Order Statistics and Applications

E. Jack Chen
BASF Corporation, USA

INTRODUCTION

Order statistics refer to the collection of sample observations sorted in ascending order and are among the most fundamental tools in non-parametric statistics and inference. Statistical inference established based on order statistics assumes nothing stronger than continuity of the cumulative distribution function of the population and is simple and broadly applicable. Important special cases of order statistics are the minimum and maximum value of samples, the sample median and other sample quantiles. Order statistics have been widely studied and applied to many real-world issues. It is often of interest to estimate the reliability of the component/system from the observed lifetime data. In many applications we want to estimate the probability that a future observation will exceed a given high level during some specified epoch. For example, a machine may break down if a certain temperature is reached. For more applications of order statistics, see (Arnold et al., 2008; David & Nagaraja, 2003; DasGupta, 2011).

It is the purpose of this chapter to present some of the more important results in the sampling theory of order statistics and of functions of order statistics and their applications, such as tests of independence, tests of goodness of fit, hypothesis tests of equivalence of means, ranking and selection, and quantile estimation. Order-statistics techniques continue to be key components of statistical analysis. For example, quantile estimates are used to build empirical distributions, tests of goodness of fit are used to check the validity of the fitted input distribution, hypothesis tests of equivalence of means, and ranking and selection procedures are used to compare the performance of multiple systems.

DOI: 10.4018/978-1-5225-2255-3.ch162

BACKGROUND

Let $X_i, i = 1, 2, \ldots, n,$ denote a sequence of mutually independent random samples from a common distribution of the continuous type having a probability density function (pdf) f and a cumulative distribution function (cdf) F. Let $X_{[u]}$ be the u^{th} smallest of these X_i such that $X_{[1]} \le X_{[2]} \le \ldots \le X_{[n]}$. Then $X_{[u]}, u = 1, 2, \ldots n,$ is called the u^{th} order statistic of the random sample $X_i,\ i = 1,2,\ldots,n.$ Note that even though the samples $X_1, X_2, \ldots, X_n$ are independently and identically distributed (iid), the order statistics $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$ are not independent because of the order restriction. The difference $R = X_{[n]} - X_{[1]}$ is called the sample range. It is a measure of the dispersion in the sample and should reflect the dispersion in the population.

Suppose $U_1, U_2 \sim U\left(0,1\right)$, where "$\sim$" denotes "is distributed as" and "$U\left(a,b\right)$" denotes a uniform distribution with range [*a,b*]. We are interested in the distribution of $U_{[2]} = \max(U_1, U_2)$, which can be viewed as the second order statistic. The cdf of $U_{[2]}$ is

$$\begin{aligned} F\left(x\right) &= P\left(\max\left(U_1, U_2\right) \le x\right) \\ &= P\left(U_1 \le x, U_2 \le x\right) \\ &= P\left(U_1 \le x\right) P\left(U_2 \le x\right) \\ &= x^2 \end{aligned}$$

Furthermore, the cdf of $U_{[1]} = \min(U_1, U_2)$, which can be viewed as the first order statistic, is

$$\begin{aligned} F(x) &= 1 - P(\min(U_1, U_2) > x) \\ &= 1 - P(U_1 > x)P(U_2 > x) \\ &= 1 - (1 - x)^2 \end{aligned}$$

Consider a device contains a series of two identical components and the lifetime of the component $L_c \sim U(0,1)$. If any of these two components fail, the device fails, such as chains. Then, the lifetime of the device $L_d \sim \min(L_{c1}, L_{c2})$. On the other hand, if these two components are configured in parallel and the device fails only when both components fail, such as a strand contains a bundle of threads. Then the lifetime of the device $L_d \sim \max(L_{c1}, L_{c2})$. These techniques can be used in reliability analysis to estimate the reliability of the component/system from the observed lifetime data.

When there are more than two samples, the distributions of sample minimum and sample maximum can be obtained via the same technique. However, the u^{th} ($u \neq 1$ and $u \neq n$) order statistic needs to be obtained via other techniques. Similarly, joint distributions of order statistics and conditional distributions of order statistics require more complicated derivation.

MAIN FOCUS

We discuss distributions of order statistics, joint and conditional distributions of order statistics, and empirical distribution functions and show how order statistics techniques are used in statistical analysis. We then present several procedures that are derived based on inference of order statistics, e.g., tests of independence, range statistics and hypothesis tests of equivalence of means, ranking and selection procedures, and quantile estimation.

Distributions of Order Statistics

Hogg et al. (2012) show that the distribution of the u^{th} order statistic of n samples of X (i.e., $X_{[u]}$) is

$$g_{u:n}(x) = \beta(F(x); u, n-u+1) f(x),$$

where

$$\beta(x; a, b) = \frac{(a+b-1)!}{(a-1)!(b-1)!} x^{a-1} (1-x)^{b-1}.$$

That is,

$$g_{u:n}(x) = n! \frac{[F(x)]^{u-1}}{(u-1)!} \frac{[1-F(x)]^{n-u}}{(n-u)!} f(x).$$

Furthermore, the cdf

$$G_{u:n}(y) = \int_{-\infty}^{y} g_{u:n}(x)\, dx.$$

In particular, the first order statistic, or the sample minimum, $X_{[1]}$ has the pdf

$$g_{1:n}(x) = n[1-F(x)]^{n-1} f(x)$$

and the cdf

$$G_{1:n}(x) = 1 - [1-F(x)]^{n}.$$

The n^{th} order statistic, or the sample maximum, $X_{[n]}$ has the pdf

$$g_{n:n}(x) = n[F(x)]^{n-1} f(x)$$

and the cdf

$$G_{n:n}(x) = [F(x)]^{n}.$$

In the cast that f is the uniform [0,1] distribution,