ESTIMATION OF POPULATION MAXIMUM USING SAMPLE ORDER STATISTICS

Limbore Jaya L.
Department of Statistics, Tuljaram Chaturchand College, Baramati, Dist. Pune (MS), India 413102
jayalimbore1702@gmail.com
Jagtap Avinash S.
Department of Statistics, Tuljaram Chaturchand College, Baramati, Dist. Pune (MS), India 413102
avinash.jagtap65@gmail.com

Abstract

Order statistics are mostly used in nonparametric methods but it is possible to use order statistics in a parametric setting and for estimating a parameter. The population maximum is a natural parameter of a discrete uniform distribution. This paper proposes to use sample order statistics for estimating the population size through population maximum. It is already known that sample maximum can be used to estimate the population maximum. This paper extends the approach and attempts to construct an estimator of the population maximum using all the sample order statistics.

Keywords: Order statistic, Sample maximum, Population maximum, Uniform distribution.

Introduction

The population sampling units can be arranged in an ascending order of magnitude to obtain population order. Similarly, when sample values are arranged in an ascending order of magnitude, sample order statistics are obtained. It may be interesting to note that when the population maximum is the parameter of interest, then, conditional on the sample maximum, the distribution of other sample order statistics is free from the parameter, namely the population maximum. In other words, conditional on the sample maximum, the other sample order statistics are ancillary. Still it is possible to use these ancillary statistics in constructing an estimator for the population maximum.

The population maximum is a natural parameter of a uniform distribution. As such the sample maximum is sufficient statistic for the population maximum in the uniform distribution. In this case, all other order statistics are ancillary conditional on the sample maximum. However, the marginal distribution of every order statistic involves the population maximum as a natural parameter. This paper discusses the possibility of using all the sample order statistics for estimating the population maximum. One special case of this approach has already been discussed in Chapter 3 in the form of the German Tank problem the solution to the German Tank problem takes the sample maximum and makes necessary either to remove or to reduce the bias and obtains an unbiased or almost unbiased estimator of population maximum (Limbore, 2017).

Suppose a sample of size n is denoted by $S = \{X_1, X_2, ..., X_n\}$. If the sample values are arranged in an ascending order of magnitude, then the result is an ordered sample

$$S_0 = \{X_{(1)}, X_{(2)}, ..., X_{(n)}\},$$

where $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$. If the population has finite bounds, let L denote the lower population bound and let U denote the upper population bound. Then it can be easily shown that the sample order statistics satisfy the following condition.

$$L \le X_{(1)} \le X_{(2)} \le ... \le X_{(n)} \le U.$$

It is also trivial to claim that $X_{(1)} \to L \, and \, X_{(n)} \to U$ as the sample size n increases indefinitely. In other words, $X_{(1)}$ and $X_{(n)}$ are asymptotically unbiased for L and U respectively.

When the population distribution has no finite bounds, researchers have modified the problem to that of estimating an upper percentile, still using similar methods. This chapter does not distinguish between these two problems and attempts to address them simultaneously. There is enough literature on the properties of order statistics for use in problems like estimating an upper percentile or even a higher quintile. When the population has finite bounds, it can always be transformed to a distribution on the unit interval through an appropriate change of origin and of scale. When the population has only a finite lower bound, it can be transformed so that the population is distributed on the positive half of the real line.

The Uniform Distribution

Consider the case of the uniform distribution, not necessarily confined to the unit interval. The uniform distribution may be discrete or continuous. The method is not affected by the range of the uniform distribution except for the requirement that the lower limit of the support of the distribution should be at zero. Let us first begin with a discrete uniform distribution on integers 1, 2, ..., N, so that the random variable X can take any of the N possible values with probability 1/N. Suppose $X_1, X_2, ..., X_n$ is a random sample from this distribution. That is, $X_1, X_2, ..., X_n$ are independent and identically distributed uniformly over the positive integers 1, 2, ..., N. The probability mass function of each of them is given by

$$P[X_r = i] = \frac{1}{N} \; for \; r = 1, 2, ..., n \; and \; i = 1, 2, ..., N \qquad (1)$$

The distribution function of each of the n sample values is given by

$$F_r(i) = P[X_r \le i] = \frac{i}{N} \; for \; i = 1, 2, ..., N \; and \; r = 1, 2, ..., n \qquad (2)$$

Let $X_{(1)}, X_{(2)}, ..., X_{(n)}$ be the order statistics of this sample so that $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$

The probability mass function of the i-th order statistics is then given by

$$P[X_{(r)} = i] = \frac{n!}{(i-1)!(n-i)!} \left(\frac{i}{N}\right)^{r-1} \left(1 - \frac{i}{N}\right)^{n-r} \frac{1}{N}, i = 1, 2, ..., N; r = 1, 2, ..., n \qquad (3)$$

The distribution function of the r-th order statistics is then given by

$$F_r(i) = P[X_{(r)} \le i] = \sum_{j=i}^{n} \binom{n}{j} \left(\frac{i}{N}\right)^{j} \left(1 - \frac{i}{N}\right)^{n-j} \; i = 1, 2, ..., N; r = 1, 2, ..., n \qquad (4)$$

Consequently, the expected value of the r-th order statistic $X_{(r)}$ is given by

$$E[X_{(r)}] = \frac{r}{n+1} N, \quad r = 1, 2, ..., n \qquad (5)$$

Further the variance of the r-th order statistic $X_{(r)}$ is given by

$$Var[X_{(r)}] = \frac{r(n-r+1)}{(n+1)^2 (n+2)} N^2 \qquad (6)$$

For r ≠s, 1 ≤ r, s ≤ n, the covariance between $X_{(r)}$ and $X_{(s)}$ is given by

$$\operatorname{cov}\left[X_{(r)}, X_{(s)}\right] = \frac{r(n-s+1)}{(n+1)^2(n+2)} N^2 \qquad (7)$$

It is easy to obtain the following from Equation (5) for every r = 1, 2, ..., N.

$$E\left[(n+1).\frac{X_{(r)}}{r}\right] = N, \quad r = 1,2,...,n \qquad (8)$$

Combining the above result for r = 1, 2, ..., n, we obtain

$$E\left[\frac{(n+1)}{n}\sum_{r=1}^{n} \frac{X_{(r)}}{r}\right] = \frac{(n+1)}{n}\sum_{r=1}^{n} \frac{X_{(r)}}{r} = N, \qquad (9)$$

giving an unbiased estimator of the population maximum N.

Further, the variance of this unbiased estimator of N is given by

$$Var\left[\frac{n+1}{n}\sum_{r=1}^{n} \frac{X_{(r)}}{r}\right]$$

$$= \frac{(n+1)^2}{n^2} Var\left[\sum_{r=1}^{n} \frac{X_{(r)}}{r}\right]$$

$$= \frac{(n+1)^2}{n^2}\sum_{r=1}^{n} \frac{Var\left[X_{(r)}\right]}{r^2} + \frac{(n+1)^2}{n^2}\sum_{r\neq s}^{n} \frac{\operatorname{cov}\left[X_{(r)}, X_{(s)},\right]}{r.s} \qquad (10)$$

First Consider

$$\sum_{r=1}^{n} \frac{Var\left[X_{(r)}\right]}{r^2}$$

$$= \sum_{r=1}^{n} \frac{r(n-r+1)}{r^2(n+1)^2(n+2)} N^2$$

$$= \frac{1}{(n+1)^2(n+2)}\sum_{r=1}^{n} \frac{n-r+1}{r} N^2$$

$$= \frac{N^2}{(n+1)^2(n+2)}\left[(n+1)\left(1+\frac{1}{2}+...+\frac{1}{n}\right)-n\right] \qquad (11)$$

Next Consider

$$\sum_{\substack{r=1 \\ r\neq s}}^{n}\sum_{s=1}^{n} \frac{\operatorname{cov}\left[X_{(r)}, X_{(s)}\right]}{r.s}$$

$$= \sum_{\substack{r=1 \\ r\neq s}}^{n}\sum_{s=1}^{n} \frac{r(n-s+1)}{r.s(n+1)^2(n+2)} N^2$$

$$= \frac{N^2}{(n+1)^2(n+2)}\sum_{\substack{r=1 \\ r\neq s}}^{n}\sum_{s=1}^{n} \frac{r(n-s+1)}{r.s}$$

$$= \frac{N^2}{(n+1)^2(n+2)}\sum_{\substack{r=1 \\ r\neq s}}^{n}\sum_{s=1}^{n} \frac{n-s+1}{s}$$

$$= \frac{(n-1)N^2}{(n+1)^2(n+2)} \sum_{s=1}^{n} \frac{n-s+1}{s}$$

$$= \frac{(n-1)N^2}{(n+1)^2(n+2)} \left[(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right) - n \right] \qquad (12)$$

Putting the terms together, we obtain

$$Var\left[\frac{(n+1)}{n}\sum_{r=1}^{n}\frac{X_{(r)}}{r}\right]$$

$$= \frac{N^2(n+1)^2}{n^2}\left\{\frac{1}{(n+1)^2(n+2)}\left[(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)-n\right]+\frac{N^2(n-1)}{(n+1)^2(n+2)}\left[(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)-n\right]\right\}$$

$$= \frac{N^2}{n^2(n+2)}\left\{\left[(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)-n\right]+(n-1)\left[(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)-n\right]\right\}$$

$$= \frac{N^2}{n^2(n+2)}\left\{n(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)-n^2\right\}$$

$$= \frac{N^2}{n(n+2)}\left\{(n+1)\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)-n\right\}$$

Table 1: Exact expressions for the expected value of sample maximum of order statistics from discrete uniform distribution

n	$\mu_{n:n}$
1	$(1/2)(N+1)$
2	$(1/6)N^{-1}(4N^2+3N-1)$
3	$(1/4)N^{-1}(3N^2+2N+1)$
4	$(1/30)N^{-3}(24N^4+15N^3-10N^2+1)$
5	$(1/12)N^{-3}(10N^4+6N^3-5N^2+1)$
6	$(1/42)N^{-5}(36N^6+21N^5-21N^4+7N^2-1)$
7	$(1/24)N^{-5}(21N^6+12N^5-14N^4+7N^2-2)$
8	$(1/90)N^{-7}(80N^8+45N^7-60N^6+42N^4-20N^2-3)$
9	$(1/20)N^{-7}(18N^8+10N^7-15N^6+14N^4-10N^2+3)$
10	$(1/66)N^{-9}(6N^{10}+33N^9-55N^8+66N^6-66N^4+33N^2-5)$
11	$(1/24)N^{-9}(22N^{10}+12N^9-22N^8+66N^6-44N^4+33N^2-10)$
12	$(1/2730)N^{-11}(2520N^{12}+1365N^{11}-2730N^{10}+5005N^8-858N^6+9009N^4-4550N^2+691)$
13	$(1/420)N^{-11}(390N^{12}+210N^{11}-455N^{10}+1001N^8-2145N^6+3003N^4-2275N^2+691)$
14	$(1/90)N^{-13}(84N^{14}+45N^{13}-105N^{12}+273N^{10}-715N^8+1287N^6-1365N^4+691N^2-105)$
15	$(1/48)N^{-13}(45N^{14}+24N^{13}-60N^{12}+182N^{10}-572N^8+1287N^6-1820N^4+1382N^2-420)$

Table 2: Exact expressions for the variance of sample maximum of order statistics from discrete uniform distribution

n	$\sigma = \sigma^2_{n:n}$
1	$(1/12)(N^2-1)$
2	$(1/36)N^{-2}\,(2N^2+1)\,(N^2-1)$
3	$(1/240)N^{-2}\,(9N^2-1)\,(N^2-1)$
4	$(1/900)N^{-6}\,(24N^6-21N^4-19N^2+1)\,(N^2-1)$
5	$(1/1008)N^{-6}\,(20N^6-36N^4-15N^2+7)\,(N^2-1)$
6	$(1/1764)N^{-10}\,(27N^{10}-78N^8+6N^6+78N^4-13N^2+1)\,(N^2-1)$
7	$(1/2880)N^{-10}\,(35N^{10}-145N^8+93N^6+213N^4-120N^2+20)\,(N^2-1)$
8	$(1/8110)N^{-14}\,(80N^{14}-445N^{12}+575N^{10}+715N^8-1449N^6+541N^4-111N^2+9)\,(N^2-1)$
9	$(1/13200)N^{-14}\,(108N^{14}-772N^{12}+1571N^{10}+911N^8-5821N^6+4389N^4-1683N^2+297)\,(N^2-1)$
10	$(1/4356)N^{-18}\,(30N^{18}-267N^{16}+767N^{14}-25N^{12}-3622N^{10}+5690N^8-3572N^6+1444N^4-305N^2+25)(N^2-1)$
11	$(1/262080)N^{-18}(1540N^{18}-16660N^{16}+63420N^{14}-42400N^{12}-349707N^{10}+958873N^8-980525N^6)+641095N^4-254800N^2+45500)(N^2-1)$
12	$(1/7452900)N^{-22}\,(37800N^{22}-487725N^{20}+2359665N^{18}-3195885N^{16}-13921600N^{14}+63345590N^{12}-104252950N^{10}+99659850N^8-66497141N^6+27342319N^4-5810619N^2+477481)(N^2-1)$
13	$(1/176400)N^{-22}(780N^{22}-11820N^{20}+70535N^{18}-148255N^{16}-399506N^{14}+3051214N^{12}-7462327N^{10}+10192303N^8-9968838N^6+6659202N^4-22666569N^2+477481)(N^2-1)$
14	$(1/16200)N^{-26}(63N^{26}-110N^{24}+7965N^{22}-23235N^{20}-36300N^{18}+515160N^{16}-1785320N^{14}+3428080N^{12}-4587230N^{10}+4530710N^8-3053308N^6+1260092N^4-268170N^2+22050)\,(N^2-1)$
15	$(1/195840)N^{-26}\,(675N^{26}-13605N^{24}+115935N^{22}-440985N^{20}-266395N^{18}+10536085N^{16}+130345829N^{12}-231687560N^{10}+313890720N^8-310871860N^6+208610740N^4-83680800N^2+14994000)(N^2-1)$

It is interesting to note in this regard that the population maximum can also be estimated with help of the sample maximum alone. Further, after correcting for the bias, the sample maximum provides the minimum variance unbiased estimator of the population maximum because the sample maximum is complete and sufficient for the population maximum. More specifically,

$$E\left[\frac{(n+1)}{n}X_{(n)}\right]=N \qquad (13)$$

and

$$\mathrm{var}\left[\frac{(n+1)}{n}X_{(n)}\right]=\frac{N^2}{n(n+2)} \qquad (14)$$

It should be noted that the variance of the previous estimator based on all sample order statistics has a large variance partly due to the fact that sample order statistics are positively correlated and therefore inflate the variance of a linear combination. It would be better if order statistics could be made stochastically independent, because that would eliminate the covariance terms from the variance of their linear combination.

Conclusion:

The main conclusion of this paper is that the population maximum can be estimated using the sample order statistics. When the population maximum is unknown and is treated like a parameter, the sample maximum is a complete sufficient statistic and therefore the estimator of the population maximum

based on the sample maximum is the most efficient estimator of the population maximum. All other sample order statistics are ancillary and hence their conditional distribution on the sample maximum does not depend on the population maximum. Nevertheless, the marginal distribution of every order statistics depends on the population maximum as the parameter. Sample order statistics for a random sample drawn from any distribution can be transformed to order statistics for a sample drawn from the uniform distribution on the unit interval [0, 1].

References:

1. David H. A. and Nagaraja H. N. (2003). Order Statistics, third ed., John Wiley and Sons, New York.
2. Gokhan Gokdere , Fahrettin Ozbey, and Mehmet gungor (2011). On Order Statistics of Discrete random Variables, e-Journal of New World Science Academy 2011, vol.6, No.2, Article No: 3A0035, pp.54-58.
3. J. L. Limbore (2017). Investigating Statistical Properties of Sample Maximum for Estimating Population Maximum (Thesis). JJT University, Rajasthan, India.
4. M. Gungor, Y. Bulut, and B. Yuzbasi (2012). On joint distributions of order statistics for a discrete case, Journal of Inequalities and Applications, 2012, Online.
5. M. Güngör1, Y. Bulut, and S. Çalık (2009). Distributions of Order Statistics, Applied Mathematical Sciences, Vol. 3, No. 16, pp. 795 – 802.
6. Suxia Yao, Yu Miao, Saralees Nadarajah (2015). Exponential Convergence for the k-th Order Statistics, Filomat 29:5 (2015), 977–984 DOI 10.2298/FIL1505977Y.