ON THE APPROXIMATION OF AN INTEGRAL
BY A SUM OF RANDOM VARIABLES

JOHN H.J. EINMAHL
Eindhoven University of Technology
Department of Mathematics and Computing Science
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

MARTIEN C.A. VAN ZUIJLEN
University of Nijmegen
Department of Mathematics
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

(Received October, 1996; Revised September, 1997)

We approximate the integral of a smooth function on $[0, 1]$, where values are only known at $n$ random points (i.e., a random sample from the uniform-(0, 1) distribution), and at 0 and 1. Our approximations are based on the trapezoidal rule and Simpson's rule (generalized to the non-equidistant case), respectively. In the first case, we obtain an $n^2$-rate of convergence with a degenerate limiting distribution; in the second case, the rate of convergence is as fast as $n^{3/2}$, whereas the limiting distribution is Gaussian then.

Key words: Numerical Integration, Order Statistics, Spacings.

AMS subject classifications: 60F05, 62G30, 65D30.

1. Introduction and Main Results

Suppose we (can) only observe the values of a smooth function $f: [0, 1] \rightarrow \mathbb{R}$ at the points $U_0, U_1, \ldots, U_n, U_{n+1}$, where $U_1, U_2, \ldots, U_n$ are the order statistics ($U_1 \leq U_2 \leq \ldots \leq U_n$) of $n$ independent uniformly-(0, 1) distributed random variables and $U_0; = 0, U_{n+1}; = 1$. It is our aim to estimate the integral

$$I: = \int_{0}^{1} f(x) dx$$

from these observations, i.e., by only using $(U_i, f(U_i)), i = 0, 1, \ldots, n + 1$. The first
An estimator we will employ is constructed by using the ‘trapezoidal rule’ on each subinterval $[U_{i-1}, U_i], \ i = 1, \ldots, n+1$. This rule approximates an integral $\int_a^b g(x)dx$ simply by $\frac{1}{2}(b-a)(g(a)+g(b))$ and it can easily be shown (see, e.g., Isaacson and Keller [2], p. 304) that

$$\frac{1}{2}(b-a)(g(a)+g(b)) - \int_a^b g(x)dx = \frac{1}{12}(b-a)^3 g^{(3)}(\eta), \quad (2)$$

where $\eta \in (a,b)$. Writing $D_i = U_i - U_{i-1}, \ i = 1, \ldots, n+1$, for the spacings of the $U_i$'s, our estimator of $I$ becomes

$$I_n^* = \sum_{i=1}^{n+1} \frac{1}{2} D_i (f(U_{i-1}) + f(U_i)). \quad (3)$$

Using (2), we will prove the following limiting result for the standardized difference of $I_n$ and $I$:

**Theorem 1:** If $|f^{(3)}|$ is bounded, then

$$n^2 (I_n - I) \xrightarrow{d} \frac{1}{2} (f'(1) - f'(0)), \quad \text{as } n \to \infty. \quad (4)$$

A much better and probabilistically more interesting estimator is obtained by applying a 3-points formula, i.e., for a given $c \in (a,b)$, we approximate $\int_a^b g(x)dx$ by

$$w_1 g(a) + w_2 g(c) + w_3 g(b)$$

in such a way that the approximation error is zero in the case $g$ is a polynomial of second degree. If the 3 points are equidistant, this approximation is known as Simpson’s rule. It is not hard to show that

$$w_1 = \frac{1}{6}(b-a) \left( 2 - \frac{b-c}{c-a} \right), \quad w_2 = \frac{1}{6} \frac{(b-a)^3}{(c-a)(b-c)}, \quad w_3 = \frac{1}{6}(b-a) \left( 2 - \frac{c-a}{b-c} \right),$$

and it follows (see again Isaacson and Keller [2], p. 304) that

$$w_1 g(a) + w_2 g(c) + w_3 g(b) - \int_a^b g(x)dx = -\frac{1}{6} \int_a^b (x-a)(x-c)(x-b)g^{(3)}(\eta)dx, \quad (6)$$

where $\eta = \eta(x) \in (a,b)$. Hence, our estimator of $I$ in (1), again denoted by $I_n$, becomes

$$I_n = \sum_{i=1}^{n+1} \frac{1}{6} (D_{2i-1} + D_{2i}) \left( (2 - \frac{D_{2i}}{D_{2i-1}}) f(U_{2i-2}) + \frac{(D_{2i-1} + D_{2i})^2}{D_{2i-1} D_{2i}} f(U_{2i-1}) + (2 - \frac{D_{2i-1}}{D_{2i}}) f(U_{2i}) \right), \quad (7)$$

where, for convenience, $n$ is taken to be odd. Formula (6) will be used to prove our main result:

**Theorem 2:** Let $n$ be odd. If $|f^{(5)}|$ is bounded, then

$$n^{3/2} (I_n - I) \xrightarrow{d} \sqrt{\frac{35}{3}} \int_0^1 (f^{(3)}(x))^2 dx Z, \quad \text{as } n \to \infty.$$

\[ \]
where $Z$ is a standard normal random variable.

**Remark 1:** The present techniques can be easily adapted to cover the situation where the $U_i$'s are the order statistics of $n$ independent random variables with common distribution function $G$ (on $(0,1)$) having a smooth density $g$. The adaptation is based on the quantile transform, transforming a uniform random variable $V$ into a random variable $G^{-1}(V)$ with distribution function $G$. In this case, under regularity conditions on $g$, we obtain that the weak limit in Theorem 1 becomes

$$\frac{1}{2} \int_0^1 \left( f''(x) / g^2(x) \right) dx$$

instead of $\frac{1}{2} \int f''(x) dx = \frac{1}{2}(f'(1) - f'(0))$. In Theorem 2, the limiting random variable is again centered normal but now the standard deviation becomes

$$\sqrt{\frac{35}{3} \int_0^1 \left( f^{(3)}(x) \right)^2 / g^3(x) dx}.$$

On the other hand, the uniform distribution seems very relevant because of the following. Since $\int_0^1 f(x) dx$ can be considered as the mean ‘output’, given that the $x$-values are ‘equally important’, it seems desirable to estimate $\int_0^1 f(x) g(x) dx = \int_0^1 f(G^{-1}(y)) dy$ in the case the random variables are distributed according to $G$. But if $G$ is known, we can replace the pairs $(U_i, f(U_i))$ (just below (1)), with $U_i$’s being the order statistics from $G$, by $(G(U_i), f(U_i)) = (G(U_i), f(G^{-1}(G(U_i))))$. This brings us back to the ‘uniform distribution setup’ with $f$ replaced by $f \circ G^{-1}$, but that is just the function whose integral we wanted to estimate as argued above!

This idea leads to possible ways of applying the results. Suppose $U_i$ represents some uncontrollable physical random quantity, like temperature, humidity or light intensity with a known distribution function $G$ having density $g$. Suppose also that we can measure $f$ (the output or yield) only at the $U_i$ and that we are interested in the mean output $I_g = \int_0^1 f(x) g(x) dx$. Then one can use our theorems to obtain rapidly converging estimators of $I_g$. In particular, when measuring the $f$-values is hard or expensive, one can get good estimators based on a few observations.

Also note that for the trapezoidal rule in Theorem 1 and $f''$ being constant, the uniform distribution is optimal, since $\int_0^1 g^{-2}(x) dx \geq \int_0^1 1 dx = 1$. (This can be easily seen by using Jensen’s inequality:

$$\int_0^1 \frac{1}{g^2(x)} dx = \int_0^1 \frac{1}{g^3(x)} g(x) dx = E\frac{1}{g(X)} \geq \left(E\frac{1}{g(X)}\right)^3 = \left(\int_0^1 \frac{1}{g(x)} g(x) dx\right)^3 = 1,$$

where $X$ is a random variable with density $g$.) A similar remark applies to Theorem 2 with $f^{(3)}$ being constant.

**Remark 2:** There are various other ways to extend our results, which we will not pursue here, e.g., applying $m$-points formulas for $m > 3$ (Simpson’s rule is ‘by far the
most frequently used in obtaining approximate integrals', Davis and Rabinowitz [1], p. 45), combining trapezoidal rules to eliminate the bias \( \frac{1}{2}(f'(1) - f'(0)) \), proving a 'second order' limit result for \( n^2(I_n - I) - \frac{1}{2}(f'(1) - f'(0)) \) in Theorem 1, or treating the case \( n \) 'even' in Theorem 2. We are not pursuing these extensions because we believe they are not very interesting and/or they do not give good results.

**Remark 3:** We briefly compare our results with the deterministic, equidistant case, i.e., \( U_i = \frac{i}{n+1} \), \( i = 0, 1, \ldots, n+1 \). It is well-known that the limit in Theorem 1 is \( \frac{1}{12}(f'(1) - f'(0)) \) in that case, which means that we loose a factor of 6 by having random \( U_i \)'s. (Essentially, this 6 is coming from the third moment of a standard exponential random variable.) From Theorem 2, it is well-known that in the equidistant case (Simpson’s rule), the rate is \( n^4 \). So, there our loss is of order \( n^{1/2} \). Nevertheless, from statistical point of view, \( n^{3/2} \) is a remarkably fast rate of convergence.

## 2. Proofs

The following well-known lemma will be used frequently; it can be found in, e.g., Shorack and Wellner [3], p. 721.

**Lemma 1:** Let \( E_1, \ldots, E_{n+1} \) be independent exponential random variables with mean 1 and \( S_{n+1} \) be their sum. With \( D_i, i = 1, \ldots, n+1 \), as before, we have

\[
(D_1, \ldots, D_{n+1}) \overset{d}{=} \left( \frac{E_1}{S_{n+1}}, \ldots, \frac{E_{n+1}}{S_{n+1}} \right).
\]

**Proof of Theorem 1:** Using (3), (1) and (2) we see that

\[
n^2(I_n - I) = \frac{n^2}{12} \sum_{i=1}^{n+1} D_i^3 f'''(\bar{U}_i)
\]

for some \( \bar{U}_i \in (U_{i-1}, U_i) \), and hence,

\[
n^2(I_n - I) = \frac{n^2}{12} \sum_{i=1}^{n+1} D_i^3 f'''\left( \frac{i}{n+1} \right) + \frac{n^2}{12} \sum_{i=1}^{n+1} D_i^3 (\bar{U}_i - \frac{i}{n+1}) f'''(\bar{U}_i),
\]

with \( \bar{U}_i \) between \( \bar{U}_i \) and \( \frac{i}{n+1} \). From the boundedness of \( |f'''| \) (by \( M \), say) and the weak convergence (to a Brownian bridge) of the uniform quantile process (see, e.g., Shorack and Wellner [3]), it is readily seen that

\[
\left| \frac{n^2}{12} \sum_{i=1}^{n+1} D_i^3 (\bar{U}_i - \frac{i}{n+1}) f'''(\bar{U}_i) \right| \leq \frac{n^2}{12} M \sup_{i \in \{1, \ldots, n+1\}} \left| \bar{U}_i - \frac{i}{n+1} \right| \sum_{i=1}^{n+1} D_i^3 = O_p\left(n^{\frac{11}{2}}\right) \sum_{i=1}^{n+1} D_i^3.
\]

But

\[
\sum_{i=1}^{n+1} D_i^3 \overset{d}{=} \frac{1}{S_{n+1}^3} \sum_{i=1}^{n+1} E_i^3,
\]

by Lemma 1, and by two applications of the weak law of large numbers, this last expression is \( O_p(n^{-2}) \). Combining this with (10) and (11) yields that the second term on the right in (9) converges to zero in probability. Hence, it remains to consider the first term.
\[ \frac{n^2}{12} \sum_{i=1}^{n+1} D_i^3 f''(\frac{i}{n+1}) = \frac{1}{12n} \sum_{i=1}^{n+1} f''(\frac{i}{n+1}) E_i^3, \]

or, since \((n/S_{n+1})^3 \rightarrow 1,\)

\[ \frac{1}{12n} \sum_{i=1}^{n+1} f''(\frac{i}{n+1}) E_i^3. \]

By Chebysev’s inequality, it follows that

\[ \frac{1}{12n} \sum_{i=1}^{n+1} f''(\frac{i}{n+1}) E_i^3 - \frac{1}{2n} \sum_{i=1}^{n+1} f''(\frac{i}{n+1}) \rightarrow 0. \]

The proof is complete by noting that

\[ \frac{1}{2n} \sum_{i=1}^{n+1} f''(\frac{i}{n+1}) \rightarrow \frac{1}{2} \int_0^1 f''(x) dx = \frac{1}{2}(f'(1) - f'(0)). \]

The proof of Theorem 2 is heavily based on the following two lemmas.

**Lemma 2:** Let \( E_1, \ldots, E_{n+1}, \ n \text{ odd,} \) be independent exponential random variables with mean 1. Write

\[ X_i = (E_{2i-1} + E_{2i})^3 (E_{2i} - E_{2i-1}), \ i = 1, 2, \ldots, \frac{n+1}{2}, \]

\[ Y_i = X_i \sum_{j=1}^{2i-2} (E_j - 1), \ i = 2, 3, \ldots, \frac{n+1}{2}. \]

Then,

\[ \mathbb{E} X_i = 0, \ \text{Var} X_i = 120960, \ \mathbb{E} Y_i = 0, \ \text{Var} Y_i = 120960(2i - 2), \]

\[ \text{Cov}(Y_i, Y_k) = 0, \ \text{for} \ i \neq k. \]

**Proof:** By symmetry, we see that \( \mathbb{E} X_i = 0; \) a straightforward computation yields \( \text{Var} X_i = \mathbb{E} X_i^2 = 120960. \) For the \( Y_i \)'s we have

\[ \mathbb{E} Y_i = \mathbb{E} X_i \mathbb{E} \sum_{j=1}^{2i-2} (E_j - 1) = 0, \]

\[ \text{Var} Y_i = \mathbb{E} Y_i^2 = \mathbb{E} X_i^2 \mathbb{E} \left( \sum_{j=1}^{2i-2} (E_j - 1) \right)^2 \]

\[ = \text{Var} X_i \text{Var} \left( \sum_{j=1}^{2i-2} E_j \right) = 120960(2i - 2), \]

and for \( i < k,\)

\[ \text{Cov}(Y_i, Y_k) = \mathbb{E} Y_i Y_k \]

\[ = \mathbb{E} X_k \left( \sum_{j=1}^{2i-2} (E_j - 1) \right) \mathbb{E} X_i \left( \sum_{j=1}^{2i-2} (E_j - 1) \right) \]

\[ = \mathbb{E} X_k \mathbb{E} \left( \sum_{j=1}^{2k-2} (E_j - 1) \right) \mathbb{E} X_i \left( \sum_{j=1}^{2i-2} (E_j - 1) \right) \]

\[ = 0. \]
Lemma 3: Under the conditions of Theorem 2, we have, as \( n \to \infty \),

\[
\left| \frac{3^{1/2}}{n} (I_n - I) - \frac{3^{1/2}}{72} \sum_{i=1}^{n+1} f^{(3)}(\frac{2i-2}{n+1}) (D_{2i-1} + D_{2i})^3 (D_{2i} - D_{2i-1}) \right| = o_p(1).
\]

Proof: By (7), (1) and (6) we have

\[
n^{3/2}(I_n - I) = -\frac{n^{3/2}}{6} \sum_{i=1}^{n+1} \int_{U_{2i-2}}^{U_{2i}} \frac{1}{U_{2i-2}} (x - U_{2i-2})(x - U_{2i-1})(x - U_{2i}) f^{(3)}(\tilde{U}_{2i}) dx, \tag{12}
\]

for some \( \tilde{U}_{2i} = \tilde{U}_{2i}(x) \in (U_{2i-2}, U_{2i}) \) and hence for some \( \tilde{U}_{2i} = \tilde{U}_{2i}(x) \in (U_{2i-2}, U_{2i}) \), the right-hand side of (12) is equal to

\[
-\frac{n^{3/2}}{6} \sum_{i=1}^{n+1} \int_{U_{2i-2}}^{U_{2i}} (x - U_{2i-2})(x - U_{2i-1})(x - U_{2i}) f^{(3)}(\tilde{U}_{2i}) (\tilde{U}_{2i} - U_{2i-2}) f^{(4)}(\tilde{U}_{2i}) dx
\]

\[
= \frac{n^{3/2}}{72} \sum_{i=1}^{n+1} f^{(3)}(U_{2i-2})(D_{2i-1} + D_{2i})^3 (D_{2i} - D_{2i-1})
\]

Let \( M \) be a bound on \( |f^{(5)}| \) and all lower order derivatives of \( f \). Then the absolute value of this last term is bounded from above by

\[
M \frac{n^{3/2}}{6} \sum_{i=1}^{n+1} \int_{U_{2i-2}}^{U_{2i}} (x - U_{2i-2}) |x - U_{2i-1}| |(U_{2i} - x)(\tilde{U}_{2i} - U_{2i-2})| dx
\]

\[
\leq \frac{M}{6} \frac{n^{3/2}}{2} \sum_{i=1}^{n+1} (D_{2i-1} + D_{2i})^5 \leq \frac{M}{6} \frac{n^{3/2}}{2} \frac{1}{S_n^{5/2}} \sum_{i=1}^{n+1} (E_{2i-1} + E_{2i})^5 = o_p(1),
\]

due to Lemma 1 and two applications of the weak law of large numbers.

So, it suffices to show the convergence to zero in probability of

\[
\frac{n^{3/2}}{72} \sum_{i=1}^{n+1} \left( f^{(3)}(U_{2i-2}) - f^{(3)}(\frac{2i-2}{n+1}) \right) (D_{2i-1} + D_{2i})^3 (D_{2i} - D_{2i-1})
\]

\[
= \frac{n^{3/2}}{72} \sum_{i=2}^{n+1} \left( U_{2i-2} - \frac{2i-2}{n+1} \right) f^{(4)}(\frac{2i-2}{n+1}) (D_{2i-1} + D_{2i})^3 (D_{2i} - D_{2i-1})
\]
Approximation of an Integral by a Sum of Random Variables

\[ + \frac{3}{144} \sum_{i=2}^{n+1} \left( U_{2i-2} - \frac{2i-2}{n+1} \right)^2 f^{(5)}(\bar{U}_{2i-2})(D_{2i-1} + D_{2i})^3(D_{2i} - D_{2i-1}) \]

= : T_{1,n} + T_{2,n},

for some \( \bar{U}_{2i-2} \) between \( U_{2i-2} \) and \( \frac{2i-2}{n+1} \). By the weak convergence of the uniform quantile process,

\[ |T_{2,n}| \leq \frac{3}{144} M \sup_{i \in \{2,3,\ldots,n+1\}} \left( U_{2i-2} - \frac{2i-2}{n+1} \right)^2 \sum_{i=2}^{n+1} (D_{2i-1} + D_{2i})^4 \]

= \( O_p(n^{2}) \) \( \sum_{i=2}^{n+1} (D_{2i-1} + D_{2i})^4 \).

By Lemma 1 and twice the weak law of large numbers, this last expression is easily seen to be \( o_p(1) \). Hence, the proof of Lemma 3 is complete if we show \( T_{1,n} = o_p(1) \).

From Lemma 1 we obtain

\[ T_{1,n} \iff \frac{3}{72} \sum_{i=2}^{n+1} \left( \frac{2i-2}{S_{n+1}} - \frac{2i-2}{n+1} \right)^2 f^{(4)}(\frac{2i-2}{n+1}) \left( \frac{E_{2i-1} + E_{2i}}{S_{n+1}} \right)^3 \left( \frac{E_{2i} - E_{2i-1}}{S_{n+1}} \right) \]

\[ = \frac{3}{72} S_{n+1}^5 \sum_{i=2}^{n+1} \left( \sum_{j=1}^{2i-2} (E_{j-1}) \right) f^{(4)}(\frac{2i-2}{n+1}) (E_{2i-1} + E_{2i})^3 (E_{2i} - E_{2i-1}) \]

\[ + \frac{3}{72} S_{n+1}^5 \left( 1 - \frac{S_{n+1}}{n+1} \right) \sum_{i=2}^{n+1} (2i-2) f^{(4)}(\frac{2i-2}{n+1}) (E_{2i} - E_{2i-1}) \]

= : T_{3,n} + T_{4,n}.

It is immediate from the central limit theorem for \( S_{n+1}/(n+1) \) that

\[ T_{4,n} = O_p(n^{-2}) \sum_{i=2}^{n+1} (2i-2) f^{(4)}(\frac{2i-2}{n+1}) X_i, \]

where the \( X_i \)’s are as in Lemma 2. Now using that lemma in conjunction with Chebysev’s inequality, it readily follows that \( T_{4,n} = o_p(1) \). Finally, in the notation of Lemma 2,

\[ T_{3,n} = \frac{3}{72} S_{n+1}^5 \sum_{i=2}^{n+1} f^{(4)}(\frac{2i-2}{n+1}) Y_i = O_p(n^{-\frac{1}{2}}) \sum_{i=2}^{n+1} f^{(4)}(\frac{2i-2}{n+1}) Y_i. \]

From Lemma 2, we have

\[ \sum_{i=2}^{n+1} f^{(4)}(\frac{2i-2}{n+1}) Y_i = 0, \]

\[ \sum_{i=2}^{n+1} f^{(4)}(\frac{2i-2}{n+1}) Y_i = 0 \]

\[ \sum_{i=2}^{n+1} \left( f^{(4)}(\frac{2i-2}{n+1}) \right)^2 Y_i = O(n^2). \]
Now, Chebysev's inequality yields $T_{3,n} = o_p(1)$ and hence $T_{1,n} = o_p(1)$. □

**Proof of Theorem 2:** Given the lemmas, especially Lemma 3, the proof of Theorem 2 is rather easy. If \( \int_0^1 (f^{(3)}(x))^2 \, dx = 0 \), then $f^{(3)}(x) = 0$ for all $x \in [0,1]$ and hence trivially $I_n = I$, because $f$ is a polynomial of second degree. Therefore, we assume now \( \int_0^1 (f^{(3)}(x))^2 \, dx > 0 \). Using Lemma 1 we have

\[
\frac{n^{3/2}}{72} \sum_{i=1}^{n+1} f^{(3)}(\frac{2i-2}{n+1})(D_{2i} - D_{2i-1})^3(D_{2i} - D_{2i-1})
\]

\[
= \frac{1}{72} \sqrt{2} \left( \frac{n}{S_n + 1} \right)^4 \sum_{i=1}^{n+1} f^{(3)}(\frac{2i-2}{n+1}) (E_{2i} - E_{2i-1})^3(E_{2i} - E_{2i-1})
\]

By the weak law of large numbers and Lemma 3, it now remains to show Theorem 2 with $n^{3/2} (I_n - I)$ replaced by $W_n$. By Lemma 2, we see that $E W_n = 0$ and

\[
\text{Var} W_n = \frac{1}{2(72)^2} \sum_{i=1}^{n+1} \left( f^{(3)}(\frac{2i-2}{n+1}) \right)^2 120960 \to 35 \int_0^1 (f^{(3)}(x))^2 \, dx.
\]

Now, the Lindeberg central limit theorem applies, because of the boundedness of $|f^{(3)}|$, and it yields the result. □

**Acknowledgements**

The authors are indebted to A.O.H. Axelsson and B.J.W. Polman for helpful discussions and their useful comments.

**References**


