

# Simple Nonparametric Upper and Lower Tolerance Bounds Based on Order Statistics

Bradley A. Warner

James H. Rutledge

HQ USAFA/DFMS  
2354 Fairchild Drive, Suite 6D2A  
USAF Academy, CO. 80840

## Summary

Confidence intervals for population parameters such as mean and variance are a common statistical tool. These limits are used to make conclusions about population parameters. However, sometimes we are interested in developing limits for individual observations from a population. A tolerance interval determines the proportion of the population that falls within the interval. A simple example of this distinction is the landing distance of an aircraft on a runway. An upper confidence bound for the mean would give us an upper limit on the average stopping distance of all landings. A tolerance interval, on the other hand, determines an upper bound on the distance in which some percentage of individual landings will not exceed. The first establishes limits on the average landing distance, while the second develops limits for the proportion of all individual landings.

Using order statistics, we will discuss a distribution free method to develop upper and lower tolerance intervals based on any one of the order statistics. These limits relate to the proportion of the population that exceed or fall under the particular order statistic. It turns out that the distribution of this proportion is a Beta distribution and for the case of the lowest order statistic, the cumulative distribution function has a relatively simple form. This procedure will be applied to a real data set.

## 1. Introduction

Confidence intervals are a popular statistical technique for establishing interval estimates for parameters of a population. For example, a 95% upper confidence bound for the mean establishes an upper bound for the population mean. However, sometimes we are interested in bounds for individual observations of a population and not its parameters. For example, suppose a tire manufacturer claimed their tires had a minimum life of 30,000 miles. The

manufacturer is claiming that each tire from the population of tires will last more than 30,000 miles, which is much different from claiming that the mean life of the tires will exceed 30,000 miles. Tolerance intervals place interval estimates around individual observations.

If we knew the observations came from a normal distribution and we knew the mean and standard deviation of this distribution, then, back to the tire example, we could say that 97.5% of the tires will have a life greater than  $\mu - 1.96\sigma$  miles. Often we do not know the parameters of the distribution or even the distribution itself. A tolerance interval uses sample statistics in place of the population parameters. Since the sample statistics are only estimates of the population parameters, the interval developed from them will have some uncertainty in the proportion  $p$  of the population covered by the interval. Therefore, we assign a confidence level to the coverage of the interval. The interpretation of the tolerance interval is that we are  $(1 - \alpha)\%$  confident that tolerance interval contains at least a proportion  $p$  of the population.

Unfortunately, in the preceding discussion we assumed that the population was normally distributed. In practice, this is often unknown. The appeal of confidence intervals for population parameters is that often we can use the central limit theorem to conclude the distribution of the sample mean is approximately normal even though we do not know the distribution of the parent population. With the use of order statistics, we will demonstrate a method to develop an upper or lower tolerance bound. These bounds will not depend on the distribution of the parent population.

## 2. Distribution of the Proportion of the Population Exceeding the $r^{th}$ Order Statistic

To free ourselves of the requirement to know the distribution of the parent population to develop a tolerance interval, we will use order statistics. The  $r^{th}$  order statistic is simply the  $r^{th}$  smallest number in our sample. David (1981) gives the distribution of the  $r^{th}$  order statistic as

$$g(y_r) = \frac{n!}{(r-1)!(n-r)!} \left[ \int_{-\infty}^{y_r} f(x) dx \right]^{r-1} f(y_r) \left[ \int_{y_r}^{\infty} f(x) dx \right]^{n-r} \quad (1)$$

where  $n$  is the sample size and  $f(x)$  is the probability density function of the random variable  $X$ .

To develop a tolerance interval, define  $p$  as the proportion of the population that is greater than the  $r^{th}$  order statistic,  $y_r$ :

$$p = \int_{y_r}^{\infty} f(x) dx \quad (2)$$

Using Leibnitz's rule (see Casella and Berger (1990)) and Equation 2, Equation 1 is transformed into the probability density of  $p$

$$g(p) = \frac{n!}{(r-1)!(n-r)!} [1-p]^{r-1} [p]^{n-r} \quad 0 \leq p \leq 1 \quad (3)$$

Note that Equation 3 is a Beta distribution with parameters  $n-r+1$  and  $r$ . Also, notice that the distribution of  $p$  does not depend on the original distribution of  $X, f(x)$ . Equation 3 is the probability density function of the proportion of the population that exceed the  $r^{th}$  order statistic. Thus, to find the proportion  $p$  of the population that exceeds the  $r^{th}$  order statistic with  $(1-\mathbf{a})\%$  confidence, we must solve the following equation for  $\beta$

$$\int_b^1 g(p) dp = 1 - \mathbf{a}$$

This is equivalent to

$$\int_0^b g(p)dp = \mathbf{a} \quad (4)$$

Equation 4 is the cumulative Beta distribution with parameters  $n-r+1$  and  $r$ . The solution for the proportion of the population exceeding the  $r^{\text{th}}$  order statistic with  $(1-\mathbf{a})\%$  confidence is simply the inverse of the cumulative Beta distribution, written **InverseBeta(a ,n-r+1,r)**.

For the case of the first order statistic ( $r = 1$ ), Equation 3 simplifies to

$$g(p) = np^{n-1} \quad 0 \leq p \leq 1 \quad (5)$$

For the first order statistic, Equation 5 is substituted into Equation 4 and, after integration, has the simple form

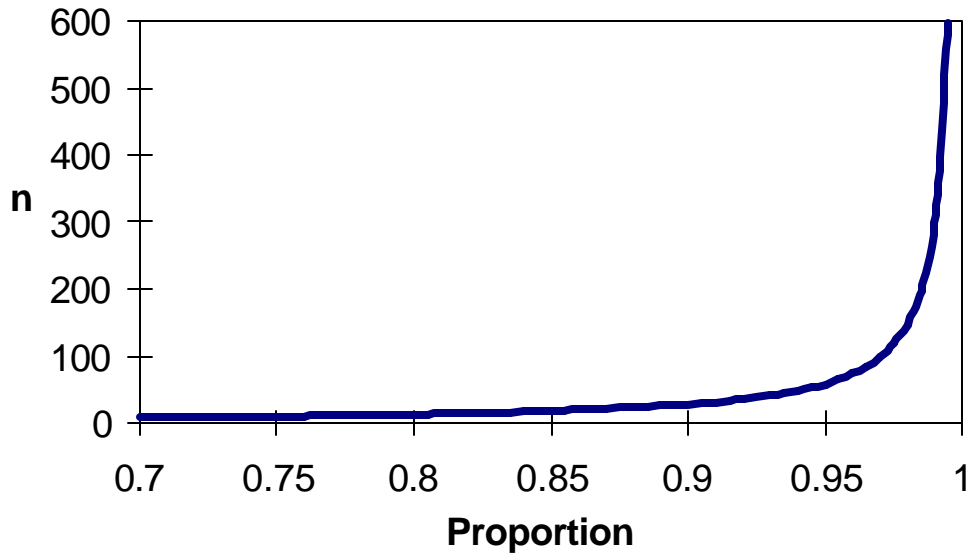
$$\mathbf{b}^n = \mathbf{a} \quad (6)$$

The proportion of the population that exceeds the first order statistic is the  $n^{\text{th}}$  root of the level of significance. As an example, if our sample size is 20 with  $\alpha = .05$ , then we can claim that we are 95% confident that at least 86% ( $\sqrt[20]{.05}$ ) of the population exceeds the smallest order statistic.

Equation 6 can also be used to determine sample size. The sample size equation is

$$n = \frac{\ln(\mathbf{a})}{\ln(\mathbf{b})} \quad (7)$$

Figure 1 summarizes these results for  $\alpha = .05$ . It is clear that to achieve  $\mathbf{b} = 1.0$ , all observations will exceed the lowest order statistic, the entire population would have to be sampled. Because of the logarithmic nature of the curve, a modest sample size of 29 is required to have  $\mathbf{b}$  exceed 0.9, but much larger samples are required to achieve small increases in  $\mathbf{b}$ .



**Figure 1: Sample size curve for a 95% lower tolerance limit of the first order statistic. The x-axis represents the proportion of the population that exceeds the first order statistic. The y-axis is the sample size required to achieve this coverage.**

A similar analysis for the upper tolerance bound yields the distribution of the proportion of the population that is less than the  $r^{\text{th}}$  order statistic. The distribution of this proportion is a Beta distribution with parameters  $r$  and  $n-r+1$ . Notice that the proportion of the population that is less than the largest order statistic,  $r = n$ , is given by Equation 5. Thus Figure 1 can be used to calculate the sample size for lower tolerance interval based on the first order statistic or a upper tolerance bound based on the largest order statistic.

### 3. Applications

Nabisco issued a challenge for their Chips Ahoy! cookies, claiming every bag of their Chips Ahoy! cookies contain more than 1000 chips. They have challenged the world to test this

claim. The introductory statistic classes at the United States Air Force Academy obtained bags of Chips Ahoy! cookies from all across America to meet the challenge. Since the challenge states that each bag contains more than 1000 chips, Nabisco is making a statement about each bag of cookies. Thus each bag is a single observation. A lower tolerance limit is the ideal tool for this analysis. Remember that a confidence interval would not be appropriate since it gives limits for a population parameter such as the mean. It would not answer Nabisco's claim to report that on **average** the bags have more than 1000 chips.

Forty-two bags were tested. Each cookie in the sample was dissolved in water and the number of chips of chocolate, a chip is defined as any unique piece of chocolate, were counted. The first order statistic, the smallest number of chips in the forty-two bags, was 1087. Using Equation 6, we claim that we are 95% confident that at least 93% of the bags of Chips Ahoy! cookies will contain more than 1087 chips. This is a powerful statement because it makes no assumptions about the distribution of chips in a bag and its conclusion relates to the population of individual bags.

There are a couple of observations about our claim. The first is that we claim at **least** 93% of the bags will have more than 1087 chips. It is possible that more than 93% of the bags will contain more than 1087 chips. Second, our lowest order statistic was 1087 which is 87 chips more than the minimum of 1000 claimed by Nabisco. We know that the percentage of bags which contain more than 1000 chips will be larger than 93%. Finally, we can use Figure 1 or Equation 7 to calculate how many bags are necessary to increase the proportion to a number larger than 93%. For example, to raise the proportion that exceed the lowest order statistic from 93% to 99% would require 300 bags of cookies. The knee in Figure 1 indicates a point of

diminishing returns where a prohibitively large sample is required for modest gain in the proportion of the population that exceeds the lowest order statistic.

#### **4. Summary**

In this paper we have demonstrated a method to develop nonparametric one-sided tolerance limits based on order statistics. For the case of the first order statistic, the distribution of the proportion of the population that exceeds the lowest order statistic has a simple probability density function. The cumulative density function can be calculated and used to calculate the tolerance bound or determine sample sizes. We demonstrated this method on the Chips Ahoy! challenge where we were able to claim with 95% confidence that at least 93% of all bags of Chips Ahoy! cookies have more than 1087 chips.

The method discussed in this paper also has potential use in military applications. For example, in specifying the lower limit of an altimeter, we would want to use a lower tolerance limit and not a confidence interval. The confidence interval would give a bound for the average altitude while the tolerance interval would give a bound for each flight.

#### **REFERENCES**

- Casella, G. and Berger, R. (1990). Statistical inference. *Wadsworth & Brooks/Cole Statistics/Probability Series*, Pacific Grove, California.
- David, H. A. (1981). Order Statistics. *Wiley & Sons*, New York, New York.