S-Statistics and Their Basic Properties

Marek Gągolewski[1,2] and Przemysław Grzegorzewski[1,2]

This is a revised version of the paper:
Gągolewski M., Grzegorzewski P., S-Statistics and Their Basic Properties, In: Borgelt C. et al (Eds.), *Combining Soft Computing and Statistical Methods in Data Analysis*, Springer-Verlag, 2010, 281–288.

Abstract Some statistical properties of the so-called S-statistics, which generalize the ordered weighted maximum aggregation operators, are considered. In particular, the asymptotic normality of S-statistics is proved and some possible applications in estimation problems are suggested.
Keywords Aggregation, L-statistics, OWA, OWMax operators.

1 Introduction

The process of aggregation, i.e. combining many numerical values into a single one, plays an important role in many areas of practical human activities, such as statistics, decision making, computer science, operational research, etc. Operators projecting multidimensional state space into a single dimension are often called *aggregation functions* [5]. Among well-known examples are: the sample maximum and other quantiles, arithmetic mean, ordered weighted averaging (OWA) [11] and ordered weighted maximum (OWMax) [2] operators.

The OWA operators are a particular case of L-statistics. Their basic statistical properties were widely discussed, see e.g. [7, 10].

In this paper we consider another useful class of aggregation operators called S-statistics, which generalize OWMax. We show that S-statistics are

[1] Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland, `{gagolews,pgrzeg}@ibspan.waw.pl` .[2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Plac Politechniki 1, 00-661 Warsaw, Poland.

consistent estimators of the so-called κ-index (Sec. 3). Moreover, they are asymptotically normally distributed (Sec. 4). Regarding similar constructions it seems that S-statistics would be useful in many situations, e.g. in scientometrics to construct reliable tools for scientific activity assessment (see [3, 4, 9]), pattern matching [2] and decision making [1].

2 S-Statistics

Let $(X_1, \ldots, X_n)$ denote a sample of i.i.d. random variables, while $X_{(1)}, \ldots, X_{(n)}$ are order statistics corresponding to this sample. Assume that the c.d.f. F of X_i is continuous and strictly increasing in interval (a, b), where $a = \inf\{x : F(x) > 0\}$, $b = \sup\{x : F(x) < 1\}$.

Let $\kappa : [0, 1] \to [a, b]$ be a strictly increasing continuous function such that $\kappa(0) = a$ and $\kappa(1) = b$. Further on we will call such function a *control function*.

A linear combination of order statistics, called L-statistics, is a well-known tool applied especially in robust estimation or testing. Typical examples of L-statistics are trimmed and Winsorized means that are useful in situations when data follow a heavy-tailed distribution. Its subclass is known in decision making as the ordered weighted averaging (OWA) operator [11]. Below we propose another function of ordered statistics which has some interesting statistical properties.

Definition 1. An S-*statistic* associated with a control function κ and a random sample $(X_1, \ldots, X_n)$ is a function

$$V_{n,\kappa}(X_1, \ldots, X_n) = \bigvee_{i=1}^{n} \kappa\left(\tfrac{i}{n}\right) \wedge X_{(n-i+1)}, \tag{1}$$

where $\vee$ and $\wedge$ denote the *s*upremum (hence the name) and infimum operators, respectively.

It can be seen that the S-statistic is a generalization of the ordered weighted maximum operator (OWMax) defined firstly in [2]. Moreover, for any control function κ, the corresponding S-statistic is a function $V_{n,\kappa} : [a, b]^n \to [a, b]$ which satisfies the following requirements:

1. $V_{n,\kappa}$ is non-decreasing in each variable, i.e. $(\forall \mathbf{x}, \mathbf{y} \in [a, b]^n)\ \mathbf{x} \le \mathbf{y} \Rightarrow V_{n,\kappa}(\mathbf{x}) \le V_{n,\kappa}(\mathbf{y})$,
2. $V_{n,\kappa}$ fulfills the lower boundary condition, i.e. $\inf_{\mathbf{x}\in[a,b]^n} V_{n,\kappa}(\mathbf{x}) = a$,
3. $V_{n,\kappa}$ fulfills the upper boundary condition, i.e. $\sup_{\mathbf{x}\in[a,b]^n} V_{n,\kappa}(\mathbf{x}) = b$.

Therefore, according to the definition given e.g. in [5], $V_{n,\kappa}$ is an aggregation function. Hence, $V_{n,\kappa}$ may have (at least potentially) — like other aggregation functions — many applications in different areas. In this paper we restrict ourselves to their statistical properties related to their asymptotic distribution and estimation of a population location parameter.

Note that

$$V_{n,\kappa}(X_1,\dots,X_n) = \kappa\left(\bigvee_{i=1}^{n} \tfrac{i}{n} \wedge \kappa^{-1}\left(X_{(n-i+1)}\right)\right). \tag{2}$$

Hence, without loss of generality, we will consider S-statistics of a form

$$V_n(Y_1,\dots,Y_n) = \bigvee_{i=1}^{n} \tfrac{i}{n} \wedge Y_{(n-i+1)}, \tag{3}$$

where $(Y_1,\dots,Y_n) = (\kappa^{-1}(X_1),\dots,\kappa^{-1}(X_n))$ is a sequence of i.i.d. random variables given by the continuous c.d.f. $G := F \circ \kappa$ defined on $[0,1]$. In other words, $V_n := V_{n,\mathrm{id}}$, where id is the identity function.

3 κ-index

Consider the following definition.

Definition 2. A κ-*index* of a random variable given by a c.d.f. F with respect to the control function κ is a number $\varrho_\kappa \in [0,1]$ such that

$$\varrho_\kappa = 1 - F(\kappa(\varrho_\kappa)). \tag{4}$$

If $S(x) = 1 - F(x)$ is a survival function then, of course, a κ-index ϱ_κ satisfies

$$\varrho_\kappa = S(\kappa(\varrho_\kappa)) = \Pr(X > \kappa(\varrho_\kappa)). \tag{5}$$

Thus κ-index has an intuitive interpretation: it is such a number that the probability of assuming a value greater than $\kappa(\varrho_\kappa)$ is equal to ϱ_κ.

Example 1. If Y follows the Type-II Pareto distribution, i.e. $F(x) = 1 - 1/(1+x)$ and the control function is the identity function, i.e. $\kappa(x) = x$, then $\varrho_\kappa = (\sqrt{5}-1)/2 = 1/\varphi = \varphi - 1 \simeq 0.618034$, where φ is the *golden ratio*.

It appears that the S-statistic is a strongly consistent estimator of the id-index $\varrho := \varrho_{\mathrm{id}}$ for any c.d.f. G defined on $[0,1]$. However, to prove it we need some lemmas given below.

Lemma 1. *For any sample $Y_1,\dots,Y_n$ of i.i.d. random variables defined on $[0,1]$ with a continuous c.d.f. G we have*

$$\begin{aligned} V_n(Y_1,\dots,Y_n) &= \inf\left\{x : \hat{G}_n(x) \geq 1 - x\right\} && (6)\\ &= \sup\left\{x : \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{1}(Y_i \geq x) \geq x\right\}, && (7)\end{aligned}$$

where $\hat{G}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Y_i \le x)$ *denotes the empirical distribution function and* $\mathbf{1}$ *is the indicator function.*

Proof. Since $\sum_{i=1}^n \mathbf{1}(Y_i \ge x) = \max\{i : Y_{(n-i+1)} \ge x\}$ we get

$$\begin{aligned} V_n(Y_1,\ldots,Y_n) &= \max\{\tfrac{i}{n} : \tfrac{i}{n} \le Y_{(n-i+1)}\} \vee \max\{Y_{(n-i+1)} : Y_{(n-i+1)} \le \tfrac{i}{n}\} \\ &= \max\left\{x : \tfrac{1}{n}\max\{i : Y_{(n-i+1)} \ge x\} \ge x\right\}. \end{aligned}$$

Implication from (6) to (7) is obvious and the proof is complete. □

Recall that $(\forall x)$ we have $\hat{G}_n(x) \xrightarrow{a.s.} G(x)$ and $n\hat{G}_n(x) \sim \mathrm{Bin}(n, G(x))$.

The exact distribution of V_n is given by the next lemma.

Lemma 2. *The c.d.f. of* $V_n(Y_1,\ldots,Y_n)$ *is given by*

$$\begin{aligned} D_n(x) &= 1 - \sum_{i=\lfloor xn+1 \rfloor}^{n} \binom{n}{i} \left[1 - G\left(x\right)\right]^i \left[G\left(x\right)\right]^{n-i} &(8)\\ &= I\left(G\left(x\right); n - \lfloor xn \rfloor, \lfloor xn \rfloor + 1\right) &(9) \end{aligned}$$

for $x \in [0,1)$*, where* $I(p; a, b)$ *is the regularized Euler beta function and* $\lfloor y \rfloor := \max\{i \in \mathbb{N} : i \le y\}$ *is the floor function.*

Proof. The c.d.f. of the ith order statistic $Y_{i:n}$, $i = 1, 2, \ldots, n$, is given by

$$\begin{aligned} G_{i:n}(x) &= \Pr(Y_{i:n} \le x) \\ &= \frac{\Gamma(n+1)}{\Gamma(i)\,\Gamma(n-i+1)} \int_0^{G(x)} t^{i-1}(1-t)^{n-i}\,dt \\ &= I(G(x); i, n-i+1). \end{aligned}$$

Note that V_n (by Lemma 1) is equal to the greatest number such that $\lceil n\,V_n \rceil = \min\{i \in \mathbb{N} : i \ge n\,V_n\}$ observations are not less than V_n. Hence

$$\Pr(V_n > x) = \Pr(Y_{n-\lfloor xn+1 \rfloor:n} > x) = 1 - I\left(G\left(x\right); n - \lfloor xn \rfloor, \lfloor xn \rfloor + 1\right),$$

and the lemma follows immediately. □

Lemma 3. *For any* $x \in (0,1)$ *we have*

$$\Pr(V_n > x) = \Pr(1 - x > \hat{G}_n(x)). \tag{10}$$

Proof. Since $n\,\hat{G}_n(x) \sim \mathrm{Bin}(n, G(x))$ then for any $t \in (0, n)$

$$\Pr(n\,\hat{G}_n(x) \le t) = I(1 - G(x), n - \lfloor t \rfloor, 1 + \lfloor t \rfloor).$$

Now, by Lemma 2, we get for any $x \in (0,1)$

$$\begin{aligned}
\Pr(V_n > x) &= 1 - I\left(G\left(x\right); n - \lfloor xn \rfloor, \lfloor xn \rfloor + 1\right) \\
&= I\left(1 - G(x); \lfloor xn \rfloor + 1, n - \lfloor xn \rfloor\right) \\
&= I\left(1 - G(x); n - (n - \lfloor xn \rfloor - 1), 1 + (n - \lfloor xn \rfloor - 1)\right) \\
&= \Pr(n\,\hat{G}_n(x) \le n - (\lfloor xn \rfloor + 1)) \\
&= \Pr(\hat{G}_n(x) < 1 - x),
\end{aligned}$$

which holds because $\lfloor xn \rfloor \le xn < \lfloor xn \rfloor + 1$. Thus the proof is complete. □

The following lemma (see [6]) will be also useful.

Lemma 4 (Hoeffding's inequality). *Let $(Z_1, \ldots, Z_n)$ be a sequence of independent random variables with finite second moments and let $0 \le Z_i \le 1$ for $i = 1, \ldots, n$. Then for any $t > 0$ the following inequality holds*

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} Z_i - \tfrac{1}{n}\mathbb{E}\sum_{i=1}^{n} Z_i \ge t\right) \le e^{-2nt^2}. \tag{11}$$

The next lemma shows that the S-statistic V_n converges to ϱ exponentially fast.

Lemma 5. *For any $n \in \mathbb{N}$ and $\varepsilon > 0$*

$$\Pr(|V_n - \varrho| > \varepsilon) \le 2e^{-2n\delta^2}, \tag{12}$$

where $\delta = G(\varrho + \varepsilon) - (1 - (\varrho + \varepsilon)) \wedge 1 - (\varrho - \varepsilon) - G(\varrho - \varepsilon)$.

Proof. It is worth noticing that the proof of this lemma would be analogous to that of Theorem 2.3.2 [7] where a similar result on sample quantiles is discussed. Note that $\mathbf{1}(Y_i > \varrho + \varepsilon)$ has of course finite second moments. For any $\varepsilon > 0$ we get (by Lemmas 3 and 4)

$$\begin{aligned}
\Pr(V_n > \varrho + \varepsilon) &= \Pr(1 - \varrho - \varepsilon > \hat{G}_n(\varrho + \varepsilon)) \\
&= \Pr\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{1}(Y_i > \varrho + \varepsilon) > \varrho + \varepsilon\right) \\
&= \Pr\big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{1}(Y_i > \varrho + \varepsilon) - (1 - G(\varrho + \varepsilon)) \\
&\qquad > \varrho + \varepsilon - (1 - G(\varrho + \varepsilon))\big) \\
&= \Pr\big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{1}(Y_i > \varrho + \varepsilon) - \tfrac{1}{n}\mathbb{E}\textstyle\sum_{i=1}^{n} \mathbf{1}(Y_i > \varrho + \varepsilon) \\
&\qquad > G(\varrho + \varepsilon) + \varrho + \varepsilon - 1\big) \\
&\le \exp\left\{-2n\,\delta_1^2\right\}.
\end{aligned}$$

On the other hand we have

$$\begin{aligned}
\Pr(V_n < \varrho - \varepsilon) &\le \Pr(1 - \varrho + \varepsilon \le \hat{G}_n(\varrho - \varepsilon)) \\
&= \Pr\big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{1}(Y_i \le \varrho + \varepsilon) - G(\varrho - \varepsilon) \\
&\qquad \ge 1 - (\varrho - \varepsilon) - G(\varrho - \varepsilon)\big) \\
&\le \exp\left\{-2n\,\delta_2^2\right\}
\end{aligned}$$

for $\delta_2 = 1 - (\varrho - \varepsilon) - G(\varrho - \varepsilon)$.

Hence $\Pr(|V_n - \varrho| > \varepsilon) = \Pr(V_n > \varrho + \varepsilon) + \Pr(V_n < \varrho - \varepsilon) \leq 2\exp\{-2n(\min\{\delta_1, \delta_2\})^2\}$, which completes the proof. □

Now we are ready to prove the desired result.

Theorem 1. *V_n is a strongly consistent estimator of ϱ.*

Proof. $\Pr(|V_n - \varrho| > \varepsilon) \to 0$ exponentially fast (by Lemma 5) w.r.t. n and therefore we get $V_n \xrightarrow{a.s.} \varrho$ (by Theorem 1.3.4 in [7]). □

4 Asymptotic Distribution of S-statistics

Unfortunately, the practical usage of the exact distribution (9) may sometimes be problematic. Therefore we are seriously interested in its approximation. In the present section we consider the asymptotic distribution of an S-statistic.

Let us also cite a well-known result that will be needed for proving the next theorem.

Lemma 6 (Berry-Esséen). *Let $Z_1, Z_2, \ldots$ denote a sequence of i.i.d. random variables with a finite expectation μ and finite variance σ^2 and such that $(\forall i)\ \mathbb{E}\,|Z_i - \mu|^3 < \infty$. Then for all $n \in \mathbb{N}$*

$$\sup_x |H_n(x) - \Phi(x)| \leq C \frac{\mathbb{E}\,|Z_1 - \mu|^3}{\sigma^3\sqrt{n}}, \tag{13}$$

where

$$H_n(x) = \Pr\left(\frac{\sum_{i=1}^n Z_i - n\mu}{\sigma\sqrt{n}} \leq x\right),$$

$\Phi(x)$ denotes the c.d.f. of the standard normal distribution and C is a positive constant independent of the distribution of Z_i.

This lemma characterizes the rate of convergence in the Lindeberg-Lévy Central Limit Theorem. Let us mention that the best currently known upper bound for C is 0,7056 (see [8]). Now we can present the asymptotic distribution of the S-statistic.

Theorem 2. *If G is a c.d.f. differentiable at ϱ, then*

$$V_n \xrightarrow{D} \mathrm{N}\left(\varrho, \frac{1}{1 + G'(\varrho)}\sqrt{\frac{\varrho(1-\varrho)}{n}}\right). \tag{14}$$

Proof. Let $x \in (0, 1)$ and $A > 0$ be a positive constant which will be determined later. Let

$$K_n(x) = \Pr\left(\frac{V_n - \varrho}{A}\sqrt{n} \le x\right).$$

We will show that $K_n(x) \to \Phi(x)$ as $n \to \infty$.

By Lemma 3 we have

$$\begin{aligned} K_n(x) &= \Pr\left(V_n \le \varrho + xAn^{-0,5}\right) \\ &= \Pr\left(1 - \varrho - xAn^{-0,5} \le \hat{G}_n(\varrho + xAn^{-0,5})\right). \end{aligned}$$

Assuming that $\Delta_{n,x} := \varrho + xAn^{-0,5}$ and recalling that $n\hat{G}_n(\Delta_{n,x}) \sim \mathrm{Bin}(n, G(\Delta_{n,x}))$ we obtain

$$K_n(x) = \Pr\left(\frac{n\hat{G}_n(\Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1 - G(\Delta_{n,x}))}} \ge \frac{n(1 - \Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1 - G(\Delta_{n,x}))}}\right).$$

Substituting $Z^*_{n,x}$ and $\zeta_{n,x}$ given by

$$\begin{aligned} Z^*_{n,x} &= \frac{n\hat{G}_n(\Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1 - G(\Delta_{n,x}))}} \\ \zeta_{n,x} &= \frac{n(1 - \Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1 - G(\Delta_{n,x}))}} \end{aligned}$$

into the previous equation we get $K_n(x) = \Pr(Z^*_{n,x} \ge \zeta_{n,x})$.

If $Z_1 \sim \mathrm{Bern}(G(\Delta_{n,x}))$, then $\mathbb{E}\,|Z_1 - \mathbb{E}\,Z_1|^3 = G(\Delta_{n,x})(1 - G(\Delta_{n,x}))((1 - G(\Delta_{n,x}))^2 + G(\Delta_{n,x})^2)$ (hence is finite) and $\mathrm{Var}\,Z_1 = G(\Delta_{n,x})(1 - G(\Delta_{n,x}))$.

By Lemma 6 for some $C > 0$ we obtain

$$\left|\Pr\left(Z^*_{n,x} < \zeta_{n,x}\right) - \Phi(\zeta_{n,x})\right| \le \frac{C}{\sqrt{n}}\,\frac{(1 - G(\Delta_{n,x}))^2 + G(\Delta_{n,x})^2}{\sqrt{G(\Delta_{n,x})(1 - G(\Delta_{n,x}))}} \overset{n\to\infty}{\to} 0,$$

because $G(\Delta_{n,x})(1 - G(\Delta_{n,x})) \overset{n\to\infty}{\to} (1 - \varrho)\,\varrho > 0$, and since G is continuous at ϱ. Finally we have

$$\begin{aligned} |\Phi(x) - K_n(x)| &= |\Pr\left(Z^*_n < \zeta_{n,x}\right) - (1 - \Phi(x))| \\ &= |\Phi(x) - \Phi(-\zeta_{n,x}) + \Pr\left(Z^*_n < \zeta_{n,x}\right) - \Phi(\zeta_{n,x})| \\ &\le |\Phi(x) - \Phi(-\zeta_{n,x})| + |Pr(Z^*_n < \zeta_{n,x}) - \Phi(\zeta_{n,x})| \\ &\to |\Phi(x) - \Phi(-\zeta_{n,x})|. \end{aligned}$$

Since our theorem will be proved when $|\Phi(x) - \Phi(-\zeta_{n,x})| \to 0$ we would determine A in such way that $-\zeta_{n,x} \to x$. It is seen that

$$
\begin{aligned}
-\zeta_{n,x} &= \frac{1}{\sqrt{G(\Delta_{n,x})\,(1-G(\Delta_{n,x}))}}\;\frac{1-\Delta_{n,x}-G(\Delta_{n,x})}{n^{-0,5}} \\
&= \frac{xA}{\sqrt{G(\Delta_{n,x})\,(1-G(\Delta_{n,x}))}}\;\frac{1-\varrho-xAn^{-0,5}-G(\varrho+xAn^{-0,5})}{xAn^{-0,5}} \\
&= -\frac{xA}{\sqrt{G(\Delta_{n,x})\,(1-G(\Delta_{n,x}))}}\;\frac{G(\varrho+xAn^{-0,5})-G(\varrho)+xAn^{-0,5}}{xAn^{-0,5}} \\
&\overset{n\to\infty}{\to} -\frac{xA}{\sqrt{(1-\varrho)\,\varrho}}\,(G'(\varrho)+1)
\end{aligned}
$$

and hence our desired $A = \sqrt{\varrho\,(1-\varrho)}/(1+G'(\varrho))$, QED. □

Note that Theorem 2 implies that V_n is (weakly) consistent. In practice, D_n approaches the normal distribution D_n^* quite quickly. For example if G is a beta distribution $B(0.5, 0.5)$ ($\rho = 0.5$) then for $n = 30$ we have $||D_n - D_n^*||_2 \simeq 0.013$ and $\max|D_n - D_n^*| \simeq 0.072$. and for $B(10, 3)$ ($\rho \simeq 0.713494$) $||\cdot||_2 \simeq 0.009$ and $\max|\cdot| \simeq 0.071$.

5 Conclusions

L-statistics are well-known aggregation operators useful in robust statistics. S-statistics considered in this paper possess some desired statistical properties which make them useful in many areas. Asymptotic normality proved in this paper enables interval estimation and the construction of statistical tests. Of course, some questions remain open. In particular, the problem of finding well-behaving estimators of $G'(\varrho)$ required for the above-mentioned constructions have to be considered in further research.

References

[1] Dubois D, Prade H (1986) Weighted minimum and maximum operations in fuzzy set theory. Inform. Sci. 39:205–210
[2] Dubois D, Prade H, Testemale C (1988) Weighted fuzzy pattern matching. Fuzzy Sets and Systems 28:313–331
[3] Gągolewski M, Grzegorzewski P (2010) Arity-monotonic extended aggregation operators. In: E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.), IPMU 2010, Part I, CCIS 80:693-702
[4] Gągolewski M, Grzegorzewski P (2010) Possibilistic extension of some aggregation operators. Submitted
[5] Grabisch M, Pap E, Marichal JL, Mesiar R (2009) Aggregation Functions. Cambridge

[6] Hoeffding W (1963) Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58(301):13–30

[7] Serfling RJ (1980) Approximation theorems of mathematical statistics. John Wiley & Sons, New York

[8] Shevtsova IG (2007) Sharpening of the upper bound of the absolute constant in the berry-esseen inequality. Theory of Probability and its Applications 51(3)

[9] Torra V, Narukawa Y (2008) The h-index and the number of citations: two fuzzy integrals. IEEE Transactions on Fuzzy Systems 16(3):795–797

[10] Vaart AW (2000) Asymptotic Statistics. Cambridge

[11] Yager RR (1988) On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics 18(1):183–190

Please cite this paper as: Gągolewski M., Grzegorzewski P., S-Statistics and Their Basic Properties, In: Borgelt C. et al (Eds.), *Combining Soft Computing and Statistical Methods in Data Analysis*, Springer-Verlag, 2010, 281–288.