A Note on the Construction of a Sample Variance

Nitis Mukhopadhyay[*1] **and Bhargab Chattopadhyay**[2]

[*1]Department of Statistics, University of Connecticut-Storrs, USA

[2]Department of Mathematical Sciences, University of Texas at Dallas, USA

Corresponding Author : nitis.mukhopadhyay@uconn.edu

Received: July 29,2013 / Revised: November 22, 2013 / Accepted: December 12, 2013

ABSTRACT

Mukhopadhyay and Chattopadhyay (2013) proposed a new approach to construct unbiased estimators of σ^2 by U-statistics starting with a class of non-symmetric initial kernels of degree $m(> 2)$. But, surprisingly all symmetrized final estimators in the form of U-statistics reduced to S^2 when the parent distribution function (d.f.) F remained unknown. Now, we exhibit another class of non-symmetric initial kernels to come up with U-statistics of degree $m(> 2)$ and unbiased for σ^2. Interestingly, the associated symmetrized final U-statistics again coincide with S^2 (Theorem 2.1), settling the open question from Remark 3.1 of Mukhopadhyay and Chattopadhyay (2013).

Keywords: Kernels, Sample variance, U-statistics.

1. Introduction

A popular distribution-free measure of variation is the *sample variance.* Suppose that $X_1, ..., X_n$ are independent and identically distributed random variables observed from a population *distribution function* (d.f.) F with $\mu \equiv \mu(F) = \int_{\Re} x dF(x)$ and $\sigma^2 \equiv \sigma^2(F) = \int_{\Re}(x-\mu)^2 dF(x)$. We assume that σ^2 is positive and finite, but let F be otherwise completely unknown. We denote the sample mean, $\overline{X} = n^{-1}\Sigma_{i=1}^n X_i$. The sample variance,

$$S^2 \equiv (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})^2 \text{ for } n \geq 2, \tag{1.1}$$

is an unbiased estimator of σ^2.

In a related context of measuring inequality in an income distribution F, the sample variance and other competing measures of variation appear often in a variety of fields interfacing with statistics, economics, actuarial mathematics, and financial mathematics, among others. For brevity, however, we only refer to Sen (1997) which is regarded as a classic on economic inequality and inequality indices originally due to Gini (1914,1921). Clearly, (1.1) shows that S^2 takes into account comparisons of each observed value relative to the sample average. Sen (1997) has often criticized S^2 emphatically by drawing attention to the expression (1.1). We quote from Sen (1997, p. 28): "Is it best to measure the difference of each income level from the mean only, or should the comparison be carried out between every pair of incomes? The latter will capture everyone's income difference from everyone else, and not merely from the mean, which might not be anybody's income whatsoever." Here is another quote from Sen (1997, p. 31): "In taking differences over all pairs of incomes, the Gini coefficient or the absolute mean difference avoids the total concentration on differences vis-à-vis the mean ... Undoubtedly one appeal of the Gini coefficient, or of the relative mean difference, lies in the fact that it is a very direct measure of income difference, taking note of differences between every pair of incomes." However, we may emphasize that S^2 has the following well-known equivalent representation:

$$S^2 = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} \frac{1}{2}(X_{i_1} - X_{i_2})^2. \tag{1.2}$$

It is a U-statistic introduced by Hoeffding (1948) that is associated with a symmetric kernel $g(x_{i_1}, x_{i_2}) = \frac{1}{2}(x_{i_1} - x_{i_2})^2$ of degree 2 and it is clear from (1.2) that in coming up with S^2, one compares each individual's income with everyone else's income. The cited quotes from Sen (1997) are off the mark in the sense that such criticisms against S^2 do not hold up.

1.1. Mukhopadhyay and Chattopadhyay's (2013) Original Construction

In a recent paper, Mukhopadhyay and Chattopadhyay (2013) elegantly utilized non-symmetric initial kernels of order $m > 2$ to unbiasedly estimate σ^2 when F remained completely unknown. Such a general method, however, led them back to S^2. In that regard, for completeness, we take the liberty to restate Theorem 3.1 from Mukhopadhyay and Chattopadhyay (2013) which gave rise to a new interpretation of S^2 that went much further beyond (1.2).

Mukhopadhyay and Chattopadhyay (2013) generalized the core idea from (1.2) as follows. From the set of random variables $X_1, ..., X_n$, they first considered a subset of random variables $X_{i_1}, ..., X_{i_m}$ of fixed size $m(< n)$ where the (distinct) set of indices $i_1, ..., i_m$ were held fixed for the time-being. They began comparing two arbitrary linear functions:

$$\left(\sum_{p=1}^{k} a_p\right)^{-1} \sum_{p=1}^{k} a_p X_{i_p} \text{ and } \left(\sum_{q=k+1}^{k+l} b_q\right)^{-1} \sum_{q=k+1}^{k+l} b_q X_{i_q}.$$

These were formed respectively from $X_{i_1}, ..., X_{i_k}$ and $X_{i_{k+1}}, ..., X_{i_{k+l}}$ where k, l were fixed positive integers, $k + l = m$. Mukhopadhyay and Chattopadhyay (2013) assumed that the a's, b's were non-zero real numbers and $\sum_{p=1}^{k} a_p \neq 0$, $\sum_{q=k+1}^{k+l} b_q \neq 0$. Thus, they began with the following statistic:

$$\begin{aligned} &h(X_{i_1}, ..., X_{i_m}) \\ &= \left\{\left(\sum_{p=1}^{k} a_p\right)^{-1} \sum_{p=1}^{k} a_p X_{i_p} - \left(\sum_{q=k+1}^{k+l} b_q\right)^{-1} \sum_{q=k+1}^{k+l} b_q X_{i_q}\right\}^2 \\ &= \left(\sum_{p=1}^{k} c_p X_{i_p} + \sum_{q=k+1}^{k+l} d_q X_{i_q}\right)^2, \end{aligned} \tag{1.3}$$

where $c_p = a_p \left(\sum_{i=1}^{k} a_i\right)^{-1}, d_q = -b_q \left(\sum_{j=k+1}^{k+l} b_j\right)^{-1}$.

Then, Mukhopadhyay and Chattopadhyay (2013) symmetrized the function $h(.)$ from (1.3) and made it unbiased for σ^2. This way, they came up with the following kernel of degree m:

$$\begin{aligned} &g^{(m)}(X_{i_1}, ..., X_{i_m}) \\ &= \frac{1}{m!}\left(\sum_{p=1}^{k} c_p^2 + \sum_{q=k+1}^{k+l} d_q^2\right)^{-1} \sum_{j_1 \neq ... \neq j_m} h(X_{j_1}, ..., X_{j_m}), \end{aligned} \tag{1.4}$$

where $h(.)$ came from (1.3) corresponding to the functional, $\theta \equiv \sigma^2(F)$. This amounts to averaging $h(.)$ over all permutations $j_1 \neq ... \neq j_m$ obtained from the fixed set of distinct indices $i_1, ..., i_m$. Thus far, the distinct indices $i_1, ..., i_m$ were fixed.

Theorem 1.1 (Mukhopadhyay and Chattopadhyay, 2013). *The U-statistic* $U^{(m)} \equiv \binom{n}{m}^{-1} \sum_{n,m} g^{(m)}(X_{i_1}, ..., X_{i_m})$ *with the symmetric kernel* $g^{(m)}(X_{i_1}, ..., X_{i_m})$ *of degree* m

from (1.4) reduces to the sample variance S^2 whatever may be the d.f. F associated with a population having a finite and positive variance $\sigma^2 \equiv \sigma^2(F)$. The result holds whatever be the non-zero a's and b's such that $\sum\limits_{p=1}^{k} a_p \neq 0$, $\sum\limits_{q=k+1}^{k+l} b_q \neq 0, m = k + l$ with arbitrary but fixed $k, l, 1 \leq k, l < n$ and $k + l < n$.

This striking result gave a new interpretation of S^2. The sample variance indeed compares the average income of any subgroup of individuals (of size k) with any other subgroup of individual's (of size l) average income, k and l are fixed but arbitrary, $k + l = m < n$. For a more comprehensive list of references, one may review Mukhopadhyay and Chattopadhyay (2013).

1.2. The Motivation for the Present Work

An important question came up: Is the method of construction described by Mukhopadhyay and Chattopadhyay (2013), briefly emphasized here in Theorem 1.1, the *only* possible way to finally arrive at symmetric kernels of degree $m(> 2)$ leading up to unbiased U-statistics for σ^2 when F remains arbitrary and unknown with σ^2 finite and positive? It remained an open question as stated in their Remark 3.1.

In this short paper, we show examples of interesting non-symmetric initial kernels from outside the class of non-symmetric initial kernels that was exploited by Mukhopadhyay and Chattopadhyay (2013). It is also interesting that these new kernels provide final unbiased U-statistics for σ^2 with degree $m(> 2)$ which again eventually reduce (Theorem 2.1) to the sample variance, S^2.

2. A New Class of Initial Kernels

Here, we first exhibit some new non-symmetric initial kernels (Section 2.1) outside the class of symmetric initial kernels that was exploited previously in Mukhopadhyay and Chattopadhyay (2013). The population d.f. F is left unknown other than the fact that $\sigma^2 \equiv \sigma^2(F)$ is assumed positive and finite. A general construction is given in Section 2.2 and our main finding is then briefly stated and proved as Theorem 2.1.

2.1. Illustrations and Data Analyses

Let us begin with the following basic kernels:

$$h_1(X_{i_1}, X_{i_2}, X_{i_3}) = 2X_{i_1}^2 - \frac{1}{2}(X_{i_2} + X_{i_3})^2 : m = 3; \tag{2.1}$$

$$h_2(X_{i_1},...,X_{i_4}) = \frac{3}{2}X_{i_1}^2 - \frac{1}{6}(X_{i_2} + X_{i_3} + X_{i_4})^2 : m = 4; \text{ and} \tag{2.2}$$

$$h_3(X_{i_1},...,X_{i_4}) = \frac{3}{2}(X_{i_3} + X_{i_4})^2 - \frac{2}{3}(X_{i_1} + X_{i_2} + X_{i_3})^2 : m = 4. \tag{2.3}$$

Each is an initial unbiased estimator for σ^2 with respective degrees $3, 4$ and 4. Clearly, these are outside the class of initial symmetric kernels considered in (3.1) of Mukhopadhyay and Chattopadhyay (2013) for the following simple reason: The kernels $h_1(.), h_2(.)$ and $h_3(.)$ from (2.1)-(2.3) can be negative with positive probabilities for many F's with σ^2 positive and finite. However, the initial kernels from (3.1) of Mukhopadhyay and Chattopadhyay (2013) were all non-negative w.p.1.

Next, suppose that we symmetrize each $h_i(.)$ from (2.1)-(2.3) and then average over subsets of associated size m formed from $X_1,...,X_n$ to come up with a final U-statistic denoted by $U_i^{(m)}$, $i = 1, 2, 3$. Now, consider the following data-based illustrations.

Example 2.1. We generated pseudo random data of size $n = 5$ from a $N(8, 81)$ population. The observed data were:

$$11.1309816, 1.6341749, 15.0287913, 8.6507071, 0.6121078.$$

Then, we evaluated $U_1^{(3)}, U_2^{(4)}, U_3^{(4)}$ and S^2 for this dataset and noted that their observed values coincided with 38.25063. ▲

Example 2.2. We generated pseudo random data of size $n = 10$ from an exponential distribution with population mean 10. The data were:

$$0.0349129, 0.1977668, 0.0631878, 0.0403146, 0.1707803,$$
$$0.0062840, 0.4596507, 0.0454418, 0.0146797, 0.2260883.$$

Again, we evaluated $U_1^{(3)}, U_2^{(4)}, U_3^{(4)}$ and S^2 for this dataset and noted that their observed values coincided with 0.02014379. ▲

Example 2.3. We generated pseudo random data of size $n = 15$ from an exponential distribution with population mean 10. The data were:

$$0.0349129, 0.1977668, 0.0631878, 0.0403146, 0.1707803,$$
$$0.0062840, 0.4596507, 0.0454418, 0.0146797, 0.2260883,$$
$$11.1309816, 1.6341749, 15.0287913, 8.6507071, 0.6121078.$$

When we evaluated $U_1^{(3)}, U_2^{(4)}, U_3^{(4)}$ and S^2 for this dataset, we noted that their observed values coincided with 23.57924. ▲

In other words, for each dataset, we eventually returned to S^2. Indeed, we can come up with a general construction and show that the associated resulting U-statistics would reduce to S^2.

2.2. A General Construction and a Theorem

First, we think of a fixed subset of random variables $X_{i_1}, ..., X_{i_m}$ of size $m(< n)$. Then, we pretend comparing two arbitrary functions

$$l\left(k^{-1/2}\sum_{p=1}^{k} X_{i_p}\right)^2 \text{ and } k\left(l^{-1/2}\sum_{q=1}^{l} X_{j_q}\right)^2, \tag{2.4}$$

formed respectively from subgroups $X_{i_1}, ..., X_{i_k}$ and $X_{j_1}, ..., X_{j_l}$ along the lines of (2.1)-(2.3). Here k, l are fixed positive integers such that the two subgroups combined have exactly m distinct observations in common, that is, $k + l \geq m$.

In the context of $h_1(.), h_2(.)$, and $h_3(.)$,we had (i) $k = 1, l = 2, m = 3$, (ii) $k = 1, l = 3, m = 4$, and (iii) $k = 2, l = 3, m = 4$ respectively. Clearly, we allow some common indices between the two subgroups $X_{i_1}, ..., X_{i_k}$ and $X_{j_1}, ..., X_{j_l}$ such that the two subgroups combined have exactly m distinct indices.

With such general k, l and m, we consider the basic initial non-symmetric kernel:

$$h(X_{i_1}, ..., X_{i_m}) = \frac{1}{(l-k)}\left[l\left(k^{-1/2}\sum_{p=1}^{k} X_{i_p}\right)^2 - k\left(l^{-1/2}\sum_{q=1}^{l} X_{i_q}\right)^2\right], \tag{2.5}$$

for fixed but arbitrary but fixed $k, l, k + l \geq m$. This is unbiased for σ^2 to begin with.

Next, we symmetrize the function $h(X_{i_1}, ..., X_{i_m})$ from (2.5) and make it unbiased for σ^2. This way, we would come up with the following symmetric kernel of degree m which would estimate σ^2 unbiasedly:

$$\begin{aligned} g^{(m)}(X_{i_1}, ..., X_{i_m}) &= \frac{1}{m!}\sum_{1\leq j_1\neq ...\neq j_m\leq m} h(X_{j_1}, ..., X_{j_m}). \\ &= \frac{1}{m!}\left[\frac{l}{k\,(l-k)}I - \frac{k}{l\,(l-k)}II\right], \end{aligned} \tag{2.6}$$

where we denote:

$$\begin{aligned} I &= \sum_{1\le j_1\neq \ldots \neq j_m \le m} \left(\sum_{p=1}^{k} X_{j_p} \right)^2 \\ &= k(m-1)! \sum_{j=1}^{m} X_{i_j}^2 + 2k(m-2)! \sum_{1\le j<k\le m} X_{i_j} X_{i_k}, \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} II &= \sum_{1\le j_1\neq \ldots \neq j_m \le m} \left(\sum_{q=1}^{l} X_{j_q} \right)^2 \\ &= l(m-1)! \sum_{j=1}^{m} X_{i_j}^2 + 2l(m-2)! \sum_{1\le j<k\le m} X_{i_j} X_{i_k}. \end{aligned} \tag{2.8}$$

Thus, combining (2.6)-(2.8), we can express $g^{(m)}(.)$ as follows:

$$\begin{aligned} &g^{(m)}(X_{i_1}, ..., X_{i_m}) \\ &= \frac{1}{m!} \left[\frac{l}{k\,(l-k)} \left\{ k(m-1)! \left(\sum_{j=1}^{m} X_{i_j}^2 \right) + 2k(m-2)! \sum_{1\le j<k\le m} X_{i_j} X_{i_k} \right\} \right. \\ &\left. - \frac{k}{l\,(l-k)} \left\{ l(m-1)! \left(\sum_{j=1}^{m} X_{i_j}^2 \right) + 2l(m-2)! \sum_{1\le j<k\le m} X_{i_j} X_{i_k} \right\} \right] \\ &= \frac{1}{m!} \left[(m-1)! \sum_{j=1}^{m} X_{i_j}^2 - 2(m-2)! \sum_{1\le j<k\le m} X_{i_j} X_{i_k} \right] \\ &= \frac{1}{m!} \sum_{j=1}^{m} X_{i_j}^2 - \frac{2}{m(m-1)} \sum_{1\le j<k\le m} X_{i_j} X_{i_k}. \end{aligned} \tag{2.9}$$

We let $\sum_{n,m}$ denote the summation over all indices $i_1, i_2, ..., i_m$ such that $1 \le i_1 < i_2 < ... < i_m \le n$. Next, we state and prove our main result here.

Theorem 2.1. *A U-statistic $U^{(m)} \equiv \binom{n}{m}^{-1} \sum_{n,m} g^{(m)}(X_{i_1}, ..., X_{i_m})$ with the symmetrized kernel $g^{(m)}(X_{i_1}, ..., X_{i_m})$ of degree m from (2.9) reduces to the sample variance S^2 whatever may be the d.f. F with finite and positive variance $\sigma^2 \equiv \sigma^2(F)$. The result holds for arbitrary but fixed integers $k, l, 1 \le k, l < n, m \le k + l \le n$ and the two combined subgroups of sizes k, l have exactly m* distinct *indices as in (2.5).*

Proof: We rely upon (2.9) to express the U-statistic $U^{(m)}$ as follows:

$$\begin{aligned}
&\binom{n}{m}^{-1} \sum_{n,m} g^{(m)}(X_{i_1}, ..., X_{i_m}) \\
&= \binom{n}{m}^{-1} \sum_{n,m} \frac{1}{m!} \left\{ (m-1)! \sum_{j=1}^{m} X_{i_j}^2 - 2(m-2)! \left(\sum_{1 \le j < k \le m} X_{i_j} X_{i_k} \right) \right\} \qquad (2.10) \\
&= \binom{n}{m}^{-1} \sum_{n,m} \left\{ \frac{1}{m} \sum_{j=1}^{m} X_{i_j}^2 - \frac{2}{m(m-1)} \sum_{1 \le j < k \le m} X_{i_j} X_{i_k} \right\}.
\end{aligned}$$

Then, (2.10) leads to:

$$\begin{aligned}
U^{(m)} &= \binom{n}{m}^{-1} \left\{ \frac{1}{m} \binom{n-1}{m-1} \sum_{i=1}^{n} X_i^2 - \frac{2}{m(m-1)} . \binom{n-2}{m-2} \sum_{1 \le i < j \le n} X_i X_j \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} X_i X_j.
\end{aligned} \qquad (2.11)$$

On the other hand, however, the sample variance is given by,

$$\begin{aligned}
S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \overline{X}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{2}{n(n-1)} \textstyle\sum_{1 \le i < j \le n} X_i X_j.
\end{aligned} \qquad (2.12)$$

Next, by combining (2.11)-(2.12), we complete the proof. □

Remark 2.1. One should not get an impression that initial non-symmetric kernels taking negative values with a positive probability are possible only when m exceeds two. With $m = 2$, for example, we may begin with a kernel $X_1^2 - X_1 X_2$ of degree 2. After symmetrizing and averaging, one has $\frac{1}{2}(X_1^2 - X_1 X_2 + X_2^2 - X_2 X_1)$ which obviously reduces to $\frac{1}{2}(X_1 - X_2)^2$ which leads to (1.2).

3. Further Thoughts

One may feel tempted to claim the validity of Mukhopadhyay and Chattopadhyay's (2013) original Theorem 3.1 and our present Theorem 2.1 by invoking completeness and sufficiency. But, that will not work since we have assumed an unknown

F, not necessarily admitting a probability density. This point was emphasized in Mukhopadhyay and Chattopadhyay (2013).

Next, one may come up with similar constructions for estimating a population mean $\mu \equiv \mu(F)$ unbiasedly. For brevity, we simply show a U-statistic of degree 2 that estimates μ unbiasedly. We begin with an initial unbiased kernel $(a_1 + a_2)^{-1}(a_1 X_{i_1} + a_2 X_{i_2})$ for any two distinct indices i_1, i_2 with a's fixed but non-zero real numbers, $a_1 \neq -a_2$. Then, we may symmetrize this and then average all these constructed unbiased estimators for μ from all n observations. If we denote such a U-statistic with degree 2 by $U^{(2)}$, we find that $U^{(2)}$ would eventually coincide with the sample mean, $\overline{X}$. This idea can again be pushed to construct an analogous U-statistic $U^{(m)}$ with arbitrary degree $m(> 1)$, and one can verify that such corresponding $U^{(m)}$ would again coincide with the sample mean, $\overline{X}$.

Here is a word of caution: Our kind of new interpretations for the sample variance and a sample mean may not carry over while unbiasedly estimating arbitrary functionals of F when F remains fully unknown. Mukhopadhyay and Chattopadhyay (2011) included elaborate discussions of some scenarios where the associated $U^{(m)}$ did not coincide with Gini's unbiased mean difference estimator when $\sigma \equiv \sigma(F)$ was the parameter of interest even when F corresponded to a normal population.

Acknowledgement

We appreciated the reviewers' enthusiastic feedback and we thank them for the same.

References

1. Gini, C. (1914). *L'Ammontare la Composizione della Ricchezza delle Nazioni*, Boca: Torino.

2. Gini, C. (1921). Measurement of Inequality of Incomes, *Economic Journal* 31: 124-126.

3. Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution, *Annals of Mathematical Statistics* 19: 293-325.

4. Mukhopadhyay, N. and Chattopadhyay, B. (2011). Estimating a Standard Deviation with U-statistics of Degree More Than Two: The Normal Case, *Journal of Statistics: Advances in Theory and Applications* 5:93-130.

5. Mukhopadhyay, N. and Chattopadhyay, B. (2013). On a New Interpretation of the Sample Variance, *Statistical Papers* 54: 827-837 DOI:10.1007/s00362-012-0465-y

6. Sen, Amartya (1997). *On Economic Inequality*, enlarged edition with a substantial annex by James Foster and Amartya Sen, Oxford: Clarendon Press.