Paper SD-07

Sample Size Determination for a Nonparametric Upper Tolerance Limit for any Order Statistic

Dr. Dennis 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. However, for other order statistics such as the second largest, third largest, etc., the equations used to determine the number of samples to achieve a specified confidence and coverage become more complex using the incomplete Beta function as the order statistic decreases from the maximum. This paper uses the theory of order statistics to derive the equations from the incomplete Beta distribution for calculating the sample size for a one-sided nonparametric UTL using any order statistic. SAS® code is shown that performs these calculations in a single macro. The number of samples required for various order statistics is compared for the incomplete Beta function, the normal approximation to the binomial and the binomial distribution. Examples of SAS code are shown for each method. The binomial distribution is shown to be the most accurate for calculating the proper order statistic for any number of samples. This paper is for intermediate SAS users of Base SAS who understand statistical intervals, statistical distributions and SAS macros.

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

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. Since it is nonparametric, no distributional assumptions are necessary such as normality, lognormality, gamma or any other continuous distribution. 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* 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 100x(1 - α)% where α is the Type I error rate (0 < α < 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 using the SAS code presented in this paper. For example, if α = 0.05 and *p* = 0.90 using the largest order statistic from *n* = 29 samples, then we would be 95% confident that the maximum from the 29 samples bounds at least 90% of the continuous population distribution. 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 uses an incomplete Beta function described in Beyer (Beyer 1966). The equation to solve for the number of samples (*n*) is shown in Equation 1.

$$\alpha \geq \frac{\Gamma(n+1)}{\Gamma(u)\Gamma(n+1-u)} \int_0^p x^{u-1}(1-x)^{n-u}\,dx \qquad (1)$$

where u = the order statistic of interest ($u = n$ for the maximum, $u = n - 1$ for the second to maximum, etc.),
$\Gamma(n + 1) = n!$,
α = Type I error rate (0 < α < 1),
p = coverage (0 < p < 1)

LARGEST ORDER STATISTIC

Therefore, Equation 1 reduces to Equation 2 for the largest or maximum concentration where $u = n$.

$$\alpha \geq n\int_0^p x^{n-1}dx \tag{2}$$

Integrating Equation 2 yields Equation 3.

$$\alpha \geq p^n \tag{3}$$

Solving Equation 3 for *n* yields Equation 4 using the maximum for the one-sided UTL

$$n \geq \frac{\ln \alpha}{\ln p} \tag{4}$$

For example, if α = 0.05 and *p* = 0.90, then a minimum of *n* = 29 samples from an infinitely large population are needed for a one-sided nonparametric UTL using the maximum order statistic from Equation 4. The maximum from the 29 samples bounds at least 90% of the population distribution with 95% confidence. Equation 4 is also shown and derived in Hahn and Meeker (1991).

SECOND LARGEST ORDER STATISTIC

However, suppose one suspects there is a high likelihood that an outlier could be part of the 29 samples. Including an outlier would force the outlier as the maximum to be the one-sided nonparametric UTL. This could underestimate the percentage of the population distribution that the outlier bounds with 95% confidence. Therefore, we want to calculate the number of samples required so the second largest order statistic can be used as the UTL in the event a single outlier is present in the sample.

Using Equation 1 where *u* = *n* – 1 for the second largest order statistic, Equation 1 reduces to Equation 5.

$$\alpha \geq n(n-1)\int_0^p x^{n-2}(1-x)\ dx \tag{5}$$

Integrating Equation 5 yields Equation 6 as shown in Hahn and Meeker (1991).

$$\alpha \geq np^{n-1} - (n-1)p^n \tag{6}$$

However, there is no closed form solution for solving Equation 6 for *n* as a function of *p* and α. Therefore, after specifying *p* Equation 6 can be solved using any number of standard analytical techniques. The easiest method is to insert the function from Equation 6 into a SAS do loop that increments *n* by one and evaluating Equation 6 until *n* is found so that the right hand side of Equation 6 is ≤ α. Using this technique shows that *n* = 46 is the minimum sample size that causes Equation 6 to be true for α = 0.05 for *p* = 0.90. Therefore, if the sample size increases from *n* = 29 to *n* = 46, then the second to largest result in the sample of 46 is the one-sided nonparametric UTL instead of the maximum order statistic in the event of a single outlier with 95% confidence and 90% coverage. Equations 4 and 6 can be used for any confidence 100x(1 – α)% and coverage *p*.

THIRD LARGEST ORDER STATISTIC

Suppose one suspects there could be at most two outliers that could be part of the 46 samples. Then we want to calculate the number of samples required so the third largest order statistic can be used as the UTL in the event two outliers are present in the sample.

Using Equation 1 where *u* = *n* – 2 for the third largest order statistic, Equation 1 reduces to Equation 7.

$$\alpha \geq \frac{n(n-1)(n-2)}{2!}\int_0^p x^{n-3}(1-x)^2\,dx \tag{7}$$

Integrating Equation 7 yields Equation 8.

$$\alpha \geq \frac{n(n-1)(n-2)}{2!}\left[\frac{p^{n-2}}{n-2}-\frac{2p^{n-1}}{n-1}+\frac{p^{n}}{n}\right] \tag{8}$$

There also is no closed form solution for solving Equation 8 for *n* as a function of *p* and α. Therefore, after specifying *p* Equation 8 can be solved using SAS by incrementing *n* by one and evaluating Equation 8 until *n* is found so that the right hand side of Equation 8 is ≤ α. Using this technique shows that *n* = 61 is the minimum sample size that causes Equation 8 to be true for α = 0.05 for *p* = 0.90. Therefore, if the sample size increases from *n* = 46 to *n* = 61, then the third largest result in the sample of 61 is the one-sided nonparametric UTL. Equations 4, 6 and 8 can be used for any confidence 100x(1 – α)% and coverage *p*.

FOURTH LARGEST ORDER STATISTIC

Suppose we want to calculate the number of samples required so the fourth largest order statistic can be used as the UTL in the event three outliers are present in the sample.

Using Equation 1 where *u* = *n* – 3 for the fourth largest order statistic, Equation 1 reduces to Equation 9.

$$\alpha \geq \frac{n(n-1)(n-2)(n-3)}{3!}\int_0^p x^{n-4}(1-x)^3\,dx \tag{9}$$

Integrating Equation 9 yields Equation 10.

$$\alpha \geq \frac{n(n-1)(n-2)(n-3)}{3!}\left[\frac{p^{n-3}}{n-3}-\frac{3p^{n-2}}{n-2}+\frac{3p^{n-1}}{n-1}-\frac{p^{n}}{n}\right] \tag{10}$$

There also is no closed form solution for solving Equation 10 for *n* as a function of *p* and α. Therefore, after specifying *p* Equation 10 can be solved using SAS by incrementing *n* by one and evaluating Equation 10 until *n* is found so that the right hand side of Equation 10 is ≤ α. Using this technique shows that *n* = 76 is the minimum sample size that causes Equation 10 to be true for α = 0.05 for *p* = 0.90. Therefore, if the sample size increases from *n* = 61 to *n* = 76, then the fourth largest result in the sample of 76 is the one-sided nonparametric UTL. Equations 4, 6, 8 and 10 can be used for any confidence 100x(1 – α)% and coverage *p*.

ANY ORDER STATISTIC

Using mathematical induction we can derive the equation used to derive *n* for any $(r+1)^{th}$ observation from the maximum for 0 ≤ *r* ≤ *n* - 1. So *r* = 0 corresponds to the largest order statistic (maximum), *r* = 1 is the second largest order statistic, etc.

Using Equation 1 where *u* = *n* – *r* for the $(r+1)^{th}$ largest observation, Equation 1 reduces to Equation 11.

$$\alpha \geq \frac{1}{r!}\prod_{i=0}^{r}(n-i)\int_0^p x^{n-r-1}(1-x)^r\,dx \tag{11}$$

Integrating Equation 11 yields Equation 12.

$$\alpha \geq \frac{1}{r!}\prod_{i=0}^{r}(n-i)\left[\sum_{i=0}^{r}(-1)^i\left(\frac{r!}{i!(r-i)!}\right)\frac{p^{n-i}}{n-i}\right] \tag{12}$$

Clearly there is no closed form solution for solving Equation 12 for *n* as a function of *p* and α. Therefore, after specifying *p* Equation 12 can be solved using SAS by incrementing *n* by one and evaluating Equation 12 until the right hand side of Equation 12 is ≤ α. Equations 4, 6, 8, 10 and 12 can be used for any confidence 100x(1 – α)% and coverage *p*.

SAS CODE FOR EQUATION 12

The SAS code that solves Equation 12 for any α, *p*, *n* and *r* is shown in the SAS macro UTL below.

```
%macro utl(NUM); ** order statistic (1=max, 2=second to max, 3=third to max, etc.) ;
data a&NUM;
  NUM = &NUM;
  do p = 0.95; * percent coverage desired;
    do alpha = 0.05;
      CONF = (1 - alpha) * 100;  ** confidence as an integer;
      n = num-1;
      fn = 1;  ** initialize f(n) = 1;
      do until (n=2000);
        n+1;
        %do t = 1 %to &NUM;
         T&t = (-1)**(&t+1) * comb(num-1, &t.-1) * p**(&t.-1) / (n - num + &t.);
         %end;
        fn = comb(n, n-num)*num*p**(n-num+1)*sum(of t1-t&num) - alpha; ** Eqn. 12 ;
        output;
        end;
      output;
    end;
  end;
  proc print data=a&NUM; title "&NUM";
run;
%mend utl;
%utl(1)
```

NORMAL APPROXIMATION TO THE BINOMIAL

In practice, Equation 12 has been shown to be overly conservative for estimating the UTL when *n* is large. Equation 12 is best used when $np < 5$ or $n(1-p) < 5$. Equation 12 solves correctly for *n* for $r \leq 9$, but the function has multiple roots close together for $r > 9$, making it difficult to choose the proper value of *n*.

When *n* is large enough, the normal approximation to the binomial distribution provides a more accurate order statistic than Equation 12. When $np \geq 5$ and $n(1-p) \geq 5$, Equation 13 (U.S. Environmental Protection Agency 2010) can be used to determine the proper order statistic *k* ($1 \leq k \leq n$) for a one-sided nonparametric UTL.

$$k = np + z_{1-\alpha}\sqrt{np(1-p)} + 0.5 \qquad (13)$$

The $z_{(1-\alpha)}$ term in Equation 13 is the deviate from the standard normal distribution associated with a 100x(1 – α)% one-sided confidence interval. For example, when α = 0.05 for a 95% confidence interval, $z_{0.95} = 1.645$. The 0.5 term in Equation 13 is included as a correction factor as the continuous normal distribution approximates the discrete binomial distribution.

SAS CODE FOR EQUATION 13

The SAS code that calculates the order statistics *k* from Equation 13 using the normal distribution to approximate the binomial distribution for any α, *p*, and *n* is shown below.

```
data a;
  p = 0.95;
  z = probit(p);
  do n = 10 to 4000;                    ** n = number of samples;
    k =  n*p + z*sqrt(n*p*(1-p)) + 0.5;  ** k = order statistic;
    r = n - k;                          ** r = observations below the maximum;
    output;
   end;
   proc print data=a; run;
```

THE BINOMIAL DISTRIBUTION

The order statistic k ($1 \le k \le n$) can be calculated directly and exactly from the binomial distribution. The cumulative binomial distribution is shown in Equation 14.

$$1-\alpha \le \sum_{i=0}^{k} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \qquad (14)$$

Equation 14 is used to calculate the smallest order statistic k such that the cumulative binomial distribution equals or exceeds the confidence coefficient 1-α.

SAS CODE FOR EQUATION 14

The SAS code that calculates the exact order statistics k from Equation 14 using the cumulative binomial distribution for any α, p, and n is shown below.

```
data b;
  p = 0.95;  ** coverage ;
  conf = 0.95; ** confidence as a percent;
  do n = 20 to 2000;  ** n = number of samples;
    do k = 0 to n;    ** k = order statistic (1 ≤ k ≤ n);
      r = n - k;      ** r = number of observations below maximum (0 ≤ r ≤ n-1);
      prob = probbnml(p, n, k);  ** cumulative binomial distribution;
      if prob >= conf then do;
       output;
       goto done;
       end;
      end;
    done:
   end;

 proc print data=b; run;
```

RESULTS

The results from implementing the SAS code from Equations 12, 13 and 14 for α = 0.05 (95% confidence) and p = 0.95 (95% coverage) are shown in Table 1 where r = number of observations below the maximum ($0 \le r \le 30$). For example, r = 0 is the maximum, r = 1 is the second largest order statistic, etc. Table 1 shows the incomplete Beta function (Eqn. 12) and the binomial distribution (Eqn. 14) agree exactly for $r \le 9$, while the normal approximation (Eqn. 13) ranges from 7 to 11 samples higher for $r \le 9$ (the 10^{th} largest order statistic). For r > 9, the incomplete Beta function diverges much higher from both the normal approximation and the binomial, causing the incomplete Beta function to be overly conservative by selecting a much higher order statistic than is necessary to bound the 95^{th} percentile with 95% confidence. The normal approximation consistently requires 6 or 7 more samples than the binomial for r > 9, but requires fewer samples than the incomplete Beta.

Figure 1 shows the normal approximation to the binomial distribution plots consistently slightly above the binomial distribution for $0 \le r \le 30$. However, the incomplete Beta function diverges from both the normal approximation to the binomial and the binomial beginning at r = 10. The divergence increases as r increases.

r	Minimum *n** using Incomplete Beta (Eqn. 12)	Minimum *n** using Normal (Eqn. 13)	Minimum *n** using Binomial (Eqn. 14)
0	59	70	59
1	93	103	93
2	124	133	124
3	153	161	153
4	181	189	181
5	208	216	208
6	234	242	234
7	260	268	260
8	286	293	286
9	311	318	311
10	345	343	336
11	436	368	361
12	577	392	386
13	745	417	410
14	840	441	434
15	1007	465	458
16	1151	489	482
17	1287	513	506
18	1418	536	530
19	1535	560	554
20	1682	584	577
21	1799	607	601
22	1916	630	624
23	2058	654	647
24	2186	677	671
25	2293	700	694
26	2421	723	717
27	2559	746	740
28	2679	769	763
29	2788	792	786
30	2939	815	809

* Number of samples assuming 95% confidence with 95% coverage

Table 1. Minimum Sample Sizes for One-Sided Nonparametric UTLs

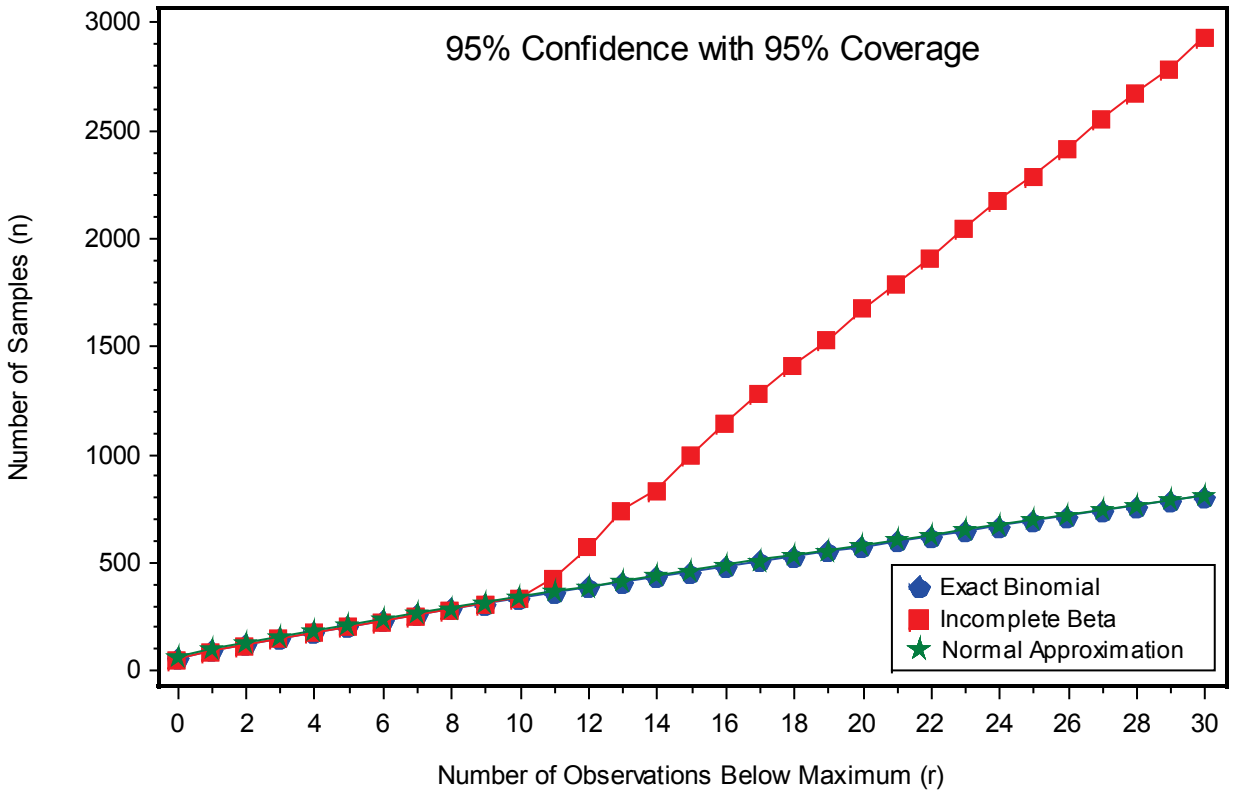

Figure 1. Number of Samples (*n*) with the Number of Observations Below the Maximum (*r*) by Method

CONCLUSION

While the most commonly used one-sided nonparametric UTL is based on the largest order statistic (the maximum), the incomplete Beta function can be used with other order statistics such as the second largest, third largest, etc., to determine the number of samples to achieve a specified confidence and coverage. The equation to use for any order statistic was derived in general for the incomplete Beta function. SAS code was presented in a single macro to calculate the number of samples required for specified confidence and coverage for any order statistic. The incomplete Beta function performed well for the 10 largest order statistics, but provided overly conservative estimates beginning with the 11th largest order statistic. The normal approximation to the binomial consistently requires 6 or 7 more samples than the cumulative binomial distribution. SAS code was presented to calculate the order statistics for the incomplete Beta, normal approximation to the binomial and the binomial distribution. Since the cumulative binomial distribution can be calculated easily in SAS, the binomial distribution has been shown to be the preferred method for calculating the proper order statistic for any number of samples. Examples of calculations using the SAS code for the three methods were shown for the 31 largest order statistics for 95% confidence and 95% coverage.

REFERENCES

Beyer, W. 1966. *Handbook of Tables for Probability and Statistics*. 251. Boca Raton, Florida: CRC Press, Inc.

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

U.S. Environmental Protection Agency (May 2010). *ProUCL Version 4.1.00 Technical Guide: Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations*. 88. (EPA/600/R-07/041). Washington, DC

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
151 Lafayette Drive
Oak Ridge, Tennessee 37831
phone: 865-481-8736
e-mail: beald@saic.com