Miss Cox

AN UPPER BOUND FOR THE VARIANCE OF CERTAIN STATISTICS

by

Wassily Hoeffding
University of North Carolina

This research was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under Contract No. AF 18(600)-458. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Institute of Statistics
Mimeograph Series No. 193
March, 1958

AN UPPER BOUND FOR THE VARIANCE OF CERTAIN STATISTICS[1]

by

Wassily Hoeffding
University of North Carolina

1. Results. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables (real- or vector-valued). Let $f(X_1, X_2)$ be a bounded function such that $f(X_1, X_2) = f(X_2, X_1)$. With no loss of generality we shall assume that the bounds are $0 \leq f(X_1, X_2) \leq 1$. Let

$$U = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} f(X_i, X_j).$$

Examples of statistics of this form are given below. The mean of U is

$$E(U) = p$$

and the variance is

$$\operatorname{var}(U) = \frac{2}{n(n-1)} \{2(n-2)(r-p^2) + s - p^2\} ,$$

where

$$p = E\, f(X_1, X_2) ,$$

$$r = E\, f(X_1, X_2)\, f(X_1, X_3), \qquad s = E\, f^2(X_1, X_2) .$$

As n tends to infinity, $\sqrt{n}\,(U-p)$ has a normal limiting distribution [2]. Hence if we have an upper bound for the variance of U which depends only on p and n, we can obtain an approximate confidence region for p and a lower bound for the power of a test based on U when n is large (see [3]).

---

[1] This research was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under Contract No. AF 18(600)-458. Reproduction in whole or in part is permitted for any purpose of the United States Government.

It is known [2] that $2(r-p^2) \leq s - p^2$, and since obviously $s \leq p$, we have

$$\operatorname{var}(U) \leq \frac{2}{n} p(1-p).$$

In this note we shall show that, under the stated assumptions,

$$(1) \qquad r - p^2 \leq H(p) = \begin{cases} p^{3/2} & -p^2 , \quad p \geq \frac{1}{2} \\ (1-p)^{3/2} & -(1-p)^2 , \quad p \leq \frac{1}{2} \end{cases} .$$

It is easily seen that the sign of equality holds in (1) if, with probability one, $f(X_1,X_2) = g(X_1)g(X_2)$ (for $p \geq \frac{1}{2}$) or $f(X_1,X_2) = 1 - g(X_1)g(X_2)$ (for $p \leq \frac{1}{2}$), where $g(X)$ takes the values 0 and 1 only.

Inequality (1) implies that

$$(2) \qquad \operatorname{var}(U) \leq \frac{2}{n(n-1)} \{2(n-2) H(p) + 1 - p^2\} .$$

An inequality analogous to (1) was conjectured by Daniels and Kendall [1] for the variance of the finite population analogue of the statistic t defined in Example 1 below. A proof of this conjecture suggested by Sundrum [4] does not seem to be complete.

We now give three examples of statistics to which the present bound is applicable; in the first two examples the bound can be attained.

Example 1. Let $X_i = (Y_i,Z_i)$ be a random vector with two continuously distributed components, and $f(X_1,X_2) = 1$ or 0 according as $(Y_1-Y_2)(Z_1-Z_2)$ is positive or negative. In this case $t = 2U - 1$ is a well-known rank correlation coefficient. The condition for equality in (1) is satisfied if Z is a function of Y of a certain form, for instance positive and decreasing for $Y < 0$ and negative and increasing

for $Y > 0$ (if $p \geq \frac{1}{2}$); or negative and increasing for $Y < 0$ and positive and decreasing for $Y > 0$ (if $p \leq \frac{1}{2}$). For if we let $g(X) = 0$ or 1 according as $Y < 0$ or $Y > 0$, then, with probability one, $f(X_1,X_2) = g(X_1)g(X_2)$ in the first case and $f(X_1,X_2) = 1 - g(X_1)g(X_2)$ in the second case. These two cases correspond to the inverse canonical ranking and the canonical ranking as defined by Daniels and Kendall [1] for finite populations.

Example 2. Let X_i be a real-valued random variable, $f(X_1,X_2) = 0$ or 1 according as $X_1 + X_2 < 0$ or > 0. Then $p = \frac{1}{2}$ if X_i has the same continuous distribution as $-X_i$. Thus a test based on U might serve to detect certain deviations from symmetry. The sign of equality in (1) is attained, for instance, if X_i can take only two values, a and b, such that either $a + b < 0 < b$ (if $p \geq \frac{1}{2}$) or $a < 0 < a + b$ (if $p \leq \frac{1}{2}$). Here we may take $g(X) = 0$ or 1 according as $X < 0$ or $X > 0$ in the first case, and $g(X) = 0$ or 1 according as $X > 0$ or $X < 0$ in the second case.

Example 3. Let X_i again be real-valued, and let

$$f(X_1,X_2) = 1 - 2 \max[F_0(X_1), F_0(X_2)] + F_0^2(X_1) + F_0^2(X_2) ,$$

where $F_0(x)$ is a continuous (cumulative) distribution function. Then, if X_i has the distribution function $F(x)$,

$$p = \frac{1}{3} + 2 \int [F(x) - F_0(x)]^2 \, dF_0(x) ,$$

and $(U - \frac{1}{3})/2$ differs in large samples negligible little from the Cramer-von Mises goodness of fit criterion $\int [F_n(x) - F_0(x)]^2 \, dF_0(x)$, where $n F_n(x)$ denotes the number of observations $\leq x$. In this case the

condition for equality in (1) cannot be satisfied, and presumably the upper bound for the variance can be further improved.

2. Proof of Inequality (1). We first assume that

$$f(X_1,X_2) = 0 \text{ or } 1. \tag{3}$$

Let $X_1,X_2,\ldots,X_n$ be independent and identically distributed random variables,

$$f_{ij} = f(X_i,X_j), \quad i \neq j, \qquad f_{ii} = 0,$$

$$\hat{p} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij}, \qquad \hat{r} = n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} f_{ij} f_{ik} .$$

As $n \longrightarrow \infty$, $\hat{p}$ and $\hat{r}$ converge in probability to p and r respectively

We shall show that

$$\hat{r} - \hat{p}^2 \leq H(\hat{p}) + \epsilon_n , \tag{4}$$

where H(p) is the function defined in (1) and ϵ_n are numbers which converge to 0 as $n \longrightarrow \infty$. Since the function H(p) is continuous, inequality (4) easily implies inequality (1).

Both $\hat{p}$ and $\hat{r}$ are functions of the nxn matrix $\|f_{ij}\|$ whose elements satisfy the conditions

$$f_{ij} = 0 \text{ or } 1, \qquad f_{ii} = 0, \qquad f_{ij} = f_{ji} . \tag{5}$$

Let

$$F_i = \sum_{j=1}^{n} f_{ij} .$$

Then

$$n^2 \hat{p} = \sum_{i=1}^{n} F_i, \qquad n^3 \hat{r} = \sum_{i=1}^{n} F_i^2.$$

We first show that, in order to find an upper bound for $\hat{r}$ when $\hat{p}$ is fixed, we may assume that

(6) $\qquad F_i \geq F_j$ implies $f_{ik} \geq f_{jk}$ for all $k \neq i$.

For suppose that there are integers i, j, k such that $F_i \geq F_j$, $k \neq i$, and $f_{ik} < f_{jk}$. Thus $f_{ik} = 0$, $f_{jk} = 1$. Let $\|f'_{uv}\|$ be the nxn matrix defined by $f'_{ik} = f'_{ki} = 1$, $f'_{jk} = f'_{kj} = 0$, $f'_{uv} = f_{uv}$ otherwise. The transformed matrix satisfies (5), and the value of $\hat{p}$ is not affected by the transformation. Also, writing F'_u for the sum of the u-th row in the new matrix, $F'_i = F_i + 1$, $F'_j = F_j - 1$, $F'_u = F_u$ otherwise. Therefore

$$\Sigma(F'_u)^2 - \Sigma F_u^2 = (F_i+1)^2 - F_i^2 + (F_j-1)^2 - F_j^2$$

$$= 2(F_i-F_j) + 2 > 0.$$

Thus the value of $\hat{r}$ is increased by the transformation. By repeated application of this transformation we can obtain a matrix which satisfies (6), for instance as follows. We first take one of the rows with the largest sum to be the i-th row, and for every $j \neq i$ we apply the transformation for each $k \neq i$ with $f_{ik} < f_{jk}$. Having exhausted all k and all j, we are left with a matrix such that every column different from the i-th that has a 0 in the i-th place will consist of zeros only. This implies that any further transformations will not affect this i-th row. We next take one of the rows with the largest sum among the remaining n-1 rows, and repeat the procedure described above, and

so on. Eventually we obtain a matrix which satisfies the conditions (5) and (6) without changing the value of $\hat{p}$ and without decreasing the value of $\hat{r}$.

With no loss of generality we may assume that, in addition to (5) and (6), the matrix $\|f_{ij}\|$ satisfies the condition

$$F_1 \geq F_2 \geq \cdots \geq F_n \,. \tag{7}$$

For this can always be achieved by rearranging suitable rows and columns, which will not affect the values of $\hat{p}$ and $\hat{r}$. Thus we now restrict ourselves to symmetric matrices whose every row and every column, apart from a 0 in the main diagonal, consists of a sequence of 1's followed by a sequence of 0's. A typical matrix of this form is shown below.

0	1	1	1	1	1	0
1	0	1	1	0	0	0
1	1	0	1	0	0	0
1	1	1	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	0

For a fixed matrix $\|f_{ij}\|$ of this type let m denote the greatest integer such that $f_{m,m+1} = 1$. Then $f_{ij} = 1$ if $i \leq m+1$, $j \leq m+1$ $(i \neq j)$, and $f_{ij} = 0$ if $i \geq m+1$, $j \geq m+1$. Hence

$$n^2 \hat{p} = \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} + \sum_{i=1}^{m} \sum_{j=m+1}^{n} f_{ij} + \sum_{i=m+1}^{n} \sum_{j=1}^{m} f_{ij}$$

$$= m(m-1) + 2 \sum_{i=1}^{m} \sum_{j=m+1}^{n} f_{ij}$$

and

$$n^3 \hat{r} = \sum_{i=1}^{m} \Big(\sum_{j=1}^{m} f_{ij} + \sum_{j=m+1}^{n} f_{ij} \Big)^2 + \sum_{i=m+1}^{n} \Big(\sum_{j=1}^{m} f_{ij} \Big)^2$$

$$= \sum_{i=1}^{m} \Big(m - 1 + \sum_{j=m+1}^{n} f_{ij} \Big)^2 + \sum_{i=m+1}^{n} \Big(\sum_{j=1}^{m} f_{ij} \Big)^2$$

$$= m(m-1)^2 + 2(m-1) \sum_{i=1}^{m} \sum_{j=m+1}^{n} f_{ij} + \sum_{i=1}^{m} \Big(\sum_{j=m+1}^{n} f_{ij} \Big)^2 + \sum_{j=m+1}^{n} \Big(\sum_{i=1}^{m} f_{ij} \Big)^2 .$$

If we put

$$a_i = \frac{1}{n-m} \sum_{j=m+1}^{n} f_{ij}, \qquad b_j = \frac{1}{m} \sum_{i=1}^{m} f_{ij},$$

$$d = \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} f_{ij} = \frac{1}{m} \sum_{i=1}^{m} a_i = \frac{1}{n-m} \sum_{j=m+1}^{n} b_j,$$

we obtain

$$n^2 \hat{p} = m(m-1) + 2\, m(n-m)d \tag{8}$$

and

$$n^3 \hat{r} = (m-1)\, n^2 \hat{p} + (n-m)^2 \sum_{i=1}^{m} a_i^2 + m^2 \sum_{j=m+1}^{n} b_j^2$$

$$= (m-1)\, n^2 \hat{p} + m(n-m)n\, d^2 \tag{9}$$

$$+ m(n-m) \Big[(n-m) \frac{1}{m} \sum_{i=1}^{m} (a_i - d)^2 + m \frac{1}{n-m} \sum_{j=m+1}^{n} (b_j - d)^2 \Big]$$

Now

$$\frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \big[(f_{ij} - d) - (a_i - d) - (b_j - d) \big]^2$$

$$= \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} (f_{ij} - d)^2 - \frac{1}{m} \sum_{i=1}^{m} (a_i - d)^2 - \frac{1}{n-m} \sum_{j=m+1}^{n} (b_j - d)^2$$

$$= d - d^2 - \frac{1}{m} \sum_{i=1}^{m} (a_i - d)^2 - \frac{1}{n-m} \sum_{j=m+1}^{n} (b_j - d)^2 .$$

Since the left hand side is nonnegative, we obtain [2]

$$\frac{1}{m}\sum_{i=1}^{m}(a_i-d)^2+\frac{1}{n-m}\sum_{j=m+1}^{n}(b_j-d)^2\le d-d^2. \tag{10}$$

It now follows from (9) and (10) that

$$\begin{aligned} n^3\hat{r} &\le (m-1)\,n^2\hat{p}+m(n-m)\,n\,d^2 \\ &\quad + m(n-m)\max(m,\,n-m)(d-d^2). \end{aligned} \tag{11}$$

We thus have obtained an upper bound for $\hat{r}$ in terms of $\hat{p}$ for all matrices $\|f_{ij}\|$ with m fixed. We now put

$$c=\frac{m}{n}.$$

By equation (8), since $0\le d\le 1$, the range of c is given by

$$c^2\le\hat{p}+\frac{c}{n},\qquad (1-c)^2\le 1-\hat{p}-\frac{c}{n}. \tag{12}$$

We therefore may assume that c is bounded away from 0 and 1 as $n\longrightarrow\infty$. Indeed, inequality (1) is trivially true if $p=0$ or 1; and if $0<p<1$, since $\hat{p}$ tends to p in probability, inequalities (12) imply that c is bounded away from 0 and 1 with a probability which approaches one as n tends to infinity.

From (11) we obtain

$$\hat{r}\le c\,\hat{p}+c(1-c)d^2+c(1-c)\max c(1-c)(d-d^2)+\epsilon_n',$$

where $\epsilon_n'\longrightarrow 0$ as $n\longrightarrow\infty$. If we express d in terms of $\hat{p}$ and c by

---

[2] Inequalities (10) and (11) are closely related to the sharp upper bound for the variance of the Wilcoxon-Mann-Whitney two-sample statistic in terms of its mean; see D. van Dantzig, "On the consistency and the power of Wilcoxon's two sample test," Nederl. Akad. Wetensch. Proc. Ser. A, Vol. 54, No. 1, pp. 1-8 (1951), footnote 4.

means of (8), we have

$$\hat{r} - \hat{p}^2 \leq H(\hat{p}, c) + \epsilon_n,$$

where

$$H(p,c) = -p^2 + cp + \frac{p-c^2}{2}\left[\frac{p-c^2}{2c(1-c)} + \max(c, 1-c)\left(1 - \frac{p-c^2}{2c(1-c)}\right)\right]$$

and $\epsilon_n \longrightarrow 0$ as $n \longrightarrow \infty$.

We now maximize $H(p, c)$ with respect to c. By (12) we may assume that

$$c \leq p^{1/2}, \qquad 1-c \leq (1-p)^{1/2}.$$

We can write

$$H(p,c) = G(1-p, 1-c), \; c \leq \tfrac{1}{2}; \quad H(p,c) = G(p,c), \; c \geq \tfrac{1}{2},$$

where

$$G(p,c) = -p^2 + pc + \tfrac{1}{4} p^2 c^{-1} - \tfrac{1}{4} c^3.$$

We calculate

$$G(p, p^{1/2}) - G(p,c) = \frac{1}{4c} (p^{1/2}-c)^2 (c^2 + 2p^{1/2}c - p).$$

If $c \geq \frac{1}{2}$, the right side is positive, and hence $G(p,c) \leq G(p, p^{1/2})$ $= p^{3/2} - p^2$. It follows that

$$H(p,c) \leq \max\left[p^{3/2} - p^2, \; (1-p)^{3/2} - (1-p)^2\right],$$

and the function on the right is just $H(p)$ as defined in (1). This completes the proof for the case where $f(X_1,X_2)$ takes on the values 0 and 1 only.

Now let $f(X_1,X_2)$ be any function which satisfies the assumptions of the theorem. Then $f(X_1,X_2)$ can be approximated by a function $f'(X_1,X_2)$

which also satisfies the assumptions of the theorem and takes only finitely many values, in such a way that $p' = E\, f'(X_1,X_2)$ and $r' = E\, f'(X_1,X_2)\, f'(X_1,X_3)$ are as close as we please to p and r, respectively. It will therefore be sufficient to assume that $f(X_1,X_2)$ takes on the k values $f_1,f_2,\ldots,f_k$. Then it can be written in the form

$$f(X_1,X_2) = \sum_{i=1}^{k} f_i c_i(X_1,X_2),$$

where $0 \leq f_i \leq 1$,

$$c_i(X_1,X_2) = c_i(X_2,X_1) = 0 \text{ or } 1;$$

$$c_i(X_1,X_2)c_j(X_1,X_2) = 0, \qquad i \neq j.$$

Now let Y_1,Y_2,Y_3 be mutually independent random variables, independent of X_1,X_2,X_3, where each Y_i is uniformly distributed on the interval $0 \leq Y_i \leq 1$. The random vectors $Z_j = (X_j, Y_j)$, $j = 1,2,3$, are mutually independent and identically distributed. Now define

$$f^*(Z_1,Z_2) = \sum_{i=1}^{k} d_i(Y_1)\, d_i(Y_2)\, c_i(X_1,X_2),$$

where

$$d_i(Y) = \begin{cases} 1, & 0 \leq Y \leq f_i^{1/2} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$f^*(Z_1,Z_2) = f^*(Z_2,Z_1) = 0 \text{ or } 1.$$

The conditional expected value of $f^*(Z_1,Z_2)$ for X_1,X_2 fixed is

$$E\left[f^*(Z_1,Z_2)\,|X_1,X_2\right] = \sum_{i=1}^{k} f_i c_i(X_1,X_2) = f(X_1,X_2)\ .$$

Hence

$$(13) \qquad E\, f^*(Z_1,Z_2) = E\, f(X_1,X_2) = p.$$

Also

$$f^*(Z_1,Z_2)\, f^*(Z_1,Z_3) = \sum_{i=1}^{k} \sum_{j=1}^{k} d_i(Y_1)\, d_i(Y_2)\, c_i(X_1,X_2)\, d_j(Y_1)\, d_j(Y_3)\, c_j(X_1,X_3)$$

so that

$$E\left[f^*(Z_1,Z_2)\, f^*(Z_1,Z_3) \,|\, X_1, X_2, X_3 \right]$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} \min(f_i^{1/2}, f_j^{1/2})\, f_i^{1/2}\, f_j^{1/2}\, c_i(X_1,X_2)\, c_j(X_1,X_3)$$

$$\geq \sum_{j=1}^{k} \sum_{j=1}^{k} f_i f_j\, c_i(X_1,X_2)\, c_j(X_1,X_3)$$

$$= f(X_1,X_2)\, f(X_1,X_3) .$$

Thus

$$(14) \qquad E\, f^*(Z_1,Z_2)\, f^*(Z_1,Z_3) \geq E\, f(X_1,X_2)\, f(X_1,X_3) .$$

Since, by the first part of the proof, inequality (1) is true for f^*, it follows from (13) and (14) that it is also true for f. The proof is complete.

References

[1] M. E. Daniels and M. G. Kendall, "The significance of rank of correlations where parental correlation exists." Biometrika, Vol. 34 (1947), pp. 197-208.

[2] W. Hoeffding, "A class of statistics with asymptotically normal distribution." Ann. Math. Stat., Vol. 19 (1948), pp. 293-325.

[3] M. G. Kendall, Rank Correlation Methods. 2d ed. London (Griffin) and New York (Hafner), 1955.

[4] R. M. Sundrum, "Moments of the rank correlation coefficient τ in the general case." Biometrika, Vol. 40 (1953), pp. 409-420.