, Beal (2013) showed that Equation (3) maximizes the confidence for a given $n > 1$.

$$1 - \alpha = 1 - n^{\frac{n}{1-n}} \tag{3}$$

For the largest order statistic, Beal (2013) also showed that Equation (4) maximizes the coverage for a given $n > 1$.

$$p = \alpha^{1/n} = n^{\frac{1}{1-n}} \tag{4}$$

SECOND TO LARGEST ORDER STATISTIC

Hahn and Meeker (1991) shows that the equation that relates confidence with coverage for the second largest order statistic is given by Equation (5).

$$1-\alpha = 1-np^{n-1}+(n-1)p^{n} \tag{5}$$

Equation (5) cannot be solved for p as a function of α and n as was done for the largest order statistic in Equation (1). Although the confidence and coverage that maximizes Equation (2) cannot be obtained analytically for the second largest order statistic, Equation (2) can be modified for the second largest order statistic by substituting Equation (5) into Equation (2), as shown in Equation (6).

$$f(p) = 1-np^{n-1}+(n-1)p^{n}+p \tag{6}$$

SAS CODE

To maximize Equation (6) for the second largest order statistic, $f(p)$ is calculated for various values of p for fixed values of n as shown in the following SAS code.

```
data a;
  do n = 3 to 30 by 1, 40, 50, 75, 100;
   do p = 0.001 to 0.999 by 0.001;
    conf = 1 - n*p**(n-1) + (n-1)*p**n;
    sum = conf + p;
    output;
    end;
  end;
  label p = 'Coverage (p)'
     conf = 'Confidence'
      sum = 'Confidence + Coverage = f(p)';
run;
```

The data are then sorted by n and descending sum of confidence and coverage so the maximum sum for each n is retained in a separate data set.

```
proc sort data=a; by n descending sum;

data b;
  set a;
    by n;
  if first.n;
run;
```

GRAPH OF THE OBJECTIVE FUNCTION

Figure 1 shows line plots of the function from Equation (6) to be maximized on the vertical axis with the coverage (p) on the horizontal axis. The function $f(p)$ of confidence plus coverage for the second largest order statistic is shown for selected number of samples (n) of 10, 20, 50 and 100. Figure 1 shows the complete function for all p ($0 < p < 1$). Figure 1 shows as n increases the optimal coverage p increases as well as confidence. For these n the sum function $f(p)$ is identical until p is approximately 0.45.

Any combination of confidence and coverage along each line plot may be selected for each n. For example, for n = 50 one could choose 95% coverage with approximately 72% confidence. This would result in only 95% coverage + 72% coverage = 167% combined confidence and coverage. Selecting the optimal coverage p = 0.89 (89%) yields approximately 97.884% confidence for a total of 186.884%.

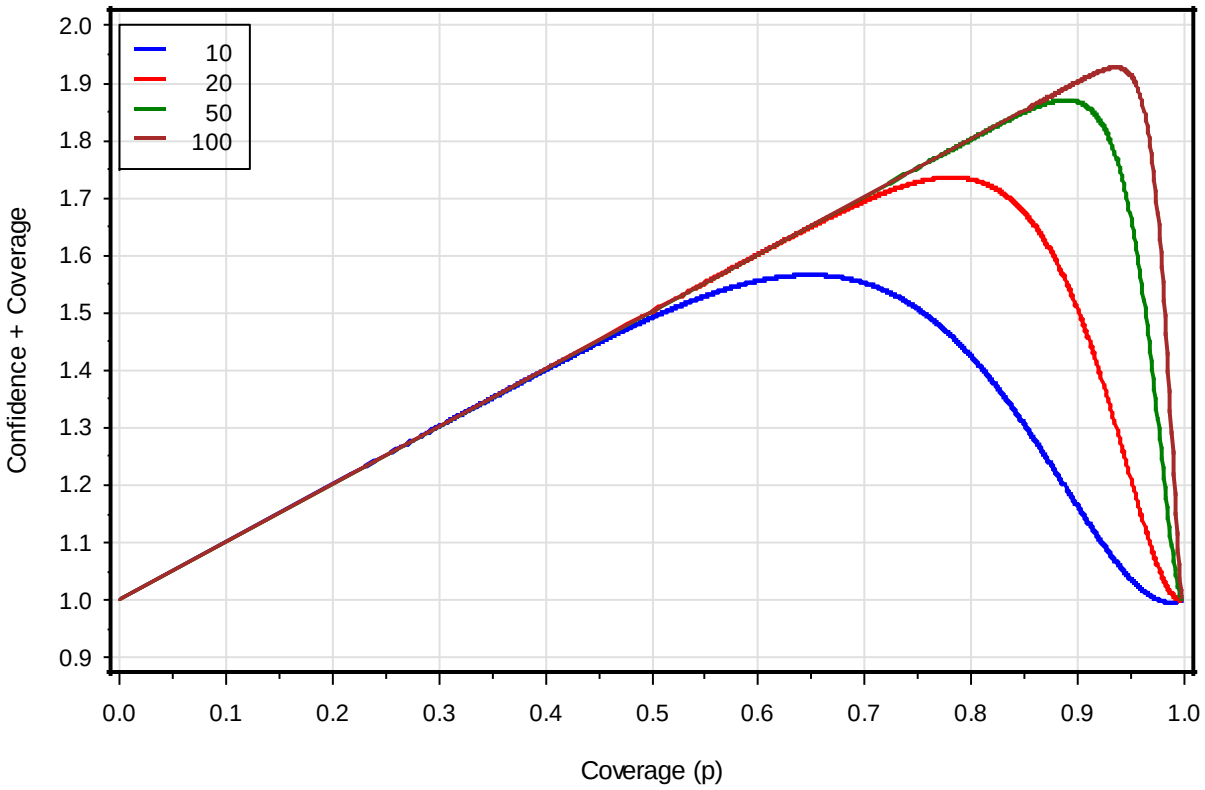

Figure 1. Line plots of confidence plus coverage objective function for *n* = 10, 20, 50, 100

RESULTS

Table 1 shows the maximized confidence and coverage for various *n* using Equations (5) and (6). For $n = 3$ the optimal confidence is approximately 88.5% while the optimal coverage is only 21.1% for a combined confidence and coverage of only 109.6%. Optimal coverage increases monotonically as *n* increases for all sample sizes. However, optimal confidence decreases slightly from $n = 3$ to $n = 4$ and then increases monotonically as *n* increases for $n \geq 4$. Clearly, as *n* increases, the sum of both confidence and coverage increases monotonically as expected.

As an example of an application using Table 1, suppose the data analyst is provided a data set with $n = 25$ from an infinitely large population. If additional samples cannot be collected in order to increase confidence or coverage, what is the maximum amount of confidence and coverage allowed from these $n = 25$ samples for the second largest order statistic? Table 1 shows the optimal confidence is 95.988% with an optimal coverage of 81.5% for a combined confidence and coverage of 177.49%. Selecting arbitrarily 90% confidence would yield 85.3% coverage. However, this would not be optimal because the combined confidence and coverage is only 175.3%.

Table 1. Optimal confidence and coverage for selected sample sizes *n*

Sample Size (*n*)	Optimal Confidence (%)	Optimal Coverage (%)	Confidence + Coverage (%)
3	88.522	21.1	109.622
4	86.277	36.1	122.377
5	86.878	45.0	131.878
6	87.997	51.1	139.097
7	89.076	55.7	144.776
8	89.976	59.4	149.376
9	90.787	62.4	153.187
10	91.405	65.0	156.405
11	92.063	67.1	159.163
12	92.556	69.0	161.556
13	93.057	70.6	163.657
14	93.416	72.1	165.516
15	93.776	73.4	167.176
16	94.168	74.5	168.668
17	94.417	75.6	170.017
18	94.643	76.6	171.243
19	94.963	77.4	172.363
20	95.091	78.3	173.391
21	95.338	79.0	174.338
22	95.514	79.7	175.214
23	95.726	80.3	176.026
24	95.882	80.9	176.782
25	95.988	81.5	177.488
26	96.148	82.0	178.148
27	96.267	82.5	178.767
28	96.348	83.0	179.348
29	96.496	83.4	179.896
30	96.613	83.8	180.413
40	97.443	86.9	184.343
50	97.884	89.0	186.884
75	98.553	92.0	190.553
100	98.949	93.6	192.549

CONCLUSION

For a given data set with fixed number of samples n, the confidence and coverage can be selected for a nonparametric UTL on any order statistic assuming an infinitely large population from which the representative samples are drawn. However, for small samples there are 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 calculates the maximum confidence and coverage for any $n > 2$ on the second to largest order statistic for a nonparametric UTL. 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 on the second largest order statistic 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*.

Beal, Dennis J. 2013. "Maximizing Confidence and Coverage for a Nonparametric Upper Tolerance Limit for a Fixed Number of Samples," *Proceedings of the 21st 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
Leidos
301 Laboratory Road
Oak Ridge, Tennessee 37831
phone: 865-481-8736
e-mail: beald@leidos.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.