Order Statistics

I Bairamov, Izmir University of Economics, Izmir, Turkey

Glossary

Censored sample – A sample appearing in the life-testing experiments when *n* items are kept under observation and only part of this sample can be observed.
Counting process – A stochastic process representing the total number of curtain event occurring up to time.
Distribution free statistics for a class of distribution function – A statistic whose distribution is the same for all distribution function in the class.
Empirical distribution function – A natural estimation of the cumulative distribution function constructed on the base of the sample.
Extreme order statistics – The maximal and minimal order statistics.
Markov property – This property states that to make predictions of the behavior of a system in the future, it suffices to consider only the present state of the system and not past history.
Order Statistics – Original random sample arranged in order of magnitude.
Probability integral transformation – A transformation that transforms an arbitrary distribution to the uniform distribution.
Robust estimators – Estimators that are efficient in presence of outliers in the sample.
Two sample problem – A hypothesis testing problem verifying whether two sample are identically distributed or not.

The Subject of Order Statistics

The independent and identically distributed random variables, which can be interpreted as results of an experiment measuring values of a certain random variable arranged in order of magnitude, are called order statistics. In the statistical model of many experiments, for instance, in reliability analysis, life time studies, the analysis of time to graduation of students, and testing of strength of materials, the realizations arise in nondecreasing order; therefore, the use of order statistics is necessary. Order statistics are extensively used in statistical inferences: in estimation theory and hypothesis testing.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population with cumulative distribution function (cdf) $F(x)$. Suppose that the elements of this sample are arranged in order of magnitude and $X_{(1)}$ denotes the smallest; $X_{(2)}$ denotes the second smallest; etc., and $X_{(n)}$ denotes the largest of the set $X_1, X_2, \ldots, X_n$. Then $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ denotes the original random sample arranged in increasing order of magnitude, and these are called the order statistics of the sample $X_1, X_2, \ldots, X_n$. We call $X_{(i)}$, for $1 \leq i \leq n$ the ith order statistic. The subject of order statistics deals with the distributional properties of $X_{(i)}$ itself. and some functions of the subset of the n order statistics and their applications. If, for example, the scores of n students in the exam are $X_1, X_2, \ldots, X_n$, then $X_{(n)}$ represents the score of the best student; $X_{(1)}$ is the score of the weakest; the sample range $W = X_{(n)} - X_{(1)}$ is a measure of dispersion; the sample median, defined as $(X_{(n/2)} + X_{(n/2+1)})/2$ for an n even and as $X_{[(n+1)/2]}$, for n odd, is a measure of location and estimate the central tendency of scores. Here, $[a]$ is the integer part of the number a. The sample midrange, defined as $(X_{(1)} + X_{(n)})/2$, is also a measure of central tendency.

Order statistics have wide applications in many areas where the use of an ordered sample is important. Order statistics are among the most fundamental tools in nonparametric statistics, because the transformation $U_{(i)} = F(X_{(i)})$ produces a random variable which is the ith order statistics from the uniform population on the interval (0, 1), and therefore $U_{(i)}$ is distribution free, that is, its distribution function is independent of the distribution function F of the original sample. This transformation is called the probability integral transformation.

It is well known from classical statistical theory that the natural estimate of an unknown distribution function is the empirical distribution function, which is a function of order statistics. Therefore, many important statistics in estimation theory and hypothesis testing appear to be an integral functional of the empirical distribution function, and can be expressed in terms of order statistics. Order statistics do not change their order under probability integral transformation, namely if $U_{(i)} = F(X_{(i)})$, $i = 1, 2, \ldots, n$, then $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$. Due to unique distribution free properties, they are widely used in nonparametric interval estimation and hypothesis testing.

Order statistics and their properties have been extensively studied since the early part of the last century, and recent years have seen a particularly rapid growth of studies. The multiauthored book *Contributions to Order*

Statistics, edited by A. H. Sarhan and B. G. Greenberg, appeared in the Wiley series in probability and statistics in 1962. The first monograph. *Order Statistics* by H. David appeared in 1970 in the same Wiley series and has served as a text, a survey of growth, and a general introduction. The second edition appeared in 1981 and the third, coauthored with H. Nagaraja, in 2003. For further reading the reader is refered to Arnold *et al.* (1992) and Balakrishnan (2007).

Basic Distribution Theory

The elements of the sample $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid), but the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are dependent random variables. The distribution of the rth order statistics can be derived using the independence of the random variables $X_1, X_2, \ldots, X_n$ and observing that the event $\{X_{(r)} \leq t\}$ occurs if and only if at least r of the observations $X_1, X_2, \ldots, X_n$ falls below t. Therefore, taking into account the fact that the probability of occurrence of exactly i of events $\{X_k \leq x\}$ in n independent Bernoully trials is $\binom{n}{i} F^i(\mathrm{x})(1-F(\mathrm{x}))^{n-i}$, the cdf of $X_{(r)}$ can be written as

$$F_r(x) = P\{X_{(r)} \leq x\} = \sum_{i=r}^{n} \binom{n}{i} F^i(x)(1-F(x))^{n-i} \qquad [1]$$

If F is absolutely continuous with probability density function (pdf) f, then [1] can also be rewritten as follows:

$$F_r(x) = \frac{n!}{(r-1)!(n-r)!} \int_0^{F(x)} u^{r-1}(1-u)^{n-r} \mathrm{d}u \qquad [2]$$

Formula [1] holds true for both discrete and continuous distribution functions. Formula [2] is true only for absolutely continuous distributions. Given the realizations of the n order statistics to be $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, the original random variables X_i are restrained to take on the values $x_{(i)}$ $(i = 1, 2, \ldots, n)$ which by symmetry assigns equal probability to each of the $n!$ permutations of $(1, 2, \ldots, n)$. Therefore, the joint pdf of all n order statistics is

$$f_{1,2,\ldots,n}(x_1, x_2, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i) \text{ for } x_2 < x_2 < \ldots < x_n$$

Since $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are dependent random variables, then their joint distributions are important. The expressions for the joint pdf's of two or more order statistics can be found in David (1981).

Order statistics from uniform distribution on [0, 1] are important when one needs to generate the order statistics from any distribution using Monte Carlo simulation. If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are order statistics from the population with cdf F and $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ is the inverse of F, then $F^{-1}(U_{(i)}) = X_{(i)}$, the equality here is in distribution. There are various methods of generating uniform random variables. Using computer simulation we generate sample $U_1, U_2, \ldots, U_n$ from the uniform distribution in [0, 1] and then order the sample. The $X_{(i)}$ value then can be calculated as $X_{(i)} = F^{-1}(U_{(i)})$. For various methods of generating order statistics, see Tadikamalla and Balakrishnan (1998).

The pdf of $W_{rs} = X_{(s)} - X_{(r)}$ when the parent population is uniform in [0, 1] depends only on $s - r$, and not on r and s individually. In addition, the pdf of the sample range $W = X_{(n)} - X_{(1)}$ is $fW_{1n}(x) = n(n-1)x^{n-2}(1-x), 0 \leq x \leq 1$.

Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics based on the sample $X_1, X_2, \ldots, X_n$ with cdf $F(x) = 1 - \exp(-\lambda x), x \geq 0$. Then the spacings $Y_1 = X_{(1)}, Y_2 = X_{(2)} - X_{(1)}, \ldots, Y_n = X_{(n)} - X_{(n-1)}$ are independent; furthermore, the random variables $Z_r = (n - r + 1)\,\lambda(X_{(r)} - X_{(r-1)}), r = 1, 2, \ldots, n$ are iid with cdf $F(x) = 1 - \exp(-x), x \geq 0$, where $X_0 = 0$. If n units are placed under strength test and $X_1, X_2, \ldots, X_n$ are independent random variables with exponential distribution and represent the life lengths of these units, then the lengths of time intervals $X_{(r)} - X_{(r-1)}$, $r = 1, 2, \ldots, n$ between two failures are independent and indentically distributed random variables. Then $X(r) = \sum_{i=1}^{r} Z_i/\lambda(n-i+1)$ that is, $\mathrm{X}_{(r)}$ can be represented as a sum of iii random variables. Then the conditional distribution of $X_{(r+1)}$ given $X_{(1)} = x_1, X_{(2)} = x_2, \ldots, X_{(r)} = x_{(r)}$ is the same with the conditional distribution of $X_{(r+1)}$ given $X_{(r)} = x_{(r)}$. This means that $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ satisfy Markov property and form an additive Markov chain. The Markov property states that to make predictions of the behavior of a system in the future, it suffices to consider only the present state of the system and not the past history. The sequence of dependent random variables satisfying the Markov property is called the Markov chain. This property helps in establishing the Markovian dependence structure of order statistics from the sample with any continuous distribution. It follows that the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ from a population with continuous cdf form a Markov chain.

There are some interesting properties of order statistics connected with truncation of these ordered observations. For instance, let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be order statistics of the sample $X_1, X_2, \ldots, X_n$ with absolutely continuous cdf F and pdf f. Then given $X_{(r)} = x$, $X_{(s)} = y$ the joint pdf of $(X_{(r+1)}, X_{(r+2)}, \ldots, X_{(s-1)})$ is the same with the joint pdf of the order statistics $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(s-r)})$ from the sample $Y_1, Y_2, \ldots, Y_{s-r}$ size $s - r$, where Y_i has pdf of the random variable X_i given $x < X_i < y$.

A counting process $\{N(t), t \geq 0\}$ representing the total number of event A occurring up to time t is called a Poisson process if it has stationary and independent increments. Note that, a stochastic process $\{N(t), t \geq 0\}$ is said to have independent increments if, for all $t_0 < t_1 < t_2 < \ldots < t_n$, the random variables $N(t_1) - N(t_0), N(t_2) - N(t_1), \ldots, N(t_n) - N(t_{n-1})$ are independent. It possesses stationary increments if $N(t + h) - N(t)$ has the same distribution for all t. There is an interesting connection between the interarrival times of the occurrence of event A and the order statistics. Let $N(t)$ be a Poisson process with rate λ, $\lambda > 0$, then $P\{N(t) = k\} = e^{-\lambda t}(\lambda t)^k/k!$ ($k = 0$, 1, 2, . . .). Denote by X_1 the time of the first event, X_2 the time between the first and the second event, X_n the time between $(n - 1)$ th and n th event. Then the sequence $\{X_n\}\, n \geq 1$ is called the sequence of interarrival times. It is well known that $X_1, X_2, \ldots, X_n$ are iid exponential random variables having a mean of $1/\lambda$. Another quantity is $S_n = \sum_{i=1}^{n} X_i$, $n \geq 1$, the arrival time of, or waiting time until, the nth event. Then given that $N(s) = n$, the n arrival times $S_1, \ldots, S_n$ have the same distribution as the order statistics corresponding to the independent random variables uniformly distributed on the interval $[0, s]$, i.e.

$$\begin{aligned} &P\{S_1 \leq t_1, S_2 \leq t_2, \ldots, S_n \leq t_n | N(s) = n\} \\ &= P\{U_{(1)} \leq t_1, \ldots, U_{(n)} \leq t_n\} \\ &= \frac{n!}{s^n}, 0 < t_1 < t_2 < \ldots < t_n, \end{aligned}$$

where $U_{(i)}$ is the ith order statistic from uniform in $[0, s]$ distribution.

Order Statistics in Statistical Inference

Order statistics are essential in several optimal inference procedures and hypothesis testing problems. In many cases when the underlying distribution has finite support, the order statistics themselves become sufficient statistics and, thus, provide minimum variance unbiased estimators and the most powerful test procedures for the unknown parameters. Note that, if the random variable X takes values from the interval $[a, b]$, then we say that this random variable has support $[a, b]$, $-\infty \leq a < b \leq \infty$. If the random variable has finite support and if the support involves the parameters of distribution of this random variable, then many of estimates of parameters involve order statistics. For example, let X be uniform in $[0, \theta]$ random variable, that is, the pdf of X is $f(x; \theta) = 1/\theta$ if $0 \leq x \leq \theta$. It is well known that $X_{(n)}$ is a sufficient, and complete statistic for θ and $\frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ.

The order statistics appear in a natural way in the inference procedures when the sample is censored and only part of the sample values are available. The censored samples appear in the life-testing experiments when n items are kept under observation until failure. These items could be technical systems or their components, patients under certain drug or clinical conditions, or candidates undergoing exams in complex conditions or under time pressure.

In measuring performance on examinations many teachers use average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, where $X_1, X_2, \ldots, X_n$ are test scores of students. It is well known that poor lecture attendance is associated with lower test scores (see, e.g., Myles and Henderson, 2002, Williams *et al.*, 2002; Balch, 1992). The presence of a student with poor attendance will probably lead to a lower test score for the class, and this outlier will affect the sample mean $\bar{X}$ and variance S^2. Therefore, the existence of outliers will result in misestimation of class performance. To avoid this kind of problem in applications, we need to find estimators only minimally affected by the presence of outliers. In statistical literature such estimators are called robust estimators. One popular robust estimator of the center of a symmetric distribution is the symmetric trimmed mean

$$\hat{\theta}_r = \frac{1}{n - 2r}\sum_{i=r+1}^{n-r} X_{(i)}, \;\; 0 \leq r \leq \left[\frac{n-1}{2}\right]$$

where we have trimmed the top r and the bottom r order statistics. The trimmed means give less weight to the sample extremes and are suggested as robust estimators. They are robust against the presence of a small number of outliers and highly efficient in their complete absence. Barnett and Lewis (1994) describe an estimator:

$$\hat{\theta} = \begin{cases} \hat{\theta}_0, & \text{if } \max\{(\bar{X} - X_{(n)}), (X_{(n)} - \bar{X})\} < c \\ \hat{\theta}_1 & \text{otherwise} \end{cases}$$

for the mean μ of a sample from an $N(\mu, 1)$, where *distribution suspect one of the values is from* $N(\mu + \delta, 1)$, and θ_0 and θ_1 are trimmed means and c can be the specified percentile of the distribution of max $|X_i - \bar{X}|$. This statistic is used for testing for a single extreme outlier. Balakrishnan (2007) computed the bias and mean square error of robust estimators constructed in the base of order statistics and presented tables of numerical vales for $n = 10$. For a further reading on robust statistics, see Huber (1981), Andrews *et al.* (1972), and David and Ghosh (1985).

Order Statistics and Education

Order statistics play an important role in educational statistics. In many statistical analyses, the information from a random sample is utilized through the ordered values of the sample. At the beginning of the course ‘nonparametric statistics’ the students face the nontrivial operation of ordering of a random sample. On the one hand, theoretically the elements of the random sample are random variables, that is, measurable functions $X_1(\omega)$, $X_2(\omega), \ldots, X_n(\omega)$ $(\omega \in \Omega)$ given in the probability space $\{\Omega, F, P\}$, where Ω is a sample space, F is a σ–algebra of subsets of the sample space, and P is a probability measure. On the other hand, the sample values $X_1, X_2, \ldots, X_n$ are considered as realizations of the experiment measuring values of the random variable X, namely, they are numbers. In deriving the cdf of rth order statistics for better understanding of the idea of proof, we consider the random sample as numbers, when they are actually random variables. Understanding the structure of order statistics presents difficulties when considering the original sample as functions. For example, the distribution function of the maximal order statistic $X_{(n)}$ is $F_n(x) = P\{X_{(n)} \leq x\}$ and if we consider $X_1, X_2, \ldots, X_n$ as numbers, then the maximum is less than or equal to x if all numbers are less than or equal to x. Therefore, the event $\{X_{(n)} \leq x\}$ occurs if and only if all the members of original sample are less than or equal to x, namely $\{X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x\}$. Now, since the original X's are independent and all have the same cdf F, we can state

$$F_n(x) = P\{X_{(n)} \leq x\}$$
$$P\{X_1 \leq x, X_2 \leq x, \ldots X_n \leq x\} = F^n(x)$$

In general, the cdf of rth order statistic, $P\{X_{(r)} \leq x\} = P\{\text{at least } j\ X\text{'s are} \leq x\} = \sum_{n}^{i=j} \binom{n}{j} P\{\text{exactly } i\ X\text{s are} \leq\ x\}$, here again we use the independence of the original sample. This approach is useful for understanding the structure of order statistics that are actually measurable functions of the elements of sample space.

As pointed out above the probability integral transformation $U = F(X)$ transforms the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ to uniform order statistics $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ and preserves the order. In hypothesis testing and confidence intervals, the field of nonparametric statistics relies on the concept known as distribution-free property. For the distribution-free hypothesis test, the significance level remains constant over a class of underlying distributional assumptions. The distribution-free or invariant confidence interval has a constant confidence level holding over a class of distribution functions. The test statistic $S = S(X_1, X_2, \ldots, X_n)$ is designated distribution free over some class of distributions, say F, if the distribution of S is the same for every distribution in F. The view that many students have difficulties understanding these concepts seems to be a common one among my colleagues teaching nonparametric statistics. However, this problem can be overcome using order statistics to construct effective examples. For example, let $\mathcal{F} = \mathcal{F}_\theta$ be a scale parameter class; this means that if $F_\theta(x) \in \mathcal{F}_\theta$ then $F_\theta(x) = F(x/\theta)$, for some distribution function F and parameter θ. Then the distribution of statistic defined as midrange divided by the range, that is,

$$T = \frac{1}{2}\frac{X_{(n)} + X_{(1)}}{X_{(n)} - X_{(1)}}$$

does not depend on parameter θ; in other words, this statistic is distribution free for the class F_θ. As another example, let $F = F_c$ be the class of all continuous distribution functions and $X_1, X_2, \ldots X_n$ be a sample from the population with distribution function $F \in F_c$, and let X_{n+1} be the $(n+1)$th observation from the same population independent of $X_1, X_2, \ldots, X_m$ then the probability that X_{n+1} falls into interval $(X_{(r)}, X_{(s)})$ is $(s-r)/(n+1)$. This probability is the same for all distribution functions F from the class F_c, which means that the interval $(X_{(r)}, X_{(s)})$ constructed by the order statistics is the distribution-free confidence interval for the future observation X_{n+1}. It is interesting to note that if the distribution is continuous under some regularity conditions, the interval $(X_{(r)}, X_{(s)})$ is the only distribution-free interval for the future observation X_{n+1} (see Bairamov and Petunin, 1990).

The importance of order statistics can also be seen in teaching the theory of the ranking statistics used in two-sample problem with unknown shift parameters. The sample observation X_i is said to have rank R_i among $X_1, X_2, \ldots, X_n$ if $X_i = X_{(R_i)}$, where $X_{(R_i)}$ is the R_ith order statistic. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples from continuous distributions with distribution functions $F(x)$ and $G(x) = F(x-\theta)$, respectively, where $-\infty < \theta < \infty$ is an unknown shift parameter. The Mann–Whitney–Wilcoxon nonparametric test for verifying the null hypothesis $H_0 : \theta = 0$ against alternative $H_1 : \theta > 0, \theta < 0$ or $\theta \neq 0$ is constructed based on the distribution-free property of the rank statistic $W = \sum_{i=1}^{n} R_i^*$ under hypothesis H_0, where R_j is the rank of R_j among the m X's and n Y's combined and treated as a single set of observations. In a general two-sample problem, when $H_0 : F = G$ is to be tested against a general class of alternatives $H_1 : F \neq G$, the Kolmogorov–Smirnov test based on the distance $D_{n,m} = \sup_{-\infty < x < \infty} |F_n(x) - G_m(x)|$ of two empirical distribution functions $F_n(x)$ and $G_m(x)$ of the samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$, respectively, is one of the most important consistent hypothesis tests. The empirical distribution function $F_n(x)$, defined as the number of observations $X_1, X_2, \ldots, X_n$ less than or equal to x divided by n, has the following expression in terms of order statistics:

$$F_{(n)}(x) = \begin{cases} 0, & \text{if} \quad x < X_{(1)} \\ \frac{k}{n} & \text{if} \quad X_{(k)} \leq x < X_{(k+1)} \quad \text{for } k = 1, 2, \ldots, n-1 \\ 1 & \text{if} \quad x \geq X_{(n)} \end{cases}$$

Example

The selection of students from different schools for a scholarship shortlist has been an issue of public interest. In many countries there is a general consensus on the existence of bias in selection of candidates. Fairness has been defined in a variety of ways; Torndike (1971) proposed a definition of fairness; Cole (1973) made a fundamental assumption in the reviewed models that the applicants are independently and identically distributed random variables. The model proposed by Olkin and Stephens (1993) used ordered scores of students, namely, order statistics. More specifically, let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be test scores for two groups of students representing two different high schools A and B, respectively. We assume that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are random samples from continuous distributions F_1 and F_2, respectively. The $m + n$ scores are pooled and jointly ranked, and the top k students are shortlisted for scholarships. Let R be the number of students from the first group that enter the shortlist. We are interested in the probability $P\{R = r\}$ that exactly r students from high school A appear in the shortlist. Olkin and Stephens (1993) provide an elegant solution to this problem. If $F_1 = F_2$, in other words, we consider two different groups of students from identical high schools, then this probability depends only on n, m, r, and k and can therefore be easily calculated. For the special case of choosing a single candidate, that is, $r = k = 1$, the probability that exactly one student from high school A will appear in the shortlist is equal to $n/(n + m)$, which is the probability that the maximal score $X_{(n)}$ of the first group is greater than the maximal score of the second group $Y_{(n)}$, where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(m)}$ are the ordered scores of two groups. A numerical analysis shows that when n is low relative to m, this probability is low, which may mean that a student from a small but prestigious school has considerably less chance of being shortlisted when in competition with a larger school. Some numerical values of probability that exactly r students from the high school A enter the shortlist for $r = 1$ and $r = 2$, and $k = 1, 2, \ldots, 10$ are given below:

1. $n = 9$, $m = 30$

k	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
$P\{1, k; 9, 30\}$	0.11	0.12	0.23	0.34	0.37	0.42	0.45	0.46	0.47	0.47

2. $n = 3$, $m = 25$

k	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
$P\{1, k; 3, 25\}$	0.	0.01	0.02	0.04	0.07	0.10	0.14	0.17	0.21	0.25

Let **P** denote the probability of the event that the best candidate from school A with score $X_{(n)}$ is included in the top group of k candidates; thus, at least one student from A is in the list. We provide a numerical example of $n = 9$ and $m = 30$:

k	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
P	0.23	0.41	0.56	0.67	0.75	0.82	0.878	0.91	0.93	0.95

From the table it can be seen that a shortlist of size at least $k = 4$ is needed in order to guarantee that the probability of at least one student from the group of size $n = 9$ being included in the shortlist is at least 0.905.

In the general case, when $F_1(\cdot) \neq F_2(\cdot)$, the probability that exactly r students from high school A appear in the shortlist is equal to $P\{X_{(n+1-r)} > Y_{(m+r-k)}\} - P\{X_{(n-r)} > Y_{(m+r-k+1)}\}$. In particular, when $F_1(x)$ is $N(\mu, 1)$ distribution and $F_2(y)$ is the $N(0, 1)$ distribution, selected values of the probability that r out of n students from the second high school are chosen on a shortlist of length k are presented in Olkin and Stephens (1993).

Summary

The theory of order statistics is essential in statistical analysis and its applications. Order statistics play an important role in inferential problems including estimation of unknown parameters of distributions in considered statistical models and in hypothesis testing. Before 1970s, most studies were on cases where order statistics originated from independent and identically distributed random variables. In the early 1970s however the robustness issues motivated the study of order statistics from outliers models. Recent years have seen the appearance of a number of studies on both single- and multiple-outlier models and more generally on order statistics from independent and nonidentically distributed random variables. The theory of order statistics from independent but nonidentically distributed random variables involves permanents which is similar to that of the determinants but without the alternating sign. Barnet and Lewis (1994) mainly discuss the single-outlier models. In an excellent review article, Balakrishnan (2007) describes more general model of order statistics from independent and nonidentically distributed random variables, including many important issues such as distributional properties, characterizations, estimation, outliers, robustness. Continuing from the International Conference on Order Statistics and Extreme Values, Theory and Applications 18–20 December 2000 in Mysore, India) organized by N. R. Mohan and H. N. Nagaraja, a series of international conferences devoted to order statistics in Warsaw, Poland (2002–04); Izmir, Turkey (2005); Mashad, Iran (2006); Amman, Jordan (2007) and Aachen, Germany (2008) provided international forums for presentation and discussion of topics related to ordered statistical data. In these conferences, both reviews of previously existing results and new results involving order statistics were presented in the context of topics such as approximations, characterizations, distribution theory and probability models, stochastic ordering, inequalities, censoring, statistical inference, applications of ordered data, information and entropies, nonparametric methods, ranked set sampling, and asymptotic theory. The special issue on ordered statistical data, approximations bounds and characterizations of Taylor and Francis' journal *Communications in Statistics-Theory and Methods*, vol. 36, no. 7 edited by I. Bairamov consists of selected articles presented at the international conference OSD-2005, Izmir, Turkey.

The elegant theory of order statistics and general models of ordered statistical data is likely to arouse the interest of many scientists working in the area of statistical theory and applications.

See also: Analysis of Extreme Values in Education; Hypothesis Testing and Confidence Intervals; Markov Chain Monte Carlo; Nonparametric Statistical Methods; Stochastic Processes; Survival Data Analysis.

Bibliography

Andrews, D. F., Bickel, P. J., and Hampel, F. R. *et al.* (1972). *Robust Estimates of Location: Survey and Advances*. Princeton, NJ: Princeton University Press.

Arnold, B., Balakrishnan, N., and Nagaraja, H. N. (1992). *A First Course in Order Statistics*. New York: Wiley.

Bairamov, I. and Petunin, Y. I. (1990). Structure of invariant confidence intervals containing the main distributed mass. (*English translation in Probability and Its Applications*, 1991, **35**(1), 15–26.) *Theorya Veroyatnostey i ee Primeneniya* **35**(1), 15–26.

Balakrishnan, N. (2007). Permanents, order statistics, outliers and robustness. *Revista Matematica Complutense* **20**(1), 7–107.

Balch, W. (1992). Effect of class standing on students' predictions of their final exam scores. *Teaching of Psychology* **19**(3), 136–141.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd edn. Chichester: John Wiley & Sons Ltd.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement* **10**, 237–255.

David, H. A. (1981). *Order Statistics*, 2nd edn. New York: Wiley.

David, H. A. and Ghosh, J. K. (1985). The effect of an outlier on L-estimatots of location and symmetric distributions. *Biometrica* **72**(1), 216–218.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, 3rd edn. New York: Wiley.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Myles, T. and Henderson, R. C. (2002). Medical licensure examination scores: Relationship to obstetrics and gynecology examination scores. *Obstetrics & Gynecology* Part 1, **100**(5), 955–958.

Olkin, I. and Stephens, M. A. (1993). On making the shortlist for the selection of candidates. *International Statistical Review* **61**(3), 477–486.

Sarhan, A. E. and Greenberg, B. G. (1962). *Contributions to Order Statistics*. New York: Wiley.

Tadikamalla, P. R. and Balakrishnan, N. (1998). Computer simulation of order statistics. In Balakrishnan, N. (ed.) *Order Statistics: Theory and Methods*, pp 65–72. Amsterdam: North Holland Elsevier.

Torndike, R. L. (1971). Concept of culture fairness. *Journal of Education Measurement* **8**, 63–70.

Williams, L., Wiebe, E., Yang, K., Ferzli, M., and Miller, C. (2002). In support of pair programming in the introductory computer science course. *Computer Science Education* **12**(3), 197–212.

Further Reading

Ahsanullah, M. and Nevzorov, V. B. (2001). *Ordered Random Variables*. Huntington, NY: Nova Science Publishers.

Ahsanullah, M. and Nevzorov, V. B. (2005). *Order Statistics*. *Examples and Exercises.* Huntington, NY: Nova Science Publishers.

Arnold, B. C. and Balakrishnan, N. (1989). *Relations, Bounds and Approximations for Order Statistics. Lecture Notes in Statistics* **53**. New York: Springer.

Arnold, B. C., Becker, A., Gather, U., and Zahedi, H. (1984). On the Markov property of order statistics. *Journal of Statistical Planning and Inference* **9**, 147–154.

Bairamov, I. (2007). Advances in exceedance statistics based on ordered random variables. In Ahsanullah, M. and Raqab, M. Z. (eds.) *Recent Developments in Ordered Random Variables*, pp 97–119. New York: Nova Science.

Balakrishnan, N. (1986). Order statistics from discrete distributions. *Communications in Statistics-Theory and Methods* **15**, 657–675.

Balakrishnan, N. and Cohen, A. C. (1991). *Order Statistics and Inference. Estimation Methods*. Academic Press.

Barlow, R. E. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing. To Begin With*. Silver Springs.

Barnett, V. (1976). The ordering of multivariate data (with discussion). *Journal of the Royal Statistical Society* A **139**, 318–354.

Davis, C. S. and Stephens, M. A. (1983). Approximate percentage points using Pearson curves. Algorithm AS192. *Applied Statistics* **32**, 322–327.

Eryilmaz, S. (2005). Concomitants in a sequence of independent nonidentically distributed random vectors. *Communication in Statistics-Theory and Methods* **34**, 1925–1933.

Nagaraja, N. (1992). Order statistics from discrete distributions (with discussion). *Statistics* **23**, 189–216.

Relevant Websites

http://en.wikipedia.org – Order statistic.
http://planetmath.org – PlanetMath. org.
http://mathworld.wolfram.com – Wolfrom Mathword, order statistic.