Ann. Inst. Statist. Math.
Vol. 44, No. 1, 169–175 (1992)

INEQUALITIES CONCERNING THE EXPECTED SELECTION DIFFERENTIALS

BRIAN J. ENGLISH[1], RAPHAEL GILLETT[2] AND MICHAEL J. PHILLIPS[1]

[1] *Department of Mathematics, University of Leicester, University Road, Leicester, U.K.*
[2] *Department of Psychology, University of Leicester, University Road, Leicester, U.K.*

(Received January 12, 1990; revised October 8, 1990)

Abstract. In many situations two populations are compared on the basis of subsets of the available data. If this is done using the same fraction of "best" records, then the expectations of the arithmetic means of these fractions are strictly ordered in magnitude by the ordering of the sample sizes. The results are illustrated with the special cases of the uniform and negative exponential distributions, for which further inequalities are derived.

Key words and phrases: Order statistics, means of extreme subsets, inequalities.

1. Introduction

In many situations when populations are compared, this is achieved on the basis of subsets of the available data. Thus in animal breeding studies, one common procedure for evaluating the genetic value of sires, is to compared known fractions of the total progeny records reported, from sub-samples comprising just the best progeny records (Scheaffer *et al.* (1970)). Other investigations, for example the UGC evaluation of research performance in British universities, have attempted to compare populations on the basis of a constant number of "best" records from each population. If these "best" records are compared using some simple index, for example their arithmetic mean or median, then clearly a bias exists in favour of the population providing the largest sample (see Gillet (1989)). However, it is not obvious that a bias in favour of the largest sample, persists when the sub-sample of best records is a constant fraction of the full sample.

Burrows (1972, 1975) considered the standardized expectation of the selection differential, $k_{(\alpha)}$. This is the expectation of the arithmetic mean of the fraction α of the "best" records. This expectation was evaluated numerically for the normal and exponential cases and an algebraic formula was given in the uniform case. Numerical results had been given earlier by Becker (1968) using the results of Harter (1961). Burrows (1975) noted that $k_{(\alpha)}$ increases as the sample size increases for the normal case, though this was based only on numerical evidence. Nagaraja

(1981, 1982, 1984) has produced some finite sample as well as asymptotic results for the selection differential.

The main purpose of this note is to establish a conjecture, due to Gillett, that $k_{(\alpha)}$ increases as the sample size increases, irrespective of the underlying (assumed common) distribution of the sampled populations.

2. Notation and main results

To establish our main result we use the following lemma.

LEMMA 2.1. *For real numbers $x_1, x_2, \ldots, x_n$, necessary and sufficient conditions for $\sum_{i=1}^n a_i x_i \geq 0$ to hold for all nondecreasing sequences of real numbers $a_1, a_2, \ldots, a_n$, are*

$$\sum_{i=1}^{r} x_i \leq 0 \quad (r = 1, 2, \ldots, n-1) \quad \textit{and} \quad \sum_{i=1}^{n} x_i = 0. \tag{2.1}$$

PROOF. Sufficiency follows immediately from Abel's partial summation formula

$$\sum_{i=1}^{n} a_i x_i = \sum_{r=1}^{n-1} \left((a_r - a_{r+1}) \sum_{i=1}^{r} x_i \right) + a_n \sum_{i=1}^{n} x_i.$$

Necessity for $\sum_{i=1}^n x_i = 0$ follows if we select $a_i = 1$ or $a_i = -1$ $(i = 1, 2, \ldots, n)$ and of $\sum_{i=1}^r x_i \leq 0$ $(r = 1, 2, \ldots, n-1)$ if we select $a_i = -1$ $(i = 1, 2, \ldots, r)$ with $a_i = 0$ $(i = r+1, \ldots, n)$. Further, if $\{a_i\}$ is a strictly increasing sequence, then the strict inequality $\sum_{i=1}^n a_i x_i > 0$ holds, provided at least one of the inequalities of (2.1) is strict. □

The above lemma is contained as a particular case of inequalities due to Popoviciu (see Mitrinović (1970), p. 38) and Minkowski (Beckenbach and Bellman (1961), p. 119).

Let $\mu_{i,n}$ denote the expected value of the i-th order statistic from a sample size of n independent and identically distributed observations. Let $S_{k,n}$ and $A_{k,n}$ denote respectively the sum and the arithmetic mean of the expected values of the k smallest order statistics from this sample.

Expect in the degenerate case, when the sampled population consists of a single point, the inequalities obtained are strict. In the subsequent discussions we assume, without further comment, that the population is non-degenerate.

THEOREM 2.1. *If k_1, k_2, n_1, n_2 are integers satisfying $n_1 > n_2$ and $k_1/n_1 \leq k_2/n_2 < 1$, then $A_{k_1,n_1} < A_{k_2,n_2}$.*

PROOF. Note, if $k_1 \leq k_2$ (and $n_1 > n_2$) then $k_1/n_1 < k_2/n_2$, and the theorem follows immediately from the simple inequalities $A_{k,n+1} < A_{k,n} < A_{k+1,n}$. Hence, we only need consider the case $k_1 > k_2$.

For an arbitrary distribution (see for example David (1981)), the expectations of order statistics for samples of size of n and $n+1$ satisfy the recurrence relationship

$$\mu_{i,n} = \{(n-i+1)\mu_{i,n+1} + i\mu_{i+1,n+1}\}/(n+1). \tag{2.2}$$

Summing (2.2) over $i = 1$ to k gives

$$S_{k,n} = \{nS_{k,n+1} + k\mu_{k+1,n+1}\}/(n+1). \tag{2.3}$$

Repeated application of (2.3) yields

$$S_{k,n} = \frac{n}{n+r} S_{k,n+r} + kn \sum_{j=1}^{r} \frac{\mu_{k+1,n+j}}{(n+j)(n+j-1)}. \tag{2.4}$$

Hence

$$\begin{aligned} \Delta &\overset{\text{def}}{=} A_{k,n} - A_{k+t,n+r} \\ &= \frac{(nt-kr)}{k(n+r)(k+t)} S_{k,n+r} \\ &\quad + n \sum_{j=1}^{r} \frac{\mu_{k+1,n+j}}{(n+j)(n+j-1)} - \frac{1}{(k+t)} \sum_{j=1}^{t} \mu_{k+j,n+r}. \end{aligned} \tag{2.5}$$

To establish the theorem we employ the following identity of Sillitto (1964).

$$\binom{n+r}{r} \mu_{i,n} = \sum_{j=0}^{r} \binom{i+j-1}{j} \binom{n+r-i-j}{r-j} \mu_{i+j,n+r}. \tag{2.6}$$

Using (2.6) with (2.5) provides an expression for Δ in terms of the expectations of order statistics from samples of common size $(n+r)$. Thus,

$$\Delta = \frac{(nt-kr)}{k(n+r)(k+t)} S_{k,n+r} + \sum_{j=1}^{r} w_j \mu_{k+j,n+r} - \frac{1}{(k+t)} \sum_{j=1}^{t} \mu_{k+j,n+r}, \tag{2.7}$$

where

$$w_j = \sum_{i=1}^{r-j+1} \frac{n}{k(n+r)} \binom{r-i}{j-1} \binom{n+i-2}{k-1} \Big/ \binom{n+r-1}{k+j-1}. \tag{2.8}$$

If k_1, k_2, n_1, n_2 satisfy the conditions of Theorem 2.1 with $k_1 > k_2$, then there are positive integers n, k, r and t such that $n_2 = n$, $k_2 = k$, $n_1 = n+r$, $k_1 = k+t$ and satisfying $nt - kr \leq 0$, $t < r$. Thus from (2.7) Δ has the form $\sum_{j=1}^{k+r} x_j \mu_{j,n+r}$ where $\{\mu_{j,n+r}; 1 \leq j \leq k+r\}$ is an increasing sequence and $\sum_{j=1}^{k+r} x_j = 0$. It follows from Lemma 2.1, that Theorem 2.1 is established if $\sum_{j=1}^{i} x_j \leq 0$ for $1 \leq i < k+r$.

For $1 \leq i \leq k$, these inequalities are a trivial consequence of $nt - kr \leq 0$, while those for $k+t < i < k+r$, follow immediately from $\sum_{j=1}^{k+r} x_j = 0$ and $w_j > 0$. The remaining inequalities are established if we can show that $\sum_{j=1}^{i}(w_j - 1/(k+t)) < 0$, which follows using (2.8), since with $i^* = r - j - i + 1$ we have

$$\begin{aligned} w_j &= \frac{n}{k(n+r)} \sum_{i^*=0}^{r-j} \binom{i^*+j-1}{j-1}\binom{n+r-j-i^*-1}{k-1} \Big/ \binom{n+r-1}{k+j-1} \\ &< \frac{n}{k(n+r)} \sum_{i^*=0}^{n-k+r-j} \binom{i^*+j-1}{j-1}\binom{n+r-j-i^*-1}{k-1} \Big/ \binom{n+r-1}{k+j-1} \\ &= \frac{n}{k(n+r)} \leq \frac{1}{k+t}. \end{aligned}$$ □

If $A^*_{k,n}$ denotes the arithmetic mean of the expectation of the k largest order statistics, then by changing the sign of each observation and applying Theorem 2.1, we obtain immediately the following corollary.

COROLLARY 2.1. *If k_1, k_2, n_1, n_2 are integers satisfying $n_1 > n_2$ and $k_1/n_1 \leq k_2/n_2 < 1$, then $A^*_{k_1,n_1} > A^*_{k_2,n_2}$.*

The basic recurrence relationship (2.2) is valid for any moment of the order statistics (if they exist), and holds also for their distribution functions. Further, the assumption of independent random variables may be relaxed to exchangeability (see David (1981), p. 104). Thus the theorem also established formal inequalities for these quantities.

COROLLARY 2.2. *If a, b and n are integers, such that $0 < a < b$, then $A_{an,bn}$ is a decreasing function of n and $A^*_{an,bn}$ is an increasing function of n.*

3. Special cases

We investigate the implications of the theorem for two particular distributions for which tractable results are available; the uniform and negative exponential distributions. We demonstrate that for these distributions $A^*_{an,bn}$, in addition to being bounded above by $A^*_{an+a,bn+b}$, is bounded below by $A^*_{an+a+1,bn+b}$. The corresponding inequality is established for $A_{an,bn}$. Thus for such samples, the direction of bias is reversed by the inclusion of a single observation in the larger sample. We present these results in the following theorem.

THEOREM 3.1. *If a, b and n are integers such that $0 < a < b$, then for the uniform and negative exponential distributions we have*

$$A_{an+a,bn+b} < A_{an,bn} < A_{an+a+1,bn+b}, \tag{3.1}$$

$$A^*_{an+a,bn+b} > A^*_{an,bn} > A^*_{an+a+1,bn+b}. \tag{3.2}$$

PROOF. The first inequality in each of (3.1) and (3.2) is of course a special case of Corollary 2.1.

(i) *Uniform distribution.* If the sample consists of observations of a random variable X having a uniform distribution on $[0, 1]$, then $\mu_{i,n} = i/(n+1)$ and $A_{an,bn} = (an+1)/(2bn+2)$. In this case (3.1) and (3.2) are easily verified directly.

(ii) *Negative exponential distribution.* If the sample consists of observations of a random variable X having a negative exponential distribution with unit mean, then it may be verified (see Feller (1971), p. 20) that $\mu_{k,n} = \sum_{i=n-k+1}^{n} 1/i$, from which

$$A^*_{k,n} = 1 + \sum_{i=k+1}^{n} \frac{1}{i} \quad \text{and} \quad A^*_{an,bn} = 1 + \sum_{i=an+1}^{bn} \frac{1}{i}. \tag{3.3}$$

Summations of the form $\sum_{i=an+1}^{bn} 1/i$ (for a, b and n satisfying the conditions of the theorem) have been investigated by Adamović and Tasković (1969) (see Mitrinović (1970)), who prove directly that such sums are increasing functions of n.

Also

$$A^*_{an,bn} - A^*_{an+a+1,bn+b} = \sum_{i=1}^{a+1} \frac{1}{an+i} - \sum_{i=1}^{b} \frac{1}{bn+i}. \tag{3.4}$$

The second summation of (3.4) is an increasing function of b, thus to establish the second inequality of (3.2) it sufficient to prove that

$$\sum_{i=1}^{a+1} \frac{1}{an+i} > \lim_{b\to\infty} \sum_{i=1}^{b} \frac{1}{bn+i} = \ln\left(1 + \frac{1}{n}\right),$$

which may be justified as follows;

$$\sum_{i=1}^{a+1} \frac{1}{an+i} > \int_0^{a+1} \frac{1}{an+x+1}dx = \ln\left(\frac{an+a+2}{an+1}\right) \geq \ln\left(1 + \frac{1}{n}\right).$$

Since the negative exponential distribution is asymmetric, the corresponding inquality for $A_{an,bn}$ does not follow immediately. However, from the identity

$$A_{k,n} = \{n\mu - (n-k)A^*_{n-k,n}\}/k, \tag{3.5}$$

where $\mu = E(X)$, the inequality of (3.2) does imply the corresponding inequality of (3.1). Using (3.5) with (3.3), we have

$$\begin{aligned} &A_{an+a+1,bn+b} - A_{an,bn} \\ &= \frac{c}{a}\sum_{i=1}^{an} \frac{1}{cn+i} - \frac{cn+c-1}{an+a+1} \sum_{i=1}^{an+a+1} \frac{1}{cn+c+i-1}. \end{aligned} \tag{3.6}$$

Now consider the function $\phi(k)$ defined by

$$\phi(k) = \frac{cn+kr}{an+k} \sum_{i=1}^{an+k} \frac{1}{cn+kr+i} \quad \text{for} \quad k = 0, 1, \ldots, a+1,$$

where $r = (c-1)/(a+1)$. To demonstrate that the right-hand side of (3.6) is positive, it suffices to show $\phi(k)$ decreases for $0 \leq k \leq a+1$.

This follows as

$$\begin{aligned}
&\phi(k-1) - \phi(k) \\
&= \frac{1}{an+k-1} \\
&\quad \cdot \sum_{i=1}^{an+k-1} \left\{ \frac{cn+kr-r}{cn+kr-r+i} - \frac{cn+kr}{an+k} \left(\frac{an+k-i}{cn+kr+i} + \frac{i}{cn+kr+i+1} \right) \right\} \\
&= \frac{r+1}{(an+k)(an+k-1)} \\
&\quad \cdot \sum_{i=1}^{an+k-1} \frac{i\{(n-1)(cn+kr)+n(i+1)\}}{(cn+kr-r+i)(cn+kr+i)(cn+kr+i+1)} > 0. \qquad \square
\end{aligned}$$

In view of (3.1) it is natural to enquire whether the deletion of a single observation from the smaller sample reverses the bias established in Corollary 2.2. Further, we note that for general n_1, n_2 $(n_1 > n_2)$ we cannot conclude from (3.1) that $A_{an_2,bn_2} < A_{an_1+1,bn_1}$, except when $n_1 = n_2 + 1$. This prompts us to seek bounds for the smallest value of k for which the inequality $A_{an_2,bn_2} < A_{an_1+k,bn_1}$ holds. These questions are addressed in the following corollary.

COROLLARY 3.1. *If the inequalities of* (3.1) *hold for some distribution then*

(i) *for all integers a, b and n such that* $a < b$ *we have* $A_{an-1,bn} < A_{an+a,bn+b}$*;*

(ii) *for integers* k_1, k_2, n_1, n_2 *such that* $k_1/n_1 = k_2/n_2$ *and* $n_1 > n_2$ *we have* $A_{k_2,n_2} < A_{k_1+m+1,n_1}$*; where m denotes the integer part of* (n_1/n_2).

PROOF. (i) From the second inequality of (3.1) $A_{an-1,bn} < A_{2an-1,2bn}$; and noting that $2bn \geq b(n+1)$ and $(2an-1)/(2bn) < (an+a)/(bn+b)$, Theorem 2.1 implies $A_{2an-1,2bn} < A_{an+a,bn+b}$.

(ii) Repeated application of the second inequality of (3.1) yields $A_{k_2,n_2} < A_{2k_2+1,2n_2} < A_{4k_2+3,4n_2} < \cdots < A_{2^t(k_2+1)-1,2^t n_2}$ and thus $A_{k_2,n_2} < A_{2^t(k_2+1),2^t n_2}$ for all positive integers t. Hence, if t' is such that $t' > m+1$ we have $A_{k_2,n_2} < A_{2^{t'}(k_2+1),2^{t'}n_2} < A_{(m+1)(k_2+1),(m+1)n_2} < A_{k_1+m+1,n_1}$. The second inequality follows from Corollary 2.2, while the third follows from Thorem 2.1 on noting that $(m+1)n_2 > n_1$ and $(k_2+1)/n_2 < (k_1+m+1)/n_1$. □

Acknowledgements

The authors are indebted to Professor R. O. Davies for some invaluable discussions of this problem, and to a referee for his comments.

References

Adamović, D. D. and Tasković, M. R. (1969). Monotony and the best possible bounds of some sequences of sums, *Publikacije Elektrotehničkog Fakulteta Univerziteta u Beogradu, Serija: Matematika i Fizika*, No. 247-273, 41–50.

Beckenbach, E. F. and Bellman, R. (1961). *Inequalities*, Springer, Berlin.

Becker, W. A. (1968). *Manual of Procedures in Quantitative Genetics*, Washington State University Press, Pullman, Washington.

Burrows, P. M. (1972). Expected selection differentials for directional selection, *Biometrics*, **28**, 1091–1100.

Burrows, P. M. (1975). Variances of selection differentials in normal samples, *Biometrics*, **31**, 125–133.

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

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed., Wiley, New York.

Gillett, R. (1989). A sampling artifact in the UGC evaluation of research performance, *Bri. J. Math. Statist. Psych.*, **42**, 127–132.

Harter, H. L. (1961). Expected values of normal order statistics, *Biometrika*, **48**, 151–159.

Mitrinović, D. S. (1970). *Analytic Inequalities*, Springer, Berlin.

Nagaraja, H. N. (1981). Some finite sample results for the selection differential, *Ann. Inst. Statist. Math.*, **33**, 437–448.

Nagaraja, H. N. (1982). Some nondegenerate limit laws for the selection differential, *Ann. Statist.*, **10**, 1306–1310.

Nagaraja, H. N. (1984). Some nondegenerate limit laws for sample selection differential and selection differential, *Sankhyā Ser. A*, **46**, 355–369.

Schaeffer, L. R., Van Vleck, L. D. and Velasco, J. A. (1970). The use of order statistics with selected records, *Biometrics*, **26**, 854–859.

Sillitto, G. P. (1964). Some relations between expectations of order statistics in samples of different sizes, *Biometrika*, **51**, 259–262.