Ann. Inst. Statist. Math.
Vol. 47, No. 1, 185–193 (1995)

# MAXIMUM VARIANCE OF ORDER STATISTICS

NICKOS PAPADATOS

*Department of Mathematics, Section of Statistics and Operations Research, University of Athens, Panepistemiopolis, 15784 Athens, Greece*

(Received June 30, 1993; revised May 17, 1994)

**Abstract.** Yang (1982, *Bull. Inst. Math. Acad. Sinica*, **10**(2), 197–204) proved that the variance of the sample median cannot exceed the population variance. In this paper, the upper bound for the variance of order statistics is derived, and it is shown that this is attained by Bernoulli variates only. The proof is based on Hoeffding's identity for the covariance.

*Key words and phrases*: Variance bounds, order statistics, Bernoulli variates, Hoeffding's identity.

## 1. Introduction

Let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be the order statistics corresponding to $n$ iid rv's $X_1, \ldots, X_n$ with df $F(x)$ and finite variance $\sigma^2$.

The purpose of this paper is to find the maximum variance of the $k$-th order statistic $X_{k:n}$ given a fixed population variance $\sigma^2$. Several authors have studied non-parametric bounds for the moments of order statistics. The earliest results in that direction are by Plackett (1947) and Moriguti (1951), generalized by Hartley and David (1954) and Gumbel (1954). More general results are obtained by a method due to Moriguti (1953). Furthermore, moment inequalities for order statistics are given by Sugiura (1962), David and Groeneveld (1982), Terrell (1983), Székely and Móri (1985), Papathanasiou (1990), Balakrishnan (1990) and Gajek and Gather (1991).

Yang (1982) proved that,

(i) for $n = 2k - 1$, $\mathrm{Var}(X_{k:n}) \leq \sigma^2$,

(ii) for $n = 2k$, $\mathrm{Var}[(X_{k:n} + X_{k+1:n})/2] \leq \sigma^2$,

(iii) for $k \neq (n+1)/2$, there exists a continuous df $F(x)$ such that $\mathrm{Var}(X_{k:n}) > \sigma^2$.

Recently Lin and Huang (1989) observed that equality is attained in (i) by the symmetric Bernoulli variate.

Here it will be shown that

$$\mathrm{Var}(X_{k:n}) \leq \sigma_n^2(k) \cdot \sigma^2, \qquad 1 \leq k \leq n, \tag{1.1}$$

where $\sigma_n^2(k)$ in a constant depending only on $k$ and $n$. Equality in (1.1) is attained if and only if $1 < k < n$ and $X$ is a two-valued (Bernoulli) distribution:

$$P[X = x_2] = p = 1 - P[X = x_1], \qquad x_1 < x_2,$$

where the parameter $p = p_n(k)$ depends on $k$ and $n$ only. Cases $k = 1$ and $k = n$ lead to strict inequality, but for every $\epsilon > 0$, there exists a df $F(x)$ such that $\mathrm{Var}(X_{1:n}) > \sigma_n^2(1) \cdot \sigma^2 - \epsilon$ for $k = 1$ and similarly for $k = n$. The key role to the derivation of these upper bounds is played by Hoeffding's identity for the covariance (see (3.2)) in combination with the unimodal property of the special function $t(x)$ (see notations (2.1)), proved in Lemma 2.1(iv).

Unfortunately, the $\sigma_n^2(k)$-values, given by:

$$\sigma_n^2(k) = \sup_{0<x<1} \left\{ \frac{I_x(k, n+1-k) \cdot (1 - I_x(k, n+1-k))}{x(1-x)} \right\} \tag{1.2}$$

do not have a simple form; $I_x(a, b)$ denotes, as usual, the incomplete beta function:

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x u^{a-1}(1-u)^{b-1} du, \qquad 0 < x < 1.$$

However, tables and figures for $\sigma_n^2(k)$ are obtained for various $n$ and $k$.

## 2. An auxiliary lemma

Let $X_{k:n}$ be the $k$-th order statistic based on a sample of size $n \geq 2$, from an arbitrary df $F(x)$ with finite variance $\sigma^2$. Then (c.f., for example, David (1981), p. 34) $\mathrm{Var}(X_{k:n})$ exists. We are interested in finding the maximum variance of $X_{k:n}$ among all df's $F$ having variance $\sigma^2$. Let us use the notations:

$$\begin{aligned} &G(x) = I_x(k, n+1-k), \qquad g(x) = G'(x), \\ &t_1(x) = \frac{G(x)}{x}, \qquad t_2(y) = \frac{1 - G(y)}{1-y}, \\ &t(x,y) = t_1(x) \cdot t_2(y) \quad \text{and} \quad t(x) = t(x,x), \qquad 0 < x \leq y < 1. \end{aligned} \tag{2.1}$$

The following lemma is required for the main result.

LEMMA 2.1. *Let $1 < k < n$. Then, there exist unique numbers $\rho_1 = \rho_1(k,n)$, $\rho_2 = \rho_2(k,n)$ satisfying*

$$0 < \rho_1 < \frac{k-1}{n-1} < \rho_2 < 1$$

*such that, for $0 < x < y < 1$:*

(i) *$t_1(x) = G(x)/x$ strictly increases in $(0, \rho_2)$ and strictly decreases in $(\rho_2, 1)$ and similarly $t_2(y) = (1 - G(y))/(1-y)$ strictly increases in $(0, \rho_1)$ and strictly decreases in $(\rho_1, 1)$.*

(ii) *If $x \geq \rho_1$ or $y \leq \rho_2$, then*

$$\frac{G(x)\cdot(1-G(y))}{x(1-y)} = t(x,y) < \max\{t(x),t(y)\}.$$

(iii) *If $x < \rho_1$ and $y > \rho_2$, then*

$$t(x,y) < t(\rho_1,\rho_2) < \max\{t(\rho_1),t(\rho_2)\}.$$

(iv) *There exists a unique $x_0 = x_0(k,n) \in (\rho_1,\rho_2)$ such that the function $\frac{G(x)\cdot(1-G(x))}{x(1-x)} = t(x)$ strictly increases in $(0,x_0)$ and strictly decreases in $(x_0,1)$.*

PROOF. (i) Obviously $t_1'(x) = (xg(x)-G(x))/x^2$. The function $xg(x)-G(x)$ has derivative $xg'(x)$ which is positive if $x < (k-1)/(n-1)$ and negative if $x > (k-1)/(n-1)$. Since $\lim_{x\to 0+}[xg(x)-G(x)] = 0$, $\lim_{x\to 1-}[xg(x)-G(x)] = -1$ we conclude that the equation

$$xg(x) - G(x) = 0, \qquad 0 < x < 1,$$

has a unique root $\rho_2 = \rho_2(k,n) \in ((k-1)/(n-1),1)$.

Hence $xg(x) - G(x) > 0$ for $x \in (0,\rho_2)$ and $xg(x) - G(x) < 0$ for $x \in (\rho_2,1)$ and the proof is complete for $t_1(x)$.

Similarly for $t_2(y)$, there exists a unique $\rho_1 = \rho_1(k,n)$ satisfying

$$0 < \rho_1 < \frac{k-1}{n-1}$$

such that $t_2(y)$ strictly increases in $(0,\rho_1)$ and strictly decreases in $(\rho_1,1)$.

Note that $\rho_1$ is the unique point in $(0,1)$ satisfying $1 - G(y) = (1-y)g(y)$.

(ii) Let $x < y$. If $\rho_1 \leq x$, we have

$$t(x,y) = t_1(x)\cdot t_2(y) < t_1(x)\cdot t_2(x) = t(x).$$

Similarly if $y \leq \rho_2$, $t(x,y) < t(y)$ and the proof is complete.

(iii) If $x < \rho_1$ and $y > \rho_2$, we have

$$t(x,y) = t_1(x)\cdot t_2(y) < t_1(\rho_1)\cdot t_2(\rho_2) = t(\rho_1,\rho_2).$$

The second inequality follows from (ii).

(iv) Obviously $\lim_{x\to 0+} t(x) = \lim_{x\to 1-} t(x) = 0$.

Furthermore, it is clear from (i) that the function $t(x)$ strictly increases in $(0,\rho_1]$ and strictly decreases in $[\rho_2,1)$. It suffices to study $t(x)$ in $(\rho_1,\rho_2)$.

It is easy to verify that for $0 < x < 1$, $x^2g^2(x) - xg(x)G(x) - x^2g'(x)G(x) > 0$. Indeed, the function $nG(x) - xg(x)$ strictly increases (because $(nG(x)-xg(x))' = (n-k)g(x)/(1-x) > 0$), so that $nG(x) - xg(x) > \lim_{x\to 0+}[nG(x) - xg(x)] = 0$. Hence, the function $x(1-x)g(x) - (k-nx)G(x)$ increases also (because $(x(1-$

$x)g(x) - (k - nx)G(x))' = nG(x) - xg(x) > 0$ by the above argument). Therefore, $x(1-x)g(x) - (k-nx)G(x) > \lim_{x\to 0+}[x(1-x)g(x) - (k-nx)G(x)] = 0$, so that

$$\begin{aligned} &x^2 g^2(x) - xg(x)G(x) - x^2 g'(x)G(x) \\ &\quad = \frac{xg(x)}{1-x}[x(1-x)g(x) - (k-nx)G(x)] > 0, \qquad 0 < x < 1. \end{aligned}$$

But, for $0 < x < \rho_2$, $G(x) < xg(x)$ (see (i)), so that

$$x^2 g^2(x) - G^2(x) - x^2 g'(x)G(x) > 0, \qquad 0 < x < \rho_2. \tag{2.2}$$

We observe that

$$\left[\log \frac{G(x)}{x}\right]'' = \frac{-1}{x^2 G^2(x)}(x^2 g^2(x) - G^2(x) - x^2 g'(x)G(x));$$

thus, from (2.2), $(\log \frac{G(x)}{x})'' < 0$, $0 < x < \rho_2$, that is, $t_1(x)$ is strictly log-concave in $(0, \rho_2)$.

By the same arguments it is proved that $t_2$ is strictly log-concave in $(\rho_1, 1)$.

Hence $t(x) = t_1(x) \cdot t_2(x)$ is a strictly log-concave function in $(\rho_1, \rho_2)$, and the proof is complete.

*Remark.* The unique point $x_0 = x_0(k, n)$ satisfies the equation

$$\frac{g(x)(1 - 2G(x))}{G(x)(1 - G(x))} = \frac{1 - 2x}{x(1 - x)} \tag{2.3}$$

and obviously

$$\frac{G(x_0)(1 - G(x_0))}{x_0(1 - x_0)} = \sup_{0<x<1} \left[\frac{G(x)(1 - G(x))}{x(1 - x)}\right]. \tag{2.4}$$

## 3. Main result

DEFINITION 3.1. We define the maximum variance function $\sigma_n^2(k)$ by the relation

$$\sigma_n^2(k) = \sup_{0<x<1} \left[\frac{G(x)(1 - G(x))}{x(1 - x)}\right], \qquad 1 \le k \le n, \quad n = 2, 3, \ldots$$

(see (1.2)).

Clearly, $\sigma_n^2(k)$ is a function of $k$ for $n \ge 2$ fixed. Moreover, from (2.4)

$$\sigma_n^2(k) = \frac{G(x_0)(1 - G(x_0))}{x_0(1 - x_0)} = t(x_0), \qquad 1 < k < n,$$

while $\sigma_n^2(1) = \sigma_n^2(n) = n$.

We can now state the main result of

THEOREM 3.1. (Maximum variance of order statistics)

$$\mathrm{Var}(X_{k:n}) \le \sigma_n^2(k)\sigma^2, \tag{3.1}$$

*and equality is attained if and only if* $1 < k < n$ *and* $F(x)$ *is a Bernoulli distribution with probability of success* $1 - x_0(k,n)$ ($x_0(k,n)$ *is defined in Lemma* 2.1(iv)). *Cases* $k = 1$ *and* $k = n$ *yield strict inequality, though* (3.1) *yields the best upper bound for* $\mathrm{Var}(X_{1:n})$ *and* $\mathrm{Var}(X_{n:n})$.

PROOF. Suppose $1 < k < n$. *Hoeffding's identity* for the covariance of $X$, $Y$ is given by the relation (for a proof see Lehmann (1966), Lemma 2)

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} [H(x,y) - H(x,\infty)H(\infty,y)]dydx, \tag{3.2}$$

where $H(x,y)$ is the bivariate df of $(X,Y)'$. Hence

$$\mathrm{Var}(X) = 2\iint_{x\le y} F(x)(1-F(y))dydx = \sigma^2 \tag{3.3}$$

and similarly

$$\mathrm{Var}(X_{k:n}) = 2\iint_{x\le y} G(F(x))(1-G(F(y)))dydx$$

(the last relation follows from (3.3) and the fact that $G(F(x))$ is just the df of $X_{k:n}$).

From Lemma 2.1 we conclude that for all $x \le y$,

$$x_0(1-x_0)G(F(x))(1-G(F(y))) \le G(x_0)(1-G(x_0))F(x)(1-F(y)) \tag{3.4}$$

and equality is attained if and only if either $F(x)(1-F(y)) = 0$ or $F(x) = F(y) = x_0$.

Integrating (3.4) over $S = \{-\infty < x \le y < +\infty\}$ we have the desired result. In order to hold (3.1) as an equality, it is necessary and sufficient that

$$x_0(1-x_0)G(F(x))(1-G(F(y))) = G(x_0)(1-G(x_0))F(x)(1-F(y)) \quad \text{a.e. in } S$$

or, equivalently

$$F(x)(1-F(y)) = 0 \quad \text{or} \quad F(x) = F(y) = x_0 \quad \text{a.e. in } S.$$

Thus, the only nondegenerate (with variance $\sigma^2$) df which attains equality is of the form:

$$F(x;x_1,x_2)=\begin{cases}0 & \text{if } x<x_1\\ x_0 & \text{if } x_1\le x<x_2\\ 1 & \text{if } x_2\le x\end{cases}$$

where $x_2=x_1+\sigma/\sqrt{x_0(1-x_0)}$, $x_1\in\mathbb{R}$, and the proof is complete for $1<k<n$.

For $k=n$ we have

$$\operatorname{Var}(X_{n:n})\le\sum_{i=1}^{n}\operatorname{Var}(X_{i:n})+2\sum\sum_{1\le i<j\le n}\operatorname{Cov}(X_{i:n},X_{j:n})=n\sigma^2,$$

and obviously equality is attained iff $\sigma^2=0$.

Therefore, for nondegenerate $F$,

$$\operatorname{Var}(X_{n:n})<n\sigma^2=\sigma_n^2(n)\sigma^2$$

and this bound is the best possible (for example take $F$ to be a two-valued distribution with probability of success close to zero).

Case $k=1$ is similar to $k=n$ and is omitted.

Note that $\sigma_n^2(k)$ is symmetric about $(n+1)/2$:

$$\sigma_n^2(k)=\sigma_n^2(n+1-k),$$

decreases for $k\le(n+1)/2$ and increases for $k\ge(n+1)/2$ taking the values $\sigma_n^2(1)=\sigma_n^2(n)=n$, $\sigma_n^2((n+1)/2)=1$ (see Fig. 1).

*Remark.* Figure 1 identifies the upper bound for each order statistic separately. Note that these upper bounds can never be achieved simultaneously. For a given (fixed) distribution from a lot of distribution families, the variances of order statistics have a bell-shaped curve, in contrast to the possible misunderstanding that the U-shaped curve, presented by Fig. 1, may creates.

Note that for $k>(n+1)/2$, the values $\sigma_n^2(k)$ and $x_0(k,n)$ can be calculated from Table 1 using the relations:

$$\sigma_n^2(k)=\sigma_n^2(n+1-k),\qquad x_0(k,n)=1-x_0(n+1-k,n)$$

and that $p=1-x_0(k,n)$ is the parameter of the Bernoulli rv which maximizes $\operatorname{Var}(X_{k:n})/\operatorname{Var}(X)$. For example, if $n=10$, $k=2$ we find from Table 1:

$$\sigma_{10}^2(2)=2.1608,\qquad 1-x_0(2,10)=0.8958=p$$

hence for any df we have

$$\operatorname{Var}(X_{2:10})\le 2.1608\cdot\sigma^2$$

Table 1. Values of $10^4 \cdot \sigma_n^2(k)$ and $10^4 \cdot (1 - x_0(k,n))$ for various $n$ and $k < (n+1)/2$.

| $n$ \ $k$ | 1.25 | 1.50 | 1.75 | 2.00 | 2.25 | 2.50 | 2.75 | 3.00 | 3.25 | 3.50 | 3.75 | 4.00 | 4.25 | 4.50 | 4.75 | 5.00 | 5.25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 11283 | | | | | | | | | | | | | | | | |
| | 7678 | | | | | | | | | | | | | | | | |
| 3 | 15524 | 11867 | 10409 | | | | | | | | | | | | | | |
| | 8901 | 7638 | 6327 | | | | | | | | | | | | | | |
| 4 | 20096 | 14716 | 12197 | 10870 | 10204 | | | | | | | | | | | | |
| | 9283 | 8464 | 7612 | 6747 | 5875 | | | | | | | | | | | | |
| 5 | 24749 | 17755 | 14351 | 12408 | 11225 | 10510 | 10123 | | | | | | | | | | |
| | 9468 | 8863 | 8234 | 7595 | 6949 | 6301 | 5651 | | | | | | | | | | |
| 6 | 29432 | 20863 | 16626 | 14141 | 12553 | 11497 | 10790 | 10337 | 10082 | | | | | | | | |
| | 9578 | 9098 | 8600 | 8093 | 7582 | 7068 | 6552 | 6035 | 5518 | | | | | | | | |
| 7 | 34132 | 24005 | 18957 | 15958 | 14004 | 12660 | 11708 | 11030 | 10555 | 10239 | 10059 | | | | | | |
| | 9650 | 9252 | 8840 | 8421 | 7998 | 7572 | 7145 | 6717 | 6288 | 5859 | 5430 | | | | | | |
| 8 | 38840 | 27165 | 21317 | 17819 | 15516 | 13907 | 12741 | 11877 | 11233 | 10756 | 10412 | 10179 | 10044 | | | | |
| | 9701 | 9362 | 9010 | 8653 | 8292 | 7929 | 7564 | 7199 | 6833 | 6467 | 6100 | 5734 | 5367 | | | | |
| 9 | 43554 | 30336 | 23694 | 19705 | 17062 | 15201 | 13384 | 12804 | 12015 | 11405 | 10936 | 10579 | 10318 | 10139 | 10034 | | |
| | 9739 | 9443 | 9137 | 8825 | 8510 | 8194 | 7876 | 7558 | 7239 | 6920 | 6600 | 6280 | 5960 | 5640 | 5320 | | |
| 10 | 48272 | 33514 | 26083 | 21608 | 18631 | 16523 | 14965 | 13778 | 12856 | 12129 | 11553 | 11096 | 10736 | 10459 | 10253 | 10111 | 10028 |
| | 9768 | 9506 | 9235 | 8958 | 8680 | 8399 | 8118 | 7836 | 7553 | 7270 | 6987 | 6703 | 6419 | 6136 | 5852 | 5568 | 5284 |

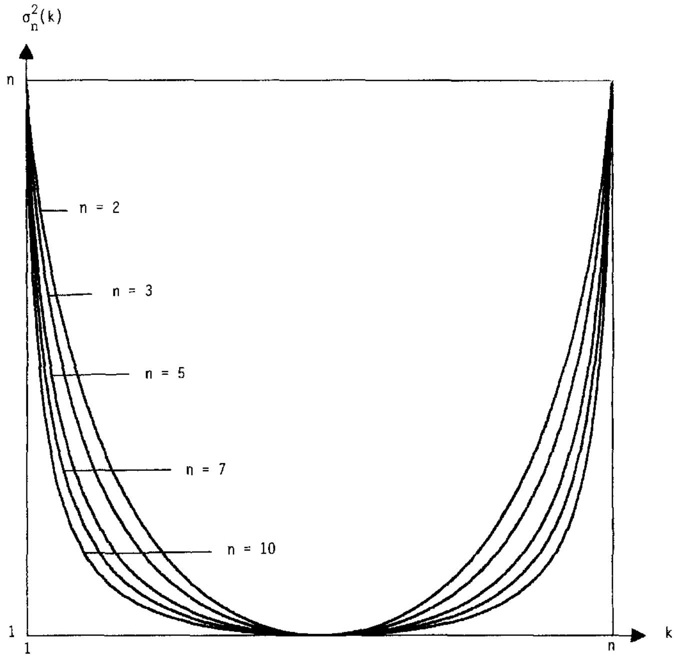


Fig. 1. The function $\sigma_n^2(k) = \sup\{\mathrm{Var}(X_{k:n})/\mathrm{Var}(X)\}$.

with equality if and only if $X_1, X_2, \ldots, X_{10}$ are 10 independent copies of the rv $X$ with

$$P[X = x_2] = p = 1 - P[X = x_1], \qquad x_1 < x_2. \tag{3.5}$$

The same inequality is true for $X_{9:10}$, but equality is attained if and only if $X$ is as in (3.5) with $x_1 > x_2$.

Throughout this paper, $k$ was assumed to be an integer in $\{1, 2, \ldots, n\}$. However, all the results continue to hold also for non-integer $k$. In this case, $X_{k:n}$ denotes the $k$-th intermediate order statistic as defined by Papadatos (1994). Values of $\sigma_n^2(k)$ and $x_0(k, n)$ for integer and non-integer $k$ are given in Table 1 and Fig. 1.


## Acknowledgements

The author would like to thank Professors T. Cacoullos, V. Papathanasiou and E. Kounias for their suggestions and helpful comments. Thanks are also due to an anonymous referee for his helpful informations about Fig. 1.


## References


Balakrishnan, N. (1990). Improving the Hartley-David-Gumbel bound for the mean of extreme order statistics, *Statist. Probab. Lett.*, **9**, 291–294.

David, H. (1981). *Order Statistics*, 2nd ed., Wiley, New York.

David, H. and Groeneveld, R. (1982). Measures of local variation in a distribution: expected length of spacing and variances of order statistics, *Biometrika*, **69**, 227–232.

Gajek, L. and Gather, U. (1991). Moment inequalities for order statistics with applications to characterizations of distributions, *Metrika*, **38**, 357–367.

Gumbel, E. (1954). The maxima of the mean largest value and of the range, *Ann. Math. Statist.*, **25**, 76–84.

Hartley, H. and David, H. (1954). Universal bounds for mean range and extreme observation, *Ann. Math. Statist.*, **25**, 85–99.

Lehmann, E. L. (1966). Some concepts of dependence, *Ann. Math. Statist.*, **37**, 1137–1153.

Lin, G. and Huang, J. (1989). Variances of sample medians, *Statist. Probab. Lett.*, **8**, 143–146.

Moriguti, S. (1951). Extremal property of extreme value distributions, *Ann. Math. Statist.*, **22**, 523–536.

Moriguti, S. (1953). A modification of Schwarz's inequality with applications to distributions, *Ann. Math. Statist.*, **24**, 107–113.

Papadatos, N. (1994). Intermediate order statistics with applications to nonparametric estimation, *Statist. Probab. Lett.* (to appear).

Papathanasiou, V. (1990). Some characterizations of distributions based on order statistics, *Statist. Probab. Lett.*, **9**, 145–147.

Plackett, R. (1947). Limits of the ratio of mean range to standard deviation, *Biometrika*, **34**, 120–122.

Sugiura, N. (1962). On the orthogonal inverse expansion with an application to the moments of order statistics, *Osaka Math. J.*, **14**, 253–263.

Székely, G. and Móri, T. (1985). An extremal property of rectangular distributions, *Statist. Probab. Lett.*, **3**, 107–109.

Terrell, G. (1983). A characterization of rectangular distributions, *Ann. Probab.*, **11**, 823–826.

Yang, H. (1982). On the variances of median and some other order statistics, *Bull. Inst. Math. Acad. Sinica*, **10**(2), 197–204.