ON ESTIMATING THE MEAN ORDINATE OF A CONTINUOUS FUNCTION OVER A SPECIFIED INTERVAL

Walter Scott Overton

Institute of Statistics
Mimeograph Series 390
April 1964

ON ESTIMATING THE MEAN ORDINATE OF A CONTINUOUS FUNCTION OVER A SPECIFIED INTERVAL

by

WALTER SCOTT OVERTON

A thesis submitted to the Graduate Faculty
of North Carolina State
of the University of North Carolina at Raleigh
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

DEPARTMENT OF EXPERIMENTAL STATISTICS

RALEIGH

1964

APPROVED BY ADVISORY COMMITTEE

Chairman	

ABSTRACT

OVERTON, WALTER SCOTT. On Estimating the Mean Ordinate of a Continuous Function over a Specified Interval. (Under the direction of ALVA LEROY FINKNER).

In many real sampling problems, estimation of the mean ordinate can be interpreted as the estimation of the area, M, between a continuous function $y(\tau)$ and the τ axis, over some interval in τ, say $[o,w]$. In the present dissertation, conventional methods of estimating mean ordinate are evaluated under this interpretation by the methods of numerical analysis. Reliance is placed on Weierstrass' Approximation Theorem, as a consequence of which the continuous function $y(\tau)$ is treated as a polynomial of arbitrarily high degree. Specific cases are studied as polynomials of prescribed degree.

The general approach is shown to have utility, and a number of useful specific results are obtained. In particular, the classical quadrature formulae (and designs based on them) are shown to be very useful. An estimator of M is described, based on the integral of the least squares fitted polynomial over the interval of interest, and quadrature formulae are developed from orthogonal polynomials for the cases in which observations are equally spaced in τ.

Conventional methods of estimating variance of the conventional estimators of systematic sampling are examined by the methods of numerical analysis. Modifications are suggested, and a procedure proposed for construction of unbiased estimators of variance of the mean of a systematic random sample, given restrictions placed on the degree of the polynomial. In the more general problem of estimating

variances of quadrature estimators, supplementary observations appear to be the most useful source of information. Several problems remain unsolved in the use of residuals from the least squares fitted polynomials for this purpose.

In general, quadrature estimators of mean ordinate are considerably more accurate than the conventional estimators (usually the observed mean) of sampling over an interval in time or space. Classical quadrature formulae (e.g., Gaussian and Tchebysheff) are very accurate for a small number of observations, and the least squares quadratures seem ideally suited to larger n and polynomials of uncertain degree. The Newton-Cotes and Centric formulae will be valuable when observations must be equally spaced and when n is small.

BIOGRAPHY

Walter Scott Overton (Jr.) was born in Farmville, Virginia, October 3, 1925, the son of Walter Scott and Alice Mottley Overton. He attended Grammar School and High School in Farmville, and entered The Virginia Polytechnic Institute in 1942. With two years spent in Military Service (European Theater of Operations, 83rd Infantry Division), he completed the B. S. Degree in Forestry and Wildlife Conservation in 1948 and the M. S. Degree in Wildlife Conservation in 1950.

From 1950-1958 he was employed by the Florida Game and Fresh Water Fish Commission as a Biologist, Project Leader. The last year (1957-1958) of the tenure was spent at North Carolina State College, during which time graduate work in Experimental Statistics was begun.

In July, 1959, he was employed by the Department of Experimental Statistics, North Carolina State College, in the capacity of Assistant Statistician, and in September, 1963, was appointed Associate Professor of Biometry at Emory University, Atlanta, Georgia.

He is married to the former Joann Frances Price of Jacksonville, Florida, and has two children, Deborah and Michael.

ACKNOWLEDGEMENTS

Acknowledgements are due the members of my Graduate Committee, Drs. A. L. Finkner, R. L. Anderson, F. S. Barkalow, Jr., W. W. Hassler, D. G. Horvitz, and H. L. Lucas. In particular, Dr. Finkner is thanked for his encouragement, his many suggestions, a great deal of his time, and his very great personal interest in the completion of this work. Dr. Horvitz also deserves special mention for his contributions, as do Drs. Anderson and Lucas. These contributions are gratefully acknowledged.

Several other persons contributed materially to the development and completion of this research, including Drs. A. E. Grandage, J. G. Baldwin, Jr., G. C. Caldwell and H. R. van der Vaart. To all of these I extend my sincere appreciation.

The majority of the work was accomplished while I was employed on the Cooperative Fish and Game Statistics Project in the Department of Experimental Statistics, North Carolina State, and work was completed during my employment in the Department of Biometry, Emory University. I am indebted to both of these departments for the considerable amount of time away from regular duties.

Finally, thanks are due my wife, Joann, for much patience and encouragement and for typing the rough draft, and to Miss Ann Smithwick for typing the final copy.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

Chapter 1. INTRODUCTION

1.1 Orientation

Systematic sampling is an enigma of sampling theory and practice. It is widely recognized as an efficient and convenient method of sampling from a sequence or function in time or space, under a wide variety of conditions, yet less efficient or less convenient methods are frequently advocated in its stead. The often-realized gain in precision is through an increase of information (in the sense of variability) within the sample, resulting in a decrease in the variation between possible samples. Yet when such a gain is achieved, variance estimators based on the sample second moment have a positive bias that increases as the precision of systematic sampling increases. An important source of information is the order of the observation, yet this has seldom been considered in estimation of precision, and almost never in estimation of the mean ordinate.

A flurry of interest in the theory of systematic sampling was highlighted by important papers by Madow and Madow (1944), Cochran (1946), Yates (1949) and Quenouille (1949). Subsequently, other problem areas in sampling occupied most of the interest of those who might have worked on the problem, and little of real importance has been done in something like ten years. Even though several theoretical questions have remained unanswered, systematic sampling is widely used among practicing statisticians, particularly when estimates of precision are not required. This use is evidence of faith in the technique, despite the lack of theoretical respectability.

The present stimulus came from problems in Field Ecology. Not only does systematic sampling appeal to the intuition of the Field Ecologist from the standpoint of precision and utility of results, but in many cases there is no other practical sampling scheme. In making airplane counts of breeding waterfowl, or deer track counts, or dove call counts or a vegetation study, to name only a few of many examples, it is clearly impractical to travel long distances to get to a "random" plot. One might as well count while traveling or measure at intervals while traveling. It has occasionally been advocated that a random set of plots be set up, and then a shortest route selected for travel to cover all plots. This has two practical disadvantages in that 1) seldom is it possible to accurately locate random points or plots in a forest or marsh, and 2) one is not getting maximum information from the sampling scheme. Also, many of the measurements are a function of time as well as space, so that spatial randomization is only a partial solution.

The germ of the present approach to the problem was planted when the writer was introduced to Simpson's Rule in a course in basic calculus. Prior to that time, he had used graphical or simple arithmetical methods of computing the area under a curve of observations taken in time or space, and numerical integration was obviously a better way to do the same thing. Since then, a number of simple applications of Simpson's Rule have been made, most of which were in measurement of fishing pressure through periodic counts of fishermen. Field application seems satisfactory.

The use of Simpson's Rule, or other quadrature formulae, involves taking measurements of the system (function) at prescribed points along the abscissa (time or space). These measurements represent the ordinate at the "sample" points. The quadrature formula then computes the area under a polynomial of prescribed degree passing through the observed ordinates. In the simpler formulae, the observations are spaced at equal intervals, so that there is a close relationship with systematic (equal interval) sampling, and so that the formulae can be used with standard systematic samples. The more complex formulae involve non-equidistantly spaced observations, and allow exact fit to polynomials of higher degree from the same number of measurements. The latter can be considered to be systematic samples of a more general kind.

In comparing estimates obtained from these quadrature methods with the conventional estimators associated with sampling a function in time or space, it was natural to define the latter in terms of area under the curve. This device was considered briefly by Yates (1949) and also by Aitken (1939) and Moran (1950), though most of the literature is in terms of the mean ordinate. This must be considered a statistical habit, as area under the curve is often of real interest.

Although there is a simple functional relationship between area and mean ordinate, so that there is no mathematical difference in estimating one or the other, there is nevertheless a distinct advantage in framing the problem in terms of area under the curve, even when mean ordinate is of primary interest. "Area under the curve" immediately evokes an image of the curve and the end points of the interval of interest, with the observations superimposed _in place_ along the time or space axis. With this picture in mind, it is natural to estimate area under the curve by fitting a function to the observation, and integrating this function. However, when one considers estimators of the mean of a universe from which a sample is taken, one almost automatically thinks in terms of the first moment of the sampling distribution, and the structure of the sample in relation to the universe (when such structure exists) is overlooked.

The main body of sampling theory is in terms of discrete sampling units, and most sampling thought is oriented in this direction. However, many observations and measurements are "instantaneous", representing a point rather than an interval in the continuous sample space. Many others, properly defined, have the statistical characteristics of an "instantaneous" measurement, and may be interpreted as an ordinate to a curve or surface. While it is true that any real world sampling scheme must employ discrete units of time or space in describing the sample, there is nevertheless distinct utility in the evaluation of the sampling process on the continuous scale.

Further, the main body of sampling theory deals primarily with completely general unbiased estimation, with no assumptions regarding the nature of the systems or function being sampled. It must be considered another habit of Sampling that such assumptions are not commonplace. Statistics has long since made its peace with restrictions of this kind in other areas, viz., Design of Experiments.

That systematic sampling, as here treated, is a special case of time series is immediately apparent, and it is necessary to justify the fact that a traditional time series approach was not made. The primary reason is that many of the systems of immediate interest are more properly interpreted as functions than as stochastic processes and these were studied first. A logical next step would be the extension of these results to the stochastic case. In this respect, it can be noted that this extension would add a third objective, that of estimating the mean value of the process over a specified interval, to the two classical objectives of time series analysis, prediction and description of the process.

In the present dissertation, four conventional estimators of mean ordinate are expressed and evaluated as area (M) under a segment of curve. The mean square errors of the estimators, $\mathcal{E}(\hat{M} - M)^2$, are developed and compared. Estimators of the squared error of a particular sample are developed in special cases. Several quadrature formulae are summarized, and their mean square errors developed. These formulae are compared with one another, and with the conventional estimators. Comparisons are made in the one stage and the two stage sampling situations.

Due to the importance of small samples in many applications, primary attention is given to characteristics of small samples, particularly with respect to the quadrature methods.

1.2 The Problem

1.2.1 The Problem

The problem considered in this thesis is outlined as follows. From a set of observations of a continuous non-negative function $y(\tau)$, defined on the closed interval $[0, w]$, $w > 0$, it is desired to estimate the area, M, between the function and the τ-axis on the interval $[0, w]$. That is, the parameter of interest is

$$M = \int_0^w y(\tau)d\tau \ .$$

Such a function can be uniformly approximated to any degree of accuracy by a polynomial of sufficiently high degree, in accordance with Weirstrass' Approximation Theorem (see for example, Simmons, 1963, p. 154, or Kopal, 1955, p. 19).

If $y(\tau)$ is a continuous real function defined on the closed interval $[0, w]$, for every $\epsilon > 0$ there exists a polynomial of degree m,

$$P_m(\tau) = \sum_{i=0}^{m} A_i \tau^i \text{ , such that}$$

$$|P_m(\tau) - y(\tau)| < \epsilon \text{ for all } \tau \text{ in } [0, w] \text{ .}$$

It follows that,

$$|M - \int_0^w P_m(\tau) d\tau| < w\epsilon$$

so that by an appropriate choice of ϵ, a polynomial $P_{m'}$ can be chosen such that

$$|M - \int_0^w P_{m'}(\tau) d\tau| = |r|$$

can be made arbitrarily small.

On the basis of this result, $y(\tau)$ shall be treated as a polynomial of (in general) relatively high degree. However, it should be emphasized that the only general restrictions placed on $y(\tau)$ are those imposed by the above theorem.

Now, if there are n known (observed) values $\{y(t_i)\}$ of $y(\tau)$ at n completely arbitrary distinct values $\{t_i\}$ of τ, $0 \leq t_i \leq w$, it is possible to uniquely describe a polynomial P of degree n-1, such that $P(t_i) = y(t_i)$ for $\{t_i\}$. That is, for any set $\{t_i\}$ there exists a set of coefficients $\{W_i\}$, $\Sigma W_i = w$, such that

$$\hat{M}_n \doteq \sum_{i=1}^{n} W_i y(t_i) = M - R$$

and such that $R = 0$ if $y(\tau)$ can be expressed exactly by a power polynomial of degree $< n$, and where $\hat{M}_n$ and R are functions of $\{t_i\}$.

1.2.2 Conventional Methods

Conventional methods of estimating mean ordinate

$$\mu = M/w$$

are evaluated by use of the above theorem and its mathematical consequences, and by this method, estimators of variance of these conventional estimators are also developed and evaluated.

Further, this method is employed, in the form of the classical ideas of numerical integration, in development of estimators of M, or mean ordinate, and evaluation of these estimators. Specifically, numerical integration formulae are considered when $y(\tau)$ is observed with error. To this end, it is specified that

$$y_i = y(t_i) + \epsilon_i$$

where $\mathcal{E}\epsilon_i = 0$, $\mathcal{E}\epsilon_i^2 = \sigma^2$ and $\mathcal{E}\epsilon_i\epsilon_j = 0$, $i \neq j$. This specification will be maintained throughout the dissertation.

1.3 Review of Literature

1.3.1 One Dimension

The key papers in which the theory of Systematic Sampling in one dimension is investigated are Madow and Madow (1944), Cochran (1946), Yates (1949) and Madow (1949). This theory is well summarized by Cochran (1953).

Yates (1949) is the author of the major paper of the one dimensional case, in which he examined systematic sampling theoretically and empirically. He considered a number of variance estimators, several of which are partially evaluated in the present dissertation. In addition, he considered the systematic sample estimator as area under the curve, but did not follow up the idea. Yates departed from the conventional mean observation estimator in his "end correction for linear trend," which is here shown to be a simple quadrature method.

1.3.2 Two Dimensions

Systematic sampling in two dimensions has been theoretically studied by Quenouille (1950) and Das (1950) and examined empirically by Osborne (1942), Milne (1959), Haynes (1948) and others. (Osborne's methods were actually one dimensional, although applied to a two dimensional problem.) The paper by Quenouille (1950) was evidently stimulated by Yates' (1949) paper, and extends the one dimensional ideas of Cochran (1946) and Yates to the two dimensional case. The paper by Das (1950) is very close in scope to that of Quenouille (1950) but not as complete.

Empirical studies by Osborne (1942) and Haynes (1948) indicated that sizeable gains in precision could be obtained from systematic sampling in some natural populations. The study by Milne (1958) gave somewhat opposing results, in that he found the mean square among observations in a centric systematic sample to closely approximate the "random" mean square, implying little gain in precision over a random sample. The crux of the matter, as is known from the theory, is that the error of the systematic sample, in any given population, or function, is dependent on the chosen sampling interval. Thus, for a systematic sample to be self-evaluating, it is necessary that inferences about the function be made from the sample. This is the general approach of Yates (1949), and many of the present results are extensions of his.

1.3.3 Numerical Integration

The connection between systematic sampling and numerical integration was casually mentioned by Yates (1949) and earlier implied by Aitken (1939). Moran (1950) was stimulated by Yates' (1949) paper to write on the subject of "Numerical Integration by Systematic Sampling," but no one seems to have considered the possibility of using the established results of numerical integration to obtain more precise estimators. The lone exception is the forementioned "end-correction" of Yates (1949), which is in reality a first degree quadrature formula (Section 4.6), although he does not identify it as such.

However, in rereading Yates (1949) and Cochran (1953 and 1946) after completing the present thesis, one has the strong feeling that both authors knew many of the present results, or at least would not be very surprised by them.

1.3.4 Numerical Analysis

Several numerical analysis texts were studied in development of the results of this dissertation. The most valuable to this study is considered to be Kopal (1955), although Rierdan (1958), Whitaker and Robinson (1944), Kunz (1957), Guest (1961) and Booth (1955), all contributed to the results. Pertinent results in numerical analysis are given in Section 1.4.

1.4 A Summary of the Mathematics

A summary of the mathematics with useful identities and the development of general results follows.

1.4.1 Finite Differences

Define:
$$\Delta_a = y(a+h) - y(a)$$
$$\Delta_a^2 = \Delta_{a+h} - \Delta_a = y(a+2h) - 2y(a+h) + y(a)$$
$$\Delta_a^k = \Delta_{a+h}^{k-1} - \Delta_a^{k-1} \tag{1.1}$$

and
$$E\, y(a) = y(a+h)$$
$$E^2\, y(a) = y(a+2h)$$
$$E^n\, y(a) = y(a+nh)\ . \tag{1.2}$$

Then
$$\Delta_a = (E-1)y(a)$$
$$\Delta_a^2 = (E-1)^2 y(a) = (E^2-2E+1)y(a)$$
$$\Delta_a^b = (E-1)^n y(a)\ . \tag{1.3}$$

1.4.2 Expansion by Difference Operators

By the operators Δ and E, it is possible to derive many of the useful relationships. For example, if it is desired to expand $y(a+th)$ in differences about $y(a)$,

$$E^t\, y(a) = y(a+th)$$
$$E^t = (1+\Delta)^t$$
$$= 1 + t\Delta + \frac{t(t-1)}{2!}\Delta^2 + \ldots$$

a. If t is an integer, the expansion has $t+1$ terms,

$$y(a+th) = y(a) + t\Delta_a + \frac{t^{[2]}}{2!}\Delta_a^2 + \ldots + \Delta_a^t \tag{1.4}$$

b. If t is a fraction, the expansion is an infinite series,

$$y(a+th) = y(a) + t\Delta_a + \frac{t^{[2]}}{2!}\Delta_a^2 + \ldots + \frac{t^{[k]}}{k!}\Delta_a^k + \ldots, \tag{1.5}$$

where $t^{[k]} = \frac{t!}{(t-k)!}$, and where (1.5) is known as the Gregory-Newton Interpolation Formula.

1.4.3 The Correspondence of Difference and Derivative Expansions

The analogy between (1.5) and Taylor's expansion is immediately apparent. If y is analytic, then, by Taylor's expansion,

$$y(a+th) \doteq y(a) + th\, y^{(1)}(a) + \frac{(th)^2\, y^{(2)}(a)}{2!} + \ldots \tag{1.6}$$

In order to facilitate the manipulation of these infinite series, define,

$$\underline{D}_a' = (D_a, \frac{1}{2!}D_a^2, \frac{1}{3!}D_a^3, \ldots)$$

$$\text{where } D_a^k = y^{(k)}(a)$$

$$\underline{\Delta}_a' = (\Delta_a, \frac{1}{2!}\Delta_a^2, \frac{1}{3!}\Delta_a^3, \ldots)$$

$$\text{where } \Delta_a^k = \Delta_{a+h}^{k-1} - \Delta_a^{k-1}$$

$$\underline{t}_o^{*\prime} = (t, t^{[2]}, t^{[3]}, \ldots)$$

$$\underline{t}_o = (t, t^2, t^3, \ldots)$$

and
$$H = \begin{bmatrix} h & & & & & \\ & h^2 & & & & \\ & & h^3 & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & \cdot \end{bmatrix}$$

Then (1.5) can be expressed as

$$y(a+th) = y(a) + \underline{\Delta}_a' \underline{t}_o^* \tag{1.7}$$

and (1.6) as,

$$y(a+th) = y(a) + \underline{D}_a' H \underline{t}_o . \tag{1.8}$$

Thus, $\underline{D}_a' H \underline{t}_o = \underline{\Delta}' \underline{t}_o^* .$

Therefore, if there is a transformation matrix, A, such that

$$H \underline{D}_a = A' \underline{\Delta}_a , \tag{1.9}$$

then

$$A \underline{t}_o = \underline{t}_o^* . \tag{1.10}$$

Now, from (1.10), the n^{th} row of A is the set of coefficients, s(n,k), such that

$$t^{[n]} = \sum_{k=1}^{n} s(n,k) t^k ,$$

where the s(n,k) are known as Stirling Numbers of the First Kind. These numbers are tabled (e.g., Riordan, 1958, p. 48), or may be generated by the recursion formula,

$$s(n+1,k) = s(n,k-1) - ns(n,k) . \tag{1.11}$$

It follows directly that the n^{th} row of A^{-1} is the set of coefficients, S(n,k), such that

$$t^n = \sum_{k=1}^{n} S(n,k) t^{[k]}$$

and where the S(n,k) are known as Stirling Numbers of the Second Kind. These are also tabled by Riordan (1958, p. 48).

1.4.4 Useful Integrals of Functions of $\underline{t}_o$

Define $\underline{\mu}_o = \int_o^1 \underline{t}_o dt$

$$\underline{\mu}_o^* = \int_0^1 \underline{t}_o^* dt$$

$$T_o = (\underline{t}_o - \underline{\mu}_o)(\underline{t}_o - \underline{\mu}_o)'$$

$$T_o^* = (\underline{t}_o^* - \underline{\mu}_o^*)(\underline{t}_o^* - \underline{\mu}_o^*)'.$$

Then
$$T_o^* = AT_oA'$$

$$\int_0^1 T_o^* dt = A\left\{\int_0^1 T_o dt\right\} A'$$

and the ij^{th} element of $\int_0^1 T_o dt$ is

$$\int_0^1 t_{ij} dt = \int_0^1 (t^i - \mu_i)(t^j - \mu_j) dt,$$

where
$$\mu_i = \int_0^1 t^i dt = \frac{1}{i+1} ,$$

so,
$$\int_0^1 t_{ij} dt = \frac{1}{i+j+1} - \frac{1}{(i+1)(j+1)} = \frac{ij}{(i+1)(j+1)(i+j+1)} .$$

Then the ℓk^{th} element of T_o^* is

$$t_\ell^* = \sum_{j=1}^{k} \sum_{i=1}^{\ell} s(\ell,i)s(k,j)t_{ij} ,$$

and the ℓk^{th} element of $\int_0^1 T_o^* dt$,

$$\int_0^1 t_{\ell k}^* dt = \sum_{j=1}^{k} \sum_{i=1}^{\ell} \frac{(ij)s(\ell,i)s(k,j)}{(i+1)(j+1)(i+j+1)} .$$

1.4.5 Useful Integrals of Functions of $\underset{\sim}{t}_c$

$$\text{Define} \quad \underset{\sim}{t}_c' = [\ (t-\tfrac{1}{2}),\ (t-\tfrac{1}{2})^2,\ \ldots.]$$

$$\underset{\sim}{\mu}_c = \int_0^1 \underset{\sim}{t}_c dt$$

$$T_c = (\underset{\sim}{t}_c - \underset{\sim}{\mu}_c)(\underset{\sim}{t}_c - \underset{\sim}{\mu}_c)'$$

Then the ij^{th} element of $\int_0^1 T_c dt$ is,

$$\int_0^1 \tau_{ij} dt = \int_0^1 (t-\tfrac{1}{2})^{i+j} dt - \int_0^1 (t-\tfrac{1}{2})^i dt \int_0^1 (t-\tfrac{1}{2})^j dt$$

$$= \begin{cases} \dfrac{ij}{(i+j+1)(i+1)(j+1)} \left(\tfrac{1}{2}\right)^{i+j} & , \quad i \text{ and } j \text{ even} \\[2ex] \dfrac{1}{(i+j+1)} \left(\tfrac{1}{2}\right)^{i+j} & , \quad i \text{ and } j \text{ odd} \\[2ex] 0 & , \quad i+j \text{ odd} \end{cases}$$

1.4.6 Definition of $\underset{\sim}{C}^*$

$$\underset{\sim}{C}^* = T_o^* \Big|_{t=1/2}$$

$$= A\ T_o \Big|_{t=1/2} A' = A\ \underset{\sim}{C}\ A'$$

so that the ij^{th} element of $\underset{\sim}{C}$ is

$$c_{ij} = [(1/2)^i - \mathcal{E}t^i]\ [(1/2)^i - \mathcal{E}t^j]$$

1.4.7 Useful Constant Matrices

Some useful constant matrices can be written from the preceding paragraphs (1.4.4, 1.4.5, 1.4.6).

1. $$\int_0^1 T_o dt = \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & \frac{3}{40} & \cdots \\ \frac{1}{12} & \frac{4}{45} & \frac{1}{12} & \\ \frac{3}{40} & \frac{1}{12} & \frac{1}{112} & \\ \vdots & & & \ddots \end{bmatrix} \tag{1.12}$$

with ij^{th} element $= \frac{ij}{(i+1)(j+1)(i+j+1)}$

2. $$A = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & 0 \\ 2 & -3 & 1 & & \\ -6 & 11 & -6 & 1 & \\ 24 & -50 & 35 & -10 & 1 \\ \vdots & & & & & \ddots \end{bmatrix} \tag{1.13}$$

3. $$\int_0^1 T_o^* dt = \begin{bmatrix} 1/12 & 0 & -1/120 & \cdots \\ 0 & 1/180 & -1/120 & \\ -1/120 & -1/120 & 23/1480 & \\ \vdots & & & \ddots \end{bmatrix} \tag{1.14}$$

4. $$\int_0^1 T_c dt = \begin{bmatrix} 1/12 & 0 & 1/80 & \cdots \\ 0 & 1/180 & 0 & \\ 1/80 & 0 & 1/448 & \\ \vdots & & & \ddots \end{bmatrix} \tag{1.15}$$

5. $$\underset{\sim}{C} = \begin{bmatrix} 0 & 0 & 0 & 0 \ldots \\ 0 & 1/144 & 1/96 & 11/960 \ldots \\ 0 & 1/96 & 1/64 & 11/640 \\ 0 & 11/960 & 11/640 & 121/6400 \\ \vdots & & & & \ddots \end{bmatrix} \tag{1.16}$$

6. $$\underset{\sim}{C}^{*} = \begin{bmatrix} 0 & 0 & 0 & 0 \ldots \\ 0 & 1/144 & -1/96 & 73/3880 \ldots \\ 0 & -1/96 & 1/64 & -73/1920 \ldots \\ 0 & 73/2880 & -73/1920 & 5329/57600 \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix} \tag{1.17}$$

and where $\underset{\sim}{C}^{*} = \underline{C}\ \underline{C}'$

where $\underline{C}' = (0, -1/12, -1/8, -11/80, \ldots)\ A'$

$$= (0, -1/12, 1/8, -73/240, \ldots) \tag{1.18}$$

1.4.8 Differences of Lag n

From the definition of Δ_a, it follows that a difference of lag n

is defined as, $$\overset{n}{\triangle}_a = \sum_{i=a}^{n+a-1} \Delta_i = E^{n+a} - E^a \tag{1.19}$$

Thus, $\overset{n}{\triangle}_o = E^n - 1$

$$= (1 + \Delta_o)^n - 1$$

$$= n\Delta_o + \frac{n(n-1)}{2!}\Delta_o^2 + \ldots + \Delta_o^n \tag{1.20}$$

$$\overset{n}{\triangle}{}_o^2 = (E^n - 1)^2$$

$$= E^{2n} - 2E^n + 1$$

$$= 2n\Delta_o + \frac{2n(2n-1)}{2!}\Delta_o^2 + \ldots + \Delta_o^{2n}$$

$$- \left[2n\,\Delta_o + \frac{2n(n-1)}{2!}\Delta_o^2 + \ldots + 2\Delta_o^n\right] + 1$$

(recalling that the 1 appearing in the last term of the sum is the identity operator),

and, in general,

$$\Delta_o^k = (E^n - 1)^k$$

$$= E^{kn} - k^{[1]}E^{n(k-1)} + \frac{k^{[2]}E^{n(k-2)}}{2!} - \ldots + (-1)^k. \quad (1.21)$$

1.4.9 Distinction Between the Symbolic Squaring and the Arithmetic Squaring Operation

A notational difficulty is encountered in using finite differences in quadratic terms, in that confusion easily occurs between the symbolic squaring operation and the arithmetic squaring operation. To illustrate, the second difference of lag 1 has been defined,

$$\Delta_a^2 = y(a) - 2y(a+h) + y(a+2h) \ ,$$

whereas the square of the first difference of lag 1 is

$$(\Delta_a^2) = [y(a+h) - y(a)]^2$$

$$= [y(a)]^2 - 2\,y(a)y(a+h) + [y(a-h)]^2$$

This point is made at this time because the following paragraph, 1.4.10, deals with the arithmetic analog of the symbolic operation treated in paragraph, 1.4.8. However, this distinction must be maintained throughout, and the arithmetic power of any difference operator will be denoted by brackets, as $(\Delta_a)^k$.

1.4.10 Identities

In this paragraph are developed several identities helpful in comparing expressions arising later, particularly in Chapter 3. These identities have not been developed in general, but rather under the restriction that $\Delta^3 \equiv 0$. A more general development would seem desirable.

$$1. \quad (\mathbf{\Delta})^2 \doteq \left(\sum_{i=0}^{n-1} \Delta_i\right)^2$$

$$= n \sum_{i=0}^{n-1} (\Delta_i)^2 - \frac{n^2(n^2-1)}{12} (\Delta^2)^2 . \tag{1.22}$$

Proof:

$$\left(\sum_{i=0}^{n-1} \Delta_i\right)^2 = \underline{1}' \begin{bmatrix} (\Delta_0)^2 & \Delta_0\Delta_1 & \cdots & \Delta_0\Delta_{n-1} \\ \Delta_1\Delta_0 & (\Delta_1)^2 & \cdots & \Delta_1\Delta_{n-1} \\ \vdots & & & \vdots \\ \Delta_{n-1}\Delta_0 & \cdots & \cdots & (\Delta_{n-1})^2 \end{bmatrix} \underline{1}$$

Then, if $\Delta^3 \equiv 0$,

$$(\mathbf{\Delta})^2 = \underline{1}' \begin{bmatrix} (\Delta_0)^2 & \Delta_0(\Delta_0+\Delta^2) \cdots & \Delta_0(\Delta_0+(n-1)\Delta^2) \\ \Delta_1(\Delta_1-\Delta^2) & (\Delta_1)^2 \cdots & \Delta_1(\Delta_1+(n-2)\Delta^2) \\ \vdots & & \vdots \\ \Delta_{n-1}(\Delta_{n-1}-(n-1)\Delta_0) & \cdots & (\Delta_{n-1})^2 \end{bmatrix} \underline{1}$$

$$= n \sum_{i=0}^{n-1} (\Delta_i)^2 + G ,$$

where, $G = \Delta^2 \left\{ \begin{array}{l} \Delta_o[1 + 2 + \ldots + (n-1)] \\ + \Delta_1[2 + \ldots + (n-2)] \\ \vdots \\ + \Delta_{n-1}[-(n-1) - (n-2) \ldots -1] \end{array} \right.$

$$= \Delta^2 \left\{ \begin{array}{l} (\Delta_o - \Delta_{n-1}) [1 + 2 + \ldots + (n-1)] \\ + (\Delta_1 - \Delta_{n-2}) [2 + \ldots + (n-2)] \\ \vdots \\ + \left\{ \begin{array}{l} \left[\Delta_{(\frac{n-2}{2})} - \Delta_{(\frac{n}{2})}\right] \left[n/2\right] \quad , n \text{ even} \\ \text{or } 0, \ n \text{ odd} \end{array} \right. \end{array} \right.$$

$$= -(\Delta^2)^2 \left\{ \begin{array}{l} (n-1) \left(\frac{n-1}{2}\right) (n) \\ + (n-3) \left(\frac{n-3}{2}\right)(n) \\ \vdots \\ + \left\{ \begin{array}{l} n/2, \ n \text{ even} \\ 0 \ , \ n \text{ odd} \end{array} \right. \end{array} \right.$$

$$= -\left\{ \begin{array}{ll} (\Delta^2)^2 \frac{n}{2} \sum_{k=1}^{n/2} (2k-1)^2 \ , & n \text{ even} \\ (\Delta^2)^2 \frac{n}{2} \sum_{k=1}^{(n-1)/2} (2k)^2 \ , & n \text{ odd} \end{array} \right.$$

Therefore, $G = -\frac{n^2}{12} (n^2 - 1) (\Delta^2)^2$, n odd or even.

2. $\Delta_i = \Delta_{i-1} + \Delta^2$ (1.23)

3. $(\Delta_i)^2 = (\Delta_{i-1})^2 + 2\Delta^2\Delta_{i-1} + (\Delta^2)^2$ (1.24)

4. $\Delta_{n-1} = \frac{1}{n-1}\left\{\sum_{i=o}^{n-2} \Delta_i + \frac{n}{2}(n-1)\Delta^2\right\}$ (1.25)

5. $(\Delta_{n-1})^2 = (\frac{1}{n-1})\left\{\sum_{i=o}^{n-2}(\Delta_i)^2 + \frac{n(n-1)(n+1)}{6}(\Delta^2)^2 + n\Delta^2\sum_{i=o}^{n-2}\Delta_i\right\}$

(1.26)

6. $\sum_{i=o}^{n-2}(\Delta_i)^2 = (\frac{n-1}{n})\sum_{i=o}^{n-1}(\Delta_i)^2 - \left\{\frac{(n-1)(n+1)}{6}(\Delta^2)^2 + \Delta^2\sum_{i=o}^{n-2}\Delta_i\right\}$

(1.27)

7. Let $X_i = \begin{cases} 1, & i = 0, 2, 4, \ldots \\ 0, & i = 1, 3, 5, \ldots \end{cases}$

Then

$$\sum_{i=o}^{n-2} X_i\Delta_i = \frac{1}{2}\left\{\sum_{i=o}^{n-1}\Delta_i - \frac{n}{2}\Delta^2\right\},\ n \text{ even.} \qquad (1.28)$$

8. $\left[\sum_{i=o}^{n-2} X_i\Delta_i\right]^2 = \frac{n}{4}\left\{\sum_{i=o}^{n-1}(\Delta_i)^2 - \Delta^2\sum_{i=o}^{n-1}\Delta_i - \frac{n}{12}(n^2-4)(\Delta^2)^2\right\}$

(1.29)

1.5 Notation and Definitions

1. $M = \int_0^w y(\tau)d\tau$

   = area under the function between 0 and w.

2. $\mu = M/w$ = mean ordinate over the interval $[0,w]$.

3. $y(\tau)$ is a continuous non-negative function of τ.

4. $y_i = y(\tau_i)$ or

   $= y(\tau_i) + \epsilon_i$, depending on the context.

5. M_r = Estimator of M based on a simple random sample of size n over an interval of length $w = hn$.

6. M_{st} = Estimator of M based on a stratified random sample, one sample at random in each of n strata of length h.

7. M_{sr} = Estimator of M based on a systematic random sample, first observation taken at random between o and h, and other n-1 at intervals of h.

8. M_{sc} = Estimator of M based on a centric systematic sample, observations taken at centers of n panels of length h.

9. $V(\hat{M})$ = Mean square error of the Estimator, $\hat{M}$.

10. δ^2 = Mean square successive difference of lag 1. (3.8)

11. δ^2_ℓ = Mean square successive difference of lag ℓ.

12. δ^2_* = Linear combination of mean square successive differences of different lag.

13. $\delta^2(Q)$ = Mean square successive difference between successive quadratures along an interval.

14. E = 1. The squared mean alternate difference 3.21, or
2. The symbolic operator, $E = \Delta + 1$, 1.4.1.

15. $\Delta_a = y(a+h) - y(a)$, the first difference of $y(a)$. It is usually understood that differences are over intervals of h, so that they are expressed in terms of the index,

$$\Delta_i = y_{i+1} - y_i, \text{ the first difference of } y_i = y(\tau_i)$$

16. $\Delta_i^k = \Delta_{i+1}^{k-1} - \Delta_i^{k-1}$, the k^{th} difference of y_i .

17. $\overset{n}{\Delta}_i = y_{i+n} - y_i = y[(i+n)h] - y[ih]$

18. $\underline{\Delta}^!_a = (\Delta_a, \frac{1}{2!}\Delta_a^2, \frac{1}{3!}\Delta_a^3, \ldots)$, 1.4.3

19. $D_i^k = y_i^{(k)} = y^{(k)}(\tau_i)$, the k^{th} derivative of $y_i = y(\tau_i)$

20. $\underline{D}^!_a = (D_a, \frac{1}{2!}D_a^2, \frac{1}{3!}D_a^3, \ldots)$, 1.4.3

21. $s(n,k)$ = Stirling numbers of the first kind, 1.4.3

22. $S(n,k)$ = Stirling numbers of the second kind, 1.4.3

23. $t^{[n]} = t!/(t-n)!$

24. t = 1. a uniform random variable (0,1) or,
    2. a sample value of τ

25. $\sim U(0,h)$ means: distributed as Uniform random variable on the interval $[0, h]$.

26. $f(t)$ = probability density function of t.

27. $\mathcal{E}_{\epsilon}$ = expectation over ϵ

28. $\mathcal{E}_{t}$ = expectation over t

29. $\sum_{i}$ indicates summation over i

30. $\sum_{i\epsilon s}$ indicates summation over all i in the sample.

31. $\underset{\sim}{t}_o' = (t, t^2, t^3, \ldots)$

32. $\underset{\sim}{t}_c' = \left[(t - h/2), (t - h/2)^2, \ldots\ldots\right]$

33. $\underline{\mu}' = \mathcal{E}(\underset{\sim}{t}_o') = (h/2, h^2/3, h^3/4, \ldots\ldots)$

34. $\underline{\mu}_c' = \mathcal{E}(\underset{\sim}{t}_c') = \left[0, \frac{1}{3}(h/2)^2, 0, \frac{1}{5}(h/2)^4, \ldots\ldots\right]$

$$= \begin{bmatrix} 0 & & & \\ & (\frac{1}{2})^2 & & \\ & & 0 & \\ & & & (\frac{1}{2})^4 \ddots \end{bmatrix} \cdot \underline{\mu}$$

35. $T_o = (\underline{t} - \underline{\mu})(\underline{t} - \underline{\mu})'$

36. $T_c = (\underset{\sim}{t}_c - \underline{\mu}_c)(\underset{\sim}{t}_c - \underline{\mu}_c)'$, equation (1.15) .

37. $\underset{\sim}{C} = \left. T_o \right|_{t = h/2}$, $\underset{\sim}{C}^* = A\underset{\sim}{C}A'$
    C = Some constant or coefficient, apparent by context.

38. $t_o^{*\prime} = (t, t^{[2]}, t^{[3]}, \ldots\ldots)$, 1.4.3

39. $T_o^* = (\underset{\sim}{t}_o^* - \underline{\mu}_o^*)\ (\underset{\sim}{t}_o^* - \underline{\mu}_o^*)'$, equation (1.14) .

40. $\underline{\mu}_o^* = \mathcal{E}\underset{\sim}{t}_o^*$

41. Q_n^A = a quadrature formula of n observation. Superscripts N, C, G, T indicate Newton-Cotes, Centric, Gaussian, and Tchebycheff.

42. $Q_{k,n}^A$ = a quadrature formula formed by successive application of a basic formula requiring k observations to a total of n observations.

43. $Q_n(t)$ = a quadrature formula of degree n, in which observations are equally spaced, with a random start.

44. $B'(n)$ is defined by formula (4.26).

45. $P(\tau)$ is a polynomial in τ of arbitrary degree.

46. $R = M - Q$, the remainder of a quadrature formula.

47. $\underline{W}' = (W_1, W_2, \ldots, W_n)$ the set of weights in a quadrature formula.

48. x is used in two ways,

    1. As a dummy variable, $x = (0, 1)$, in a number of summation formulae involving sub-sequences definable by such a variable.

    2. As a synonym for ht in the section involving results in terms of orthogonal polynomials. This definition applies to these immediately following.

49. $\underline{X}' = (1, X, X^2, \ldots\ldots)$

50. $$\underset{\sim}{X} = \begin{bmatrix} 1 & X_1 & X_1^2 & \cdots \\ 1 & X_2 & X_2^2 & \cdots \\ \vdots & \vdots & \vdots & \\ 1 & X_n & X_n^2 & \cdots \end{bmatrix}$$

51. $\underset{\sim}{\xi}$ = matrix of orthogonal polynomials, as given, for example, by Anderson and Houseman (1942). See paragraph 4.8.2.

52. $\underset{\sim}{O\lambda}$ is the orthogonalyzing transformation, defined in paragraph 4.8.2.

53. $\underset{\sim}{\psi} = (\underset{\sim}{X}\underset{\sim}{X}')^{-1} = \underset{\sim}{O\lambda}\,(\underset{\sim}{\xi}\underset{\sim}{\xi}')^{-1}\underset{\sim}{\lambda O}'$

54. k is used in a number of ways as a constant. Definition in any one section will be apparent from context.

Chapter 2. CONVENTIONAL ESTIMATORS FOR SAMPLING OVER AN INTERVAL

2.1 Introduction

The four conventional methods of sampling over an interval (simple random, stratified random, systematic with random start and centric systematic) are all conventionally used with the same estimator of mean ordinate, a simple mean of the observed ordinates. The different properties of the various schemes are due entirely to the different methods of selecting the abscissae at which the ordinate is to be observed or measured. In the present chapter, the expression for estimated area under the curve is developed for each of the sampling schemes in terms of the generalized function expansion.

The function considered is $y(\tau)$, and the observation y_i, made at the i^{th} value of τ is,

$$y_i = y(\tau_i) + \epsilon_i \ . \qquad (2.1)$$

As the contribution of ϵ to the estimate and to the variance of the estimate is additive, the terms involving ϵ will be dropped from the development of the general expressions, and reintroduced in Section 2.7, when comparisons are made.

In the following section, 2.2, the basic expressions are developed for a single sample point selected at random over the interval $[0, h]$. In subsequent sections of Chapter 2, the results of Section 2.2 are expanded for $n > 1$ to the various sampling schemes of interest over the interval $[0, w]$, where $w = nh$. The panels of length h and the results of Section 2.2 relating to them, are in reality the building blocks for subsequent results. In addition, all results are expressed in terms of differences defined over the panels of length h.

2.2 The Simple Random Sample, n = 1

2.2.1 The Difference Expansion of Sample Results

In sampling from the interval (0, h), it is natural to let $t \sim U(0, h)$. However, in using the Gregory-Newton formula, as well as in subsequent development, it is advantageous to let $t \sim U(0,1)$. A little awkwardness in phrasing the problem is more than compensated by later simplification.

Therefore, let a value of τ, ht, be selected at random between 0 and h, *i.e.*, $t \sim U(0,1)$, and let $y \doteq y(ht)$.

Define $$M = \int_0^h y(\tau)\, d\tau = h \int_0^1 y(th)dt \tag{2.2}$$

and $${}_1M_r = hy. \tag{2.3}$$

Then $$\mathcal{E}({}_1M_r) = \int_0^1 ({}_1M_r) f(t)dt = \int_0^1 ({}_1M_r)dt$$
$$= h \int_0^1 y(ht)dt \quad = M .$$

Recall that by the Gregory-Newton formula, $y(0 + th)$ can be expanded about $y(0)$ in terms of differences, where the differences are defined over the interval h.

$$y(t) = y(0) + t\Delta_0 + \frac{t(t-1)}{2!}\Delta_0^2 + \dots + \frac{t^{[k]}}{k!}\Delta_0^k + \dots$$
$$= y(0) + \underline{t}^{*\prime}\underline{\Delta}_0 \tag{2.4}$$

where $\underset{\sim}{t}_o^{*\prime} = (t, t(t-1), \ldots, t^{[k]}, \ldots)$,

$\underset{\sim}{\Delta}_o' = (\Delta_o, \frac{1}{2!}\Delta_o^2, \ldots.)$,

and $\Delta_o = y(h) - y(0)$.

Then $M = h \int_0^1 y(th)dt = h\left[y(0) + \mathcal{E}\underset{\sim}{t}_o^{*\prime}\underset{\sim}{\Delta}_o\right]$,

$$({}_1M_r - M) = h(\underset{\sim}{t}_o^* - \mathcal{E}\underset{\sim}{t}_o^*)'\underset{\sim}{\Delta}_o ,$$

$$({}_1M_r - M)^2 = h^2\underset{\sim}{\Delta}_o' T_o^* \underset{\sim}{\Delta}_o , \tag{2.5}$$

where $T_o^* = (\underset{\sim}{t}_o^* - \mathcal{E}\underset{\sim}{t}_o^*)(\underset{\sim}{t}_o^* - \mathcal{E}\underset{\sim}{t}_o^*)'$,

and $V({}_1M_r) = \mathcal{E}_t({}_1M_r - M)^2$

$$= h^2\underset{\sim}{\Delta}_o'\mathcal{E}T_o^*\underset{\sim}{\Delta}_o . \tag{2.6}$$

2.2.2 Expansion by Central Differences

It is possible to express the results of paragraph 2.2.1 in terms of central differences of interval h. That is, let

$$\Delta_c = y(\tfrac{3h}{2}) - y(\tfrac{h}{2}) ,$$

$$\underset{\sim}{\Delta}_c' = (\Delta_c, \tfrac{1}{2!}\Delta_c^2 , \ldots), \text{ etc.}$$

Some of the results of expansion in central differences are given in Section 1.4, but the expansion of the preceding paragraph is far more useful, and little use will be made of the central difference expansion.

2.2.3 Expansion by Taylor's Formula

If the function $y(\tau)$ is analytic, and therefore can be expanded in Taylors series, it is possible to express the estimator and its variance in terms of derivatives of $y(th)$ in a manner analagous to that of Paragraph 2.2.1. In fact, these results can be obtained by means of an identical development from the Taylor expansion, or by means of the transformation relationships developed in Section 1.4. The results are summarized here. In terms of expansion about the initial ordinate,

$$ {}_1M_r = h[y(0) + \underline{t}_o' \, H \, \underline{D}_o] \, , \tag{2.7}$$

where $\underline{t}_o' = (t, t^2, \ldots.)$

$$\underline{D}_o = [y^{(1)}(0), \; \frac{y^{(2)}(0)}{2!} \; , \ldots] \, ,$$

and $H = \begin{bmatrix} h & & & 0 \\ & h^2 & & \\ & & h^3 & \\ & 0 & & \cdot \\ & & & & \cdot \\ & & & & & \cdot \end{bmatrix}$.

Then,

$$({}_1M_r - M)^2 = h^2 \underline{D}_o' H T_o H \underline{D}_o \; , \tag{2.8}$$

where $T_o = (\underline{t}_o - \mathcal{E}\underline{t}_o)(\underline{t}_o - \mathcal{E}\underline{t}_o)'$,

$$t \sim U(0,1)$$

and

$$V({}_1M_r) \doteq \mathcal{E}_t({}_1M_r - M)^2$$

$$= h^2 \underline{D}_o' H \, \mathcal{E}T_o H \underline{D}_o \; . \tag{2.9}$$

This is also easily expressed in terms of expansion about the central ordinate, $y(\frac{h}{2})$.

$$ {}_1M_r = h \, [y(\tfrac{h}{2}) + \underline{t}_c' \, H \, \underline{D}_c] \tag{2.10}$$

where $\underline{t}_c' = [(t - \frac{1}{2}), (t - \frac{1}{2})^2, \ldots]$

and $\underline{D}_c' = \left[y^{(1)}(h/2), \frac{y^{(2)}(h/2)}{2!}, \ldots \right].$

Then $$({}_1M_r - M)^2 = h^2 \underline{D}_c' H T_c H \underline{D}_c \tag{2.11}$$

where $T_c = (\underline{t}_c - \mathcal{E}\underline{t}_c)(\underline{t}_c - \mathcal{E}\underline{t}_c)'$

and $$V({}_1M_r) = h^2 \underline{D}_c' H \, \mathcal{E}T_c H \underline{D}_c \ . \tag{2.12}$$

2.3 The Simple Random Sample, n > 1

2.3.1 Generalization of Section 2.2

Generalization of the results of Section 2.2 to the estimate of area under y from measurements at n random values of t, follows from the results of that section. Let n values of t be selected at random, $t \sim U(0,1)$, and let $y_i = y(wt_i)$, $w = nh$. Then

$$M_r = h \sum y_i \tag{2.13}$$

is an estimate of

$$M = \int_0^w y(t')dt' = w \int_0^1 y(wt)dt,$$

and $$\mathcal{E}M_r = h \int_0^1 \sum y_i f(t)dt = w \int_0^1 y(wt)dt$$

$$= M \ .$$

As in the previous section, M_r can be expressed in terms of a series expansion in differences,

$$M_r = h \sum \left[y(0) + t_i \underline{\Delta}_o + \ldots + \frac{t_i^{[k]}}{k!} \underline{\Delta}_o^k + \ldots \right] ,$$

$$= wy(0) + h \underline{\Delta}_o' \sum \underline{t}_{oi}^* . \qquad (2.14)$$

$$(M_r - M)^2 = \left[M_r - w(y(0) - \underline{\Delta}_o' \, \mathcal{E}\underline{t}_o^*) \right]^2$$

$$= h^2 \left[\underline{\Delta}_o' \left(\sum \underline{t}_{oi}^* - n\mathcal{E}\underline{t}_o^* \right) \right]^2$$

$$= h^2 \underline{\Delta}_o' \left[\sum (\underline{t}_{oi}^* - \mathcal{E}\underline{t}_o^*) \right] \left[\sum (\underline{t}_{oi}^* - \mathcal{E}\underline{t}_o^*) \right]' \underline{\Delta}_o .$$

$$V(M_r) = nh^2 \underline{\Delta}_o' \, \mathcal{E}T_o^* \underline{\Delta}_o , \qquad (2.15)$$

where $\underline{\Delta}_o = y(w) - y(0)$.

2.3.2 Central Differences and Derivatives

It is also possible to derive analagous expansions in terms of the central ordinate and in terms of derivatives. However, the only one of interest is,

$$M_r = wy(0) + h \sum \left(wt_i D_o + \frac{w^2 t_i^2 D_o^2}{2!} + \ldots \right) . \qquad (2.16)$$

$$(M_r - M)^2 = h^2 \underline{D}_o' H^* \left[\sum \underline{t}_{oi} - n\mathcal{E}\underline{t}_{oi} \right] \left[\sum \underline{t}_{oi} - n\mathcal{E}\underline{t}_{oi} \right]' H^* \underline{D}_o$$

$$= h^2 \underline{D}_o' H^* \left[\sum T_{oi} + \sum \sum (\underline{t} - \mathcal{E}\underline{t})_i (\underline{t} - \mathcal{E}\underline{t})_j' \right] H^* \underline{D}_o .$$

$$\mathcal{E}(M_r - M)^2 = h^2 \underline{D}_o' H^* \left[\sum \mathcal{E}T_{oi} + 0 \right] H^* \underline{D}_o .$$

$$V(M_r) = nh^2 \underline{D}_o' H^* \, \mathcal{E}T_o H^* \underline{D}_o , \qquad (2.17)$$

where $H^* = \begin{bmatrix} w & & & 0 \\ & w^2 & & \\ & & \ddots & \\ 0 & & & \cdot \end{bmatrix}$

2.4 The Stratified Random Sample, One Observation per Stratum

2.4.1 General Results

Let an interval of length w be divided into n sub-intervals of length h, and let n values of t be selected at random, $t_i \sim U(0,1)$, $\{t\} = (t_o, t_1, \ldots, t_{n-1})$.

$$\text{Define} \quad M_{st} = h\sum_{i=o}^{n-1} y(ih + ht_i) \tag{2.18}$$

$$= \sum ({}_1M_r)_i \quad .$$

$$\text{Then} \quad \mathcal{E}M_{st} = \sum M_i = M$$

$$\text{Now} \quad M_{st} = h\sum_{i=o}^{n-1} \left[y(ih) + t_i\Delta_i + \cdots + \frac{t_i^k\Delta_i^k}{k!} + \cdots \right]$$

$$= h\sum y(ih) + h\,\Sigma\,\underline{\Delta}_i'\underline{t}_{oi}$$

$$\text{And} \quad (M_{st} - M)^2 = h^2\left[\Sigma\underline{\Delta}_i'\,\underline{t}_{oi}^* - \mathcal{E}\,\Sigma\underline{\Delta}_i'\,\underline{t}_{oi}^*\right]^2$$

$$= h^2\left[\Sigma\underline{\Delta}_i'(\underline{t}_{oi}^* - \mathcal{E}\,\underline{t}_{oi}^*)\right]^2 \tag{2.19}$$

$$= h^2\left[\sum_{i=o}^{n-1}(\underline{\Delta}_i'T_{oi}^*\underline{\Delta}_i) + \sum_{j\neq i}\sum_{i=o}^{n-1}\underline{\Delta}_i'(\underline{t}_{oi}^* - \mathcal{E}\underline{t}_{oi}^*)(\underline{t}_{oj}^* - \mathcal{E}\underline{t}_{oj}^*)'\,\underline{\Delta}_j\right]$$

$$\text{Then } V(M_{st}) = h^2\sum\left[\underline{\Delta}_i'\,\mathcal{E}T_{oi}^*\underline{\Delta}_i\right] + 0 \tag{2.20}$$

$$= \sum_{i=o}^{n-1} V({}_1M_r)_i \quad .$$

2.4.2 Derivatives

Expression in terms of $\underline{D}_i$ follows directly from earlier results.

$$(M_{st} - M)^2 = h^2 \left[\sum \underline{D}_i' H(\underline{t}_i - \mathcal{E}\underline{t}_i) \right]^2 , \tag{2.21}$$

and

$$V(M_{st}) = h^2 \sum \left[\underline{D}_i' H \,\mathcal{E}\, T_o H \underline{D}_i \right] \tag{2.22}$$

$$= \sum_{i=0}^{n-1} V({}_1M_r)_i \quad .$$

2.5 Systematic Sample, with Random Start

2.5.1 General Results

Let an interval of length w be divided into n panels of length h, and let a value of t, $t \sim U(0,1)$, be selected. Then define,

$$M_{sr} = h \sum_{i=0}^{n-1} y(ih + ht) \quad . \tag{2.23}$$

Then,

$$\mathcal{E}\, M_{sr} = M \; .$$

Now $M_{sr} = h \sum \left[y(ih) + t\,\Delta_i + \ldots + \frac{t^{[k]}}{k!} \Delta_i^k + \ldots \right]$,

so that

$$(M_{sr} - M)^2 = h^2 \left[\Sigma\, \underline{\Delta}_i' (\underline{t}_o^* - \mathcal{E}\underline{t}_o^*) \right]^2$$

$$= h^2 \,(\Sigma\underline{\Delta}_i') T_o^* \,(\Sigma\underline{\Delta}_i) \;, \tag{2.24}$$

and

$$V(M_{sr}) = h^2 (\Sigma\underline{\Delta}_i')\, \mathcal{E} T_o^* (\Sigma\underline{\Delta}_i) \; . \tag{2.25}$$

2.5.2 Derivatives

Again, expression in terms of $\underline{D}_i$ follows directly from earlier results

$$(M_{sr} - M)^2 = h^2 \left[\Sigma\, \underline{D}_i' \, H(\underline{t}_o - \mathcal{E}\underline{t}_o) \right]^2 \tag{2.26}$$

$$V(M_{sr}) = h^2\ (\Sigma \underline{D}_i')\ H\ \mathcal{E} T_o H(\Sigma \underline{D}_i)\ . \tag{2.27}$$

2.6 The Systematic Centric Sample

2.6.1 General Results

Let an interval of length w be divided into n panels of length h, and define,

$$M_{sc} = h \sum_{i=o}^{n-1} y(ih + \frac{h}{2})\ . \tag{2.28}$$

This is, then, a special case of M_{sr}, with $t = \frac{1}{2}$, and the results can be written down directly from 2.5.

$$\begin{aligned}(M_{sc} - M)^2 &= \left[M_{sr}(\tfrac{1}{2}) - M\right]^2 \\ &= h^2\left[(\Sigma \underline{\Delta}_i')\ T_o^*\ (\Sigma \underline{\Delta}_i)\right]_{t = \frac{1}{2}} \\ &= h^2\left[\Sigma \underline{\Delta}_i'\ \underset{\sim}{C}^* \Sigma \underline{\Delta}_i\right]\ , \end{aligned} \tag{2.29}$$

$$\text{and}\quad V(M_{sc}) = (M_{sc} - M)^2\ .$$

2.6.2 Derivatives and Central Differences

It follows that $V(M_{sc})$ can also be written,

$$V(M_{sc}) = h^2 \left[\ (\Sigma \underline{D}_i') H T_o H (\Sigma \underline{D}_i)\right]_{t = \frac{1}{2}} \tag{2.30}$$

$$= h^2 \left[\ (\Sigma \underline{D}_{i+1/2}') H\ T_c H (\Sigma \underline{D}_{i+1/2})\right]_{t = \frac{1}{2}}\ , \tag{2.31}$$

$$\text{or,}\quad = h^2 \left[\ (\Sigma \underline{\Delta}_{i+1/2}') T_c^*\ (\Sigma \underline{\Delta}_{i+1/2})\right]_{t = \frac{1}{2}}\ ,$$

$$= h^2 \left[\ (\Sigma \underline{\Delta}_{i+1/2}') \underset{\sim}{C}^* (\Sigma \underline{\Delta}_{i+1/2})\ \right]\ . \tag{2.32}$$

2.7 Comparison of the Conventional Sampling Schemes

2.7.1 Introduction

A general comparison of the sampling schemes, based on the results of this chapter, is not possible. In fact, direct comparison is possible only for the case $\Delta^2 \equiv 0$; that is, when the function is a polynomial of first degree. When higher differences are non-zero, it is necessary to define classes of functions for which one or the other sampling scheme is best. This point is not investigated in depth, but rather only illustrated. For purpose of comparison, the contribution of measurement error to the variance of $\hat{M}$ is re-introduced.

2.7.2 Comparison when $\Delta^2 \equiv 0$

From the preceding sections and 1.4.7, the variances of the conventional estimators obtained by the four sampling schemes are,

$$V(M_r | \Delta^2 \equiv 0) = nh^2 \left[\frac{(n\Delta)^2}{12} + \sigma^2 \right]$$

$$= nh^2 \left[\frac{n^2}{12} (\Delta)^2 + \sigma^2 \right]$$

$$V(M_{st} | \Delta^2 \equiv 0) = nh^2 \left[\frac{(\Delta)^2}{12} + \sigma^2 \right]$$

$$V(M_{sr} | \Delta^2 \equiv 0) = nh^2 \left[\frac{n(\Delta)^2}{12} + \sigma^2 \right]$$

$$V(M_{sc} | \Delta^2 \equiv 0) = nh^2 \left[0 + \sigma^2 \right] .$$

Thus, it is seen that M_{sc} is best for all values of σ^2 and Δ , although the advantage decreases as σ^2 increasingly dominates $(\Delta)^2$. It is also seen that, of the three randomized schemes, the order of precision is M_{st}, M_{sr} and M_r and again the difference decreases as σ^2 increases.

2.7.3 Comparison when $\Delta^3 \equiv 0$.

To illustrate the comparison of the sampling schemes when $\Delta^2 \neq 0$, $\Delta^3 \equiv 0$, consider, for example, M_{st} and M_{sr}.

$$V(M_{st} \mid \Delta^3 \equiv 0) = h^2 \sum \left[\Delta_i, \tfrac{1}{2!}\Delta_i^2\right] \begin{bmatrix} \frac{1}{12} & 0 \\ 0 & \frac{1}{180} \end{bmatrix} \begin{bmatrix} \Delta_i \\ \frac{1}{2!}\Delta_i^2 \end{bmatrix} + nh^2\sigma^2$$

$$= h^2 \left\{ \frac{1}{12}\Sigma(\Delta_i)^2 + \frac{1}{720}\Sigma(\Delta_i^2)^2 + n\sigma^2 \right\}$$

$$= h^2 \left\{ \frac{1}{12}\Sigma(\Delta_i)^2 + \frac{n}{720}(\Delta^2)^2 + n\sigma^2 \right\},$$

and

$$V(M_{sr}|\Delta^3 \equiv 0) = h^2 \left\{\frac{1}{12}(\Sigma\Delta_i)^2 + \frac{1}{720}(\Sigma\Delta_i^2)^2 + n\sigma^2\right\}$$

$$= h^2 \left\{\frac{1}{12}(\Sigma\Delta_i)^2 + \frac{n^2}{720}(\Delta^2)^2 + n\sigma^2\right\}.$$

From Identity 1, paragraph 1.4.10, this becomes

$$V(M_{sr}|\Delta^3 \equiv 0) = h^2 \left\{ \frac{n}{12}\Sigma(\Delta_i)^2 - \frac{n^2}{720}(5n^2 - 6)(\Delta^2)^2 + n\sigma^2\right\},$$

Thus,

$$\frac{V(M_{st}|\Delta^3 \equiv 0)}{V(M_{sr}|\Delta^3 \equiv 0)} = \left(\frac{1}{n}\right)\left[\frac{\frac{1}{12}\Sigma(\Delta_i)^2 + \frac{n}{720}(\Delta^2)^2 + n\sigma^2}{\frac{1}{12}\Sigma(\Delta_i)^2 - \frac{n(5n^2-6)}{720}(\Delta^2)^2 + \sigma^2}\right] \quad (2.33)$$

$$> \left(\frac{1}{n}\right)\left[\frac{\frac{1}{12}\Sigma(\Delta_i)^2 + n\sigma^2}{\frac{1}{12}\Sigma(\Delta_i)^2 + \sigma^2}\right], \text{ for all } \Delta^2 \neq 0,\ n > 1.$$

It is seen, therefore, that the advantage of stratified random sampling over systematic random sampling is less when $\Delta^2 \neq 0$ than when $\Delta^2 = 0$. Further, for any given $\Sigma(\Delta)^2$ and Δ^2 (i.e., for any given polynomial of degree 2 and a specified interval), it is possible to solve (2.33) for a region, in n, within which systematic sampling is advantageous over stratified random sampling.

2.7.4 Summary

It would then seem that the space defined by all possible intervals of length w on power polynomials of arbitrary degree is separated into regions within which any one of the sampling schemes is better, equal to, or worse than any other of the sampling schemes considered, as regards precision of estimate. For any given polynomial and interval, it would seem possible to select the sampling scheme which is best, but it does not seem possible to satisfactorily describe the regions of superiority in general. However, it may be possible to define regions of superiority for classes of functions.

It will be shown later (Chapter 4) that when one is given a specific polynomial and interval, one is able to improve on any of the sampling plans treated in this chapter, so it seems unnecessary to pursue further the present line of investigation.

Chapter 3. ESTIMATION OF ERROR OF SYSTEMATIC SAMPLING

3.1 Introduction

Several estimators of error of estimate of the systematic sample have been proposed. In the present chapter, these are expanded and evaluated in the notational framework developed in the preceding chapter. In addition, a general method of constructing estimators of the variance is proposed, and several examples given. Attention is confined to the systematic samples, random and centric.

The model considered is, again,

$$y_i = y(\tau_i) + \epsilon_i \ , \tag{3.1}$$

where $y(\tau)$ is a continuous function of τ and $\mathcal{E}\epsilon = 0$, $\mathcal{E}\epsilon^2 = \sigma^2$ and $\mathcal{E}\epsilon_i\epsilon_j = 0$, $i \neq j$.

3.2 Quantities to be Estimated

3.2.1 General Expressions

$$V(M_{sr}|t) = \mathcal{E}_\epsilon(M_{sr} - M)^2$$

$$= h^2(\Sigma\underline{\Delta}_i') \; T_o^*(\Sigma\underline{\Delta}_i) + nh^2\sigma^2 \tag{3.2}$$

$$V(M_{sr}) = \mathcal{E}_t V(M_{sr}|t) = \mathcal{E}_t\mathcal{E}_\epsilon(M_{sr} - M)^2$$

$$= h^2(\Sigma\underline{\Delta}_i') \; \mathcal{E}T_o^*(\Sigma\underline{\Delta}_i) + nh^2\sigma^2 \tag{3.3}$$

$$V(M_{sc}) = (M_{sc} - M)^2\Big|_{t = \frac{1}{2}}$$

$$= h^2(\Sigma\underline{\Delta}_i') \; \underset{\sim}{C}^*(\Sigma\underline{\Delta}_i) + nh^2\sigma^2 \tag{3.4}$$

3.2.2 Special Cases

Let $\Delta^3 \equiv 0$.

$$V(M_{sr}|\Delta^3 \equiv 0;\, t) = h^2\left[\,\Sigma\Delta_i\,(t-\tfrac{1}{2}) + \frac{\Sigma\Delta_i^2}{2!}\,(t^{[2]}-\tfrac{1}{6})\right]^2 + nh^2\sigma^2, \qquad (3.5)$$

$$V(M_{sr}|\,\Delta^3 \equiv 0) \doteq h^2\left[\,\frac{1}{12}\,(\Sigma\Delta_i)^2 + \frac{1}{720}\,(\Sigma\Delta_i^2)^2\right] + nh^2\sigma^2$$

$$= nh^2\left\{\,\frac{1}{12}\sum_{i=0}^{n-1}(\Delta_i)^2 + \frac{n}{144}\left[\,\frac{1}{5}-(n^2-1)\right](\Delta^2)^2\right\} + nh^2\sigma^2, \qquad (3.6)$$

$$V(M_{sc}|\Delta^3 \equiv 0) = h^2\left[\,\frac{1}{576}\,(\Sigma\Delta_i^2)^2\right] + nh^2\sigma^2\,,$$

$$= nh^2\left[\frac{n}{576}\,(\Delta^2)^2\right] + nh^2\sigma^2\,. \qquad (3.7)$$

3.3 Standard Estimators of $V(M_{sr})$, with Extensions

3.3.1 Mean Square Successive Difference (Von Neumann, 1941)

$$\text{Define}\quad \delta^2 = \frac{1}{2(n-1)}\sum_{i=0}^{n-2}(y_{i+t+1} - y_{i+t})^2 \qquad (3.8)$$

By the expansion developed earlier,

$$\mathcal{E}_\epsilon\delta^2 = \frac{1}{2(n-1)}\sum_{i=0}^{n-2}\left\{(y_{i+1}-y_i) + t(\Delta_{i+1}-\Delta_i) + \frac{t^{[2]}}{2!}(\Delta_{i+1}^2-\Delta_i^2) + \dots\right\}^2 + \sigma^2$$

$$= \frac{1}{2(n-1)}\sum_{i=0}^{n-2}\left\{\Delta_i + t\Delta_i^2 + \dots + \frac{t^{[k]}}{k!}\,\Delta_i^{k+1} + \dots\right\}^2 + \sigma^2 \qquad (3.9)$$

$$\text{and}\quad \mathcal{E}_t\mathcal{E}_\epsilon\delta^2 = \frac{1}{2(n-1)}\sum_{i=0}^{n-2}\left\{\underline{\Delta}_i'\,\mathcal{E}_t\begin{bmatrix}1\\2t\\3t^{[2]}\\\vdots\end{bmatrix}\left[1,\,2t,\,3t^{[2]},\,\dots\right]\right\} + \sigma^2$$

$$= \frac{1}{2(n-1)}\sum\left\{\underline{\Delta}_i'\mathcal{E}_t T_o^a\,\underline{\Delta}_i\right\} + \sigma^2 \qquad (3.10)$$

where the $ij^{\underline{th}}$ element of $\mathcal{E}_t T_o^a$ is ,

$$\mathcal{E}_t t_{ij}^a = \mathcal{E}_t \; ij \sum_{k=1}^{i-1} \sum_{\ell=1}^{j-1} s(i+1,k)\; s(j-1,\ell)\; t^{\ell} t^{k} ,$$

$$= ij \sum \sum \frac{s(i-1,k)s(j-1,\ell)}{(k+\ell+1)} .$$

Thus,
$$\mathcal{E}(\delta^2|\Delta^3 \equiv 0) = \frac{1}{2(n-1)} \sum_{i=o}^{n-1} \left[\Delta_i, \tfrac{1}{2}\Delta_i^2\right] \begin{bmatrix} 1 & 1 \\ 1 & 4/3 \end{bmatrix} \begin{bmatrix} \Delta_i \\ \tfrac{1}{2}\Delta_i^2 \end{bmatrix} + \sigma^2$$

$$= \frac{1}{2(n-1)} \left\{ \sum_{i=o}^{n-2} (\Delta_i)^2 + \sum_{i=o}^{n-2} \Delta_i \Delta_i^2 + \frac{1}{3} \sum_{i=o}^{n-2} (\Delta_i^2)^2 \right\} + \sigma^2 , \quad (3.11)$$

$$= \frac{1}{2n} \left\{ \sum_{i=o}^{n-1} (\Delta_i)^2 + \frac{n(n-3)}{6} (\Delta^2)^2 \right\} + \sigma^2 , \quad (3.12)$$

and
$$\mathcal{E}(\delta^2|\Delta^2 \equiv 0) = \frac{1}{2n} \sum_{i=o}^{n-1} (\Delta_i)^2 + \sigma^2$$

$$= \frac{(\Delta)^2}{2} + \sigma^2 . \quad (3.13)$$

From (3.6) and (3.13) it is seen that δ^2 is an unbiased estimator of $V(M_{sr})$, letting $\Delta^2 \equiv 0$, only if

$$\frac{1}{12} = \frac{1}{2n} . \quad (3.14)$$

This follows since, when $\Delta^2 \equiv 0$,

$$\sum_{i=o}^{n-2} (\Delta_i)^2 = (n-1)(\Delta)^2 ,$$

and
$$\left(\sum_{i=o}^{n-1} \Delta_i\right)^2 = n^2 (\Delta)^2 .$$

Thus, $nh^2\delta^2$ is unbiased only when $n = 6$, and is negatively biased when $n > 6$, given that $\Delta^2 \equiv 0$.

3.3.2 The Mean Square Successive Difference of Lag ℓ

The mean square successive difference of lag ℓ, ℓ an integer, is defined as

$$\delta_\ell^2 = \frac{1}{2(n-\ell)} \sum_{i=0}^{n-\ell-1} (y_{i+t+\ell} - y_{i+t})^2 .$$

$$\mathcal{E}\delta_\ell^2 = \frac{1}{2(n-\ell)} \sum \left[\Delta_i + t\Delta_i^2 + \dots \right]^2 + \sigma^2 . \qquad (3.15)$$

Let $\Delta^3 \equiv 0$, and the expression (3.15) becomes

$$\mathcal{E}(\delta_\ell^2 | \Delta^3 \equiv 0) = \frac{1}{2(n-\ell)} \sum \left[(\Delta_i)^2 + \Delta_i \Delta_i^2 + \frac{1}{3} (\Delta_i^2)^2 \right] + \sigma^2 , \qquad (3.16)$$

and $\mathcal{E}(\delta_\ell^2 | \Delta^2 \equiv 0) = \frac{1}{2(n-\ell)} \Sigma(\Delta_i)^2 + \sigma^2$

$$= \frac{\ell^2}{2} (\Delta)^2 + \sigma^2 . \qquad (3.17)$$

Example: Let $n = 24$, and $\Delta^2 \equiv 0$. Then set $\frac{n}{12} = \frac{\ell^2}{2}$ and solve, $\ell = \sqrt{24/6} = 2$. Thus, $nh^2\delta_2^2$ is an unbiased estimator of $V(M_{sr} |\ n = 24;\ \Delta^2 \equiv 0)$.

It follows that the expectation of the average of k such estimators, δ_ℓ^2, with different ℓ and given $\Delta^2 \equiv 0$, will be of the form,

$$\mathcal{E}(\delta^2_{\bar{\ell}}) = \frac{1}{2k} \sum_{i=1}^{k} \ell_i^2 (\Delta)^2 + \sigma^2 . \qquad (3.18)$$

Thus an entire family of unbiased estimators of $V(M_{sr} | \Delta^2 \equiv 0)$ is defined by (3.18), subject to the restriction,

$$n = \frac{6}{k} \sum_{i=1}^{k} \ell_i^2 .$$

More generally, consider an estimator of the form,

$$\delta_*^2 = a_1\delta_{\ell 1}^2 + a_2\delta_{\ell 2}^2 \tag{3.19}$$

where $a_1 + a_2 = 1$

and where $\mathcal{E}(\delta_{\ell 1}^2|\Delta^2 \equiv 0) = k_1(\Delta)^2 + \sigma^2$

$$\mathcal{E}(\delta_{\ell 2}^2|\Delta^2 \equiv 0) = k_2(\Delta)^2 + \sigma^2$$

Then, it is possible to choose a_1 and a_2 such that

$$\mathcal{E}(\delta_*^2|\Delta^2 \equiv 0) = \alpha(\Delta)^2 + \sigma^2$$

Example:

Let $\delta_*^2 = a_1\delta_1^2 + a_2\delta_2^2$

from (3.17), $k_1 = \frac{1}{2}$; $k_2 = 2$

and from (3.6), $\alpha = n/12$.

Thus, choose $$a_2 = \frac{\alpha - k_1}{k_2 - k_1} = \frac{1}{6}\left(\frac{n-6}{3}\right) = \frac{n-6}{18} \tag{3.20}$$

$$a_1 = \frac{1}{6}\left(\frac{n-24}{-3}\right) = \frac{24-n}{18}$$

$$\delta_*^2 = \left(\frac{n-6}{18}\right)\delta_2^2 + \frac{24-n}{18}\,\delta_1^2$$

$$\mathcal{E}(\delta_*^2) = \frac{n}{12}(\Delta)^2 + \sigma^2 .$$

3.3.3 The Squared Mean Alternate Difference

The squared mean alternate difference was defined by Yates (1948) as:

$$E = \frac{1}{n}\,(y_t - y_{1+t} + y_{2+t} - \dots - y_{n-1+t})^2 , \tag{3.21}$$

where n is an even integer.

Thus, $E = \frac{1}{n}\left\{-\sum_{i=o}^{n-2} \Delta_{i+t} x_i\right\}^2$, $x_i = \begin{cases}1, & i = 0, 2, \ldots \\ 0, & i = 1, 3, \ldots\end{cases}$

$$\mathcal{E}_{\epsilon} E = \frac{1}{n}\left\{\sum_{i=o}^{n-2} x_i(\Delta_i + t\Delta_i^2 + \ldots\ )\right\}^2 + \sigma^2$$

$$\mathcal{E}(E|\Delta^3 \equiv 0) = \frac{1}{n}\left\{(\sum_{i=o}^{n-2} x_i\Delta_i)^2 + (\sum_{i=o}^{n-2} x_i\Delta_i)(\sum_{i=o}^{n-2} x_i\Delta_i^2) + \frac{1}{3}(\sum_{i=o}^{n-2} x_i\Delta_i^2)^2\right\}$$
$$+ \sigma^2$$

$$= \frac{1}{4}\left\{\sum_{i=o}^{n-1}(\Delta_i)^2 - \frac{n}{12}(n^2 - 2)(\Delta^2)^2\right\} + \sigma^2 , \qquad (3.22)$$

and $$\mathcal{E}(E|\Delta^2 \equiv 0) = \frac{1}{n}\left\{\Sigma x_i\Delta_i\right\}^2 + \sigma^2$$
$$= \frac{n}{4}(\Delta)^2 + \sigma^2 . \qquad (3.23)$$

Hence, $$nh^2\ \mathcal{E}(E|\Delta^2 \equiv 0) = nh^2\left[\frac{n}{4}(\Delta)^2 + \sigma^2\right]$$
$$= V(M_{sr}|\ \Delta^2 \equiv 0) + \frac{n^2h^2(\Delta)^2}{6} ,$$

and E is positively biased for all n if $\Delta^2 \equiv 0$.

3.3.4 Variations of E

Two variations of E are suggested. The first is modified from Yates (1949),

$$E_k = \frac{1}{km}\sum_{i=o}^{k-1}(y_{o+im+t} - y_{1+im+t} + \ldots - y_{(i+1)m+t-1})^2 , \qquad (3.24)$$

where km = n.

Thus E_k is the average of the k squared quantities, defined by application of E to the k successive sets of m = n/k observations.

For example, if $n = 8$,

$$E_2 = \frac{1}{n}\left\{\left[y_t - y_{1+t} + y_{2+t} - y_{3+t}\right]^2 + \left[y_{4+t} - y_{5+t} + y_{6+t} - y_{7+t}\right]^2\right\} .$$

It is seen that a limiting form of E_k, $E_{\frac{n}{2}}$, is the mean square alternate difference

$$\delta_a^2 = \frac{1}{n}\sum_i (\Delta_{i+t})^2 \tag{3.25}$$

$$\mathcal{E}(\delta_a^2|\Delta^2 \equiv 0) = \frac{1}{2}(\Delta)^2 + \sigma^2$$

and, in general,

$$\mathcal{E}(E_k|\Delta^2 \equiv 0) = \frac{m}{4}(\Delta)^2 + \sigma^2 . \tag{3.26}$$

A second variation of E is the squared mean successive difference,

$$E_s = \frac{1}{2}\left(\sum_{i=0}^{n-2} \Delta_{i+t}\right)^2 \tag{3.27}$$

$$\mathcal{E}(E_s|\Delta^2 \equiv 0) = \frac{(n-1)^2}{2}(\Delta)^2 + \sigma^2 \tag{3.28}$$

and it is seen that E_s is a limiting form of δ_ℓ^2 , $\ell = n - 1$, since

$$\sum_{i=0}^{n-2} \Delta_{i+t} = (y_{n-1+t} - y_t) .$$

3.3.5 Cochran's Quadratic Estimators

Cochran (1953, p. 180) proposes an estimator based on an average of squared second differences. Following is a modification of Cochran's estimator,

$$\delta_c^2 = \frac{1}{6(n-2)}\sum_{i=0}^{n-3} (y_{i+t} - 2y_{i+1+t} + y_{i+2+t})^2, \tag{3.29}$$

$$\mathcal{E}(\delta_c^2) = \frac{1}{6(n-2)} \sum_{i=0}^{n-3} \left\{ \Delta_i^2 + t\Delta_i^3 + \ldots \right\}^2 + \sigma^2 . \qquad (3.30)$$

Then $$\mathcal{E}(\delta_c^2 | \Delta^2 \equiv 0) = \sigma^2 , \qquad (3.31)$$

And $$\mathcal{E}(\delta_c^2 | \Delta^3 \equiv 0) = \frac{1}{6(n-2)} \sum_{i=0}^{n-3} (\Delta_i^2)^2 + \sigma^2$$

$$= \frac{1}{6}(\Delta^2)^2 + \sigma^2 . \qquad (3.32)$$

It is immediately apparent that the coefficient of $(\Delta^2)^2$ in (3.32) can be modified by manipulation of the lag, as in paragraph 3.3.2, or by squaring sums of second differences as in 3.3.4, so that it is possible to define a family of estimators based on squared second differences. For example, let

$$E_c = \frac{1}{4}(\Sigma\Delta_{i+t}^2)^2 \qquad (3.33)$$

$$\mathcal{E}(E_c | \Delta^2 \equiv 0) = \sigma^2$$

$$\mathcal{E}(E_c | \Delta^3 \equiv 0) = \frac{(n-2)^2}{4} (\Delta^2)^2 + \sigma^2 . \qquad (3.34)$$

3.3.6 Summary

All of the conventional estimators are related and can be considered to lie in the same general family of estimators, generated by linear combinations of squares of linear combinations of sample differences. An unbiased estimator of $V(M_{sr} | \Delta^2 \equiv 0)$ is developed for all n, but a more general unbiased estimator of $V(M_{sr} | \Delta^n \equiv 0)$ was not obtained.

A general comparison of δ^2 and E is of interest. In particular, it was noted that the bias of E is always greater than 0 if $\Delta^2 \equiv 0$, so that there exist situations (e.g., $n = 6$, $\Delta^2 \equiv 0$) in which δ^2 is "better" than E. Yates (1948) demonstrated that when applied to the autoregressive function, E was consistently smaller than δ^2, but larger than V. It is concluded that neither has a consistent superiority. (It should be noted that in Yates' study, E was modified by an end correction.)

However, the question of superiority among the standard variance estimators is of little importance. Direct comparison can be made only in the case of specified polynomials, for many of which (if not all) unbiased estimates of variance can be constructed. Even so, in most such cases it will be better to modify the estimates of M to take advantage of the specified model, than to utilize this knowledge in improving the estimates of error. For this reason, this line of investigation will not be continued, although it is of considerable interest.

Chapter 4. QUADRATURE DESIGNS

The term "quadrature designs" is used to designate sampling designs that yield the exact area (definite integral) of polynomials of specified degree, if measurement error is zero. The term is derived from the "quadrature formulae", appropriate to numerical integration of the polynomials, given the observation made in accordance with the designs.

4.1 Introduction

4.1.1 The Form of Quadratures

Application of quadrature, or numerical integration, methods to the general continuous function is dependent on the power polynomial approximation to that function, as were the results of earlier chapters. As before, this approximation is justified by Weierstrass' Theorem. In the present chapter, attention is given to improved estimates of area under the curve. Reference is made to Kunz(1954,Ch.7), Kopal (1955,Ch.7), Whitaker and Robinson(1944,Ch.7) and Daniell(1940).

According to Weierstrass' Theorem (see 1.2), the continuous function $y(\tau)$ can be represented,

$$y(\tau) = P_{m'}(\tau) + r(\tau) \tag{4.1}$$

where $P_{m'}(\tau)$ is a power polynomial in τ of degree m'. Thus,

$$M = \int_a^b P_{m'}(\tau)d\tau + r , \tag{4.2}$$

where r can be made arbitrarily small by appropriate choice of $P_{m'}$.

Throughout this chapter, it is assumed that $P_{m'}$ is so chosen that r can be ignored. That is, $y(\tau)$ is treated as a power polynomial of arbitrary degree.

Quadrature formulae are of the general form,

$$Q_m = \sum W_i y(t_i) \ , \tag{4.3}$$

where the W_i and t_i are selected so as to make

$$Q_m = \int_a^b P_m(\tau) d\tau \ , \tag{4.4}$$

for some $m \leq m'$.

4.1.2 Types of Quadratures

A great many specific Quadrature formulae have been described, of which several are of particular interest and will be considered here.

1. Let all of the 2n constants in (4.3) be utilized in specifying the definite integral of $P_m(\tau)$. These are called Gaussian Formulae.

2. Let the t_i be selected for convenience, and the W_i utilized in specifying the definite integral of $P_m(\tau)$. A special class of formulae in which the intervals between the t_i are equal will be considered here. These include the Newton-Cotes Formulae and the Centric Formulae.

3. Let the W_i be selected for convenience, and the t_i utilized in specifying the definite integral of $P_m(\tau)$. When the W_i are equal, the class of formulae is named after Tchebycheff.

4.1.3 Some General Results

When the n sample values of τ are arbitrary, say $\{t_1, t_2, \ldots, t_n\}$, then, by Lagrange's interpolation formula (Kunz, 1957, p. 89),

$$P_{n-1}(\tau) = \sum_{i=1}^{n} a_i(\tau)\, y(t_i) \tag{4.5}$$

where, if $a_s(\tau)$ is a particular value of $a_i(\tau)$,

$$a_s(\tau) = \frac{(\tau - t_1)(\tau - t_2)\ldots(\tau - t_n)}{(t_s - t_1)\ldots(t_s - t_{s-1})(t_s - t_{s+1})\ldots(t_s - t_n)(\tau - t_s)} \tag{4.6}$$

From (4.5),

$$Q_n = \int_a^b P_{n-1}(\tau)d\tau = \sum_{i=1}^{n} W_i y(t_i) \tag{4.7}$$

where

$$W_i = \int_a^b a_i(\tau)d\tau \ . \tag{4.8}$$

Further,

$$M = Q + R$$

$$\text{where } R = \int_a^b \frac{(\tau - t_1)(\tau - t_2)\ldots(\tau - t_n)}{n!}\, y^{(n)}(\xi)d\tau \tag{4.9}$$

4.2 The Newton-Cotes Quadratures

4.2.1 The Weights and Their Derivation

Let it be desired to integrate $P_{n-1}(\tau)$ over the range $(0,w)$, where $w/n-1 = h$, and let $\{t_i\} = \{0, h, 2h, \ldots, w\}$. Numerical integration formulae for this case will be designated, Q_n^N .

Then, let

$$u = \frac{\tau - t_o}{h} \quad ,$$

so that (4.5) can be rewritten,

$$P_{n-1}(u) = \sum_{i=o}^{n-1} a_i(u)y(t_i) \tag{4.10}$$

$$\text{where} \quad a_s(u) = \frac{(-1)^{n-1-s}u^{[n]}}{s!(n-1-s)!(u-s)} \quad . \tag{4.11}$$

$$\text{Thus,} \quad Q_n^N = \int_o^W P_{n-1}(\tau)d\tau = h\int_o^{n-1} P_{n-1}(u)du$$

$$= \sum_{i=o}^{n-1} W_i y(t_i) \quad , \tag{4.12}$$

$$\text{where } W_s = h\int_o^{n-1} a_s(u)du \; . \tag{4.13}$$

Similarly,

$$R_n = h^n \int_o^{n-1} \frac{u^{[n]}}{n!} \; y^{(n)}(\xi)du \; . \tag{4.14}$$

(For a general treatment of remainder terms, see Section 4.7)

A number of formulae belonging to this group are in common use, including the Trapezoidal Rule, Simpsons Rule and the Three-Eighths Rule. These are all defined by the basic formula

$$Q_n^N = \frac{n}{C}\sum_{i=o}^{n-1} A_i y(t_i) \text{ , where } A_i = A_{n-i+1} \text{ ,}$$

with appropriate coefficients of the formulae up to $n = 7$ given in Table 4.1. A more complete table (to $n = 21$) is given by Kopal (1955, p. 536), in a slightly different form.

Table 4.1 Coefficients of $y(t_i)$ in Newton-Cotes Quadrature Formulae

n	A_0	A_1	A_2	A_3	C
2	1				2
3	1	4			3
4	3	9			8
5	14	64	24		45
6	95	375	250		288
7	41	216	27	272	140

4.2.2 Use of Two or More Conventional Newton-Cotes Formulae in Combination

It is common in practice to construct quadrature formulae by linear combination of other formulae. Of particular interest is the practice of constructing formulae for a large number of ordinates by successive application of simple formulae to sections of the range. For example, if $n = 21$, it is possible to partition the 20 panels into, (i) ten sets of two panels each, to each set of which is applied Simpsons rule, (ii) five sets of four panels each, to each set of which is applied the rule for $n = 5$, or, (iii) four sets of three panels each, and four sets of two panels each, to each set of which is applied the appropriate formula.

The remainder term of such a constructed formula is constructed according to Property 7, paragraph 4.7.1. That is, linear combinations of Newton-Cotes Formulae of degree m have remainders of degree m with numerical coefficients which are linear combinations of the numerical coefficients of the remainders of the component formulae.

4.3 Gaussian Quadratures

The Gaussian Formulae are the most general and powerful of the quadrature formulae, as all 2n constants on the right side of (4.3) are selected so as to specify the definite integral of $P_m(\tau)$. With 2n constants, it is possible to completely specify the definite integral of a polynomial of degree 2n - 1. A sketch of one approach to the derivation of these formulae follows. (See Kopal, 1955, p. 352 and Whitaker and Robinson, 1944, p. 159.)

Given $P_{2n-1}(\tau)$ defined over the interval $[0,w]$ and n distinct arbitrary values $\{t_i\}$ in the interval, it is possible to uniquely specify a polynomial, $F_{n-1}(\tau)$, such that $F_{n-1}(t_i) = P_{2n-1}(t_i)$ for all $t_i \in \{t_i\}$. Further, $P_{2n-1}(\tau)$ can be written,

$$P_{2n-1}(\tau) = F_{n-1}(\tau) + (\tau - t_1)(\tau - t_2)\ldots(\tau - t_n)G_{n-1}(\tau) \ . \tag{4.15}$$

Hence,

$$\int_0^w P_{2n-1}(\tau)d\tau = \int_0^w F_{n-1}(\tau)d\tau \ ,$$

<u>iff</u>

$$\int_0^w (\tau - t_1)(\tau - t_2)\ldots(\tau - t_n)G_{n-1}(\tau)d\tau = 0. \tag{4.16}$$

It turns out that the $\{t_i\}$ satisfying (4.16) are the roots of the Legendre polynomial of degree n, suitably transformed from the range (-1, 1) into the range (0,w). Given the values of $\{t_i\}$, the coefficients, W_i, are obtained by formula (4.8).

Kopal, (1955, p. 523) gives Gaussian Quadrature Formulae for n = 2 to n = 16. The coefficients and appropriate $\{t_i\}$ for the first few are given below, in somewhat different form.

Table 4.2 Coefficients of $y(t_i)$ and values of $\{t_i\}$, Gaussian Quadrature Formulae, $n = 2, \ldots, 6$.

n	A_1	A_2	A_3	t_1/w	t_2/w	t_3/w
2	1/2			$\frac{1}{2}(1 - \frac{1}{3})$		
3	5/18	4/9		$\frac{1}{2}(1 - \frac{3}{5})$	.50000	
4	.17393	.32607		.06943	.33001	
5	.11846	.23931	.28444	.04691	.23077	.50000
6	.08566	.18038	.23396	.03376	.16940	.38069

Then $$Q_n^G = w \sum_{i=1}^{n} A_i y(t_i) \ , \tag{4.17}$$

where $A_i = A_{n-i+1}$

and $t_i = w - t_{n-i+1}$

4.4 Tchebycheff Quadratures

The third conventional type of quadrature of particular interest is that named after Tchebycheff. In this family of formulae, it is specified that all weight coefficients be equal, and that the $\{t_i\}$ be selected so as to specify the definite integral of a polynomial of the highest possible degree. Reference is made to Kopal (1955, p. 417).

Salient features are:

i. Degree of the formulae is n, by definition. (Actually, degree is (n) or (n+1), whichever is odd.) Only n-1 of the weight coefficients are arbitrary (equal to the other), leaving n+1 constants on the right side of (4.3) to be utilized in specifying the definite integral of interest.

ii. Tchebycheff Quadratures exist only for $n = 2,\ldots,7,9$. For $n > 9$, no integral value of n exists for which all roots of the Tchebycheff polynomials are real.

Thus, Tchebycheff Quadratures can be completely summarized in a brief Table, modified from Kopal.

Table 4.3 Values of $\{t_i\}$ for Tchebycheff Quadrature Formulae.

n	t_1/w	t_2/w	t_3/w	t_4/w	t_5/w
2	.21132				
3	.14645	.50000			
4	.10267	.40620			
5	.08375	.31273	.50000		
6	.06688	.28874	.36668		
7	.05807	.23517	.33804	.50000	
9	.04421	.19949	.23562	.41605	.50000

Then, $$Q_n^T = \frac{w}{n}\sum_{i=1}^{n} y(t_i) \tag{4.18}$$

and $t_i = w - t_{n-i+1}$.

4.5 Centric Quadratures

In several of the texts studied, e.g., Whitaker and Robinson(1944, p. 155) mention is made of Quadrature Formulae based on ordinates in the center of the panels of integration, where panels are of equal length, as in the Newton-Cotes Quadratures. However, these formulae were not found in the literature, nor was a name associated with them.

As the formulae are directly applicable to the observations obtained from a Centric Systematic Sample, Section 2.6, they are of particular interest here, and fortunately can be derived through use of methodology already established. The name used refers to the sampling scheme to which the formulae apply.

From (4.3) and (4.4), let

$$Q_n^C = \sum_{i=1}^{n} W_i y(t_i) = \int_0^w P_n(\tau)d\tau \tag{4.19}$$

where $t_i = \dfrac{(2i - 1)h}{2}$,

$h = w/n$.

Thus, from 4.1.3,

$$W_s = \int_0^w a_s(\tau)d\tau \ , \tag{4.20}$$

where $a_s(\tau)$ is defined by (4.6).

Then, let $u = (\dfrac{\tau - t_1}{h})$, so that as before,

$$a_s(u) = \frac{u^{[n]}(-1)^{n-s}}{(s-1)!(n-s)!(u-s+1)} \tag{4.21}$$

and, $W_s = h \int_{-1/2}^{n-1/2} a_s(u)du$

$$= \frac{(-1)^{n-s}\, h}{(s-1)!(n-s)!} \int \frac{u^{[n]}}{(u-s+1)}\, du \tag{4.22}$$

$$= h\underline{b}'(n,s)\underline{u} \ , \tag{4.23}$$

where $a_s(u) \doteq \underline{b}'(n,s) \begin{bmatrix} 1 \\ u \\ u_2 \\ \vdots \\ u^{n-1} \end{bmatrix}$, (4.24)

and
$$\underline{u} = \int_{-1/2}^{n-1/2} \begin{bmatrix} 1 \\ u \\ u^2 \\ \vdots \\ u^{n-1} \end{bmatrix} du = \begin{bmatrix} (n - 1/2) - \frac{1}{2} \\ \frac{1}{2}(n - \frac{1}{2})^2 - \frac{1}{2}(- \frac{1}{2})^2 \\ \vdots \\ \frac{1}{n}(n - \frac{1}{2})^n - \frac{1}{n}(- \frac{1}{2})^n \end{bmatrix} \tag{4.25}$$

It is seen, then, that $\underline{b}(n,s)$ is the vector of coefficients obtained by the algebraic expansion of $\frac{u^{[n]}}{(u-s+1)}$, multiplied by $\frac{(-1)^{n-s}}{(s-1)!(n-s)!}$.

Further, one can write,

$$B'(n) = \begin{bmatrix} \underline{b}'(n,1) \\ \underline{b}'(n,2) \\ \vdots \\ \underline{b}'(n,n) \end{bmatrix} \tag{4.26}$$

so that the vector of coefficients for the Centric Quadrature Formula can be expressed,

$$\underline{W} \doteq hB'(n)\ \underline{u}\ . \tag{4.27}$$

Expression (4.27) is particularly helpful, as B'(n) is constant for all formulae of n equally spaced abscissae, and in order to derive constants for a particular range of integration, it is necessary only to evaluate $\underline{u}$ and the product of B' and $\underline{u}$. A Table of Centric Quadrature formulae follows. Numerical values of $B'(n)$, $n = 2, 3, \ldots, 8$, are given in the Appendix.

Table 4.4 Coefficients of $y(t_i)$ for Centric Quadrature Formulae.

n	A_1	A_2	A_3	A_4	C
2	1				1
3	3	2			3/8
4	13	11			1/12
5	275	100	402		5/1152
6	2223	1251	2286		1/1920
7	1.25300	.04466	2.83600	-1.26733	1
(7)	(173,215)	(6174)	(392,049)	(-175,196)	(1/138,240)
8	1.22200	.29485	1.95011	.53304	1

$$Q_n^c = hC \sum_{i=1}^{n} A_i y(t_i) , \qquad A_i = A_{n-i+1}$$

4.6 Quadrature Formulae in Systematic Sampling with Random Start

4.6.1 General Development

As in Section 2.5 select a value of t at random, *i.e.*, $t \sim U(0,1)$, and let $\{t_i\} = \{th, th + h, \ldots, th + h(n - 1)\}$. Following the results of Section 4.5, it is possible to write,

$$Q_n(t) = \int_0^W P_{n-1}(\tau)d\tau = \sum W_i(t)y(t_i) \tag{4.28}$$

where $\quad W_s(t) = h\int_{-t}^{n-t} a_s(u)du$.

Hence, from (4.27)

$$\underline{W}(t) = hB'(n)\underline{u}(t) \tag{4.29}$$

where $B'(n)$ is defined as before, (4.26), and

$$\underline{u}'(t) = \int_{-t}^{n-t} \left[1,\ u,\ u^2,\ \ldots,\ u^{n-1}\right] du. \tag{4.30}$$

Then let

$$R_n(t) = \int_0^W P_n(\tau)d\tau - Q_n(t) .$$

Therefore, it is possible to apply numerical integration methods to data collected by systematic sample with random start. Whether this is in general an advantageous scheme cannot be here stated, as an attempt to evaluate $R(t)$ was unsuccessful. It is shown in Paragraph 4.6.4 that in a particular case, $\mathcal{E}R(t)$ is not identically zero, so it can be stated that the procedure does not provide unbiased estimators, in general.

<u>4.6.2 The Special Case, n = 2</u>

From (4.29), letting $h = 1$,

$$\underline{W}(t) = B(2)\underline{u}(t)$$

where $\underline{u}'(t) = \int_{-t}^{2-t} [1,u]\, du,$

$$\underline{u}(t) = \begin{bmatrix} 2 \\ 2 - 2t \end{bmatrix},$$

and $B'(2) = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$.

Thus,

$$\underline{W}(t) = \begin{bmatrix} 2t \\ 2 - 2t \end{bmatrix} .$$

4.6.3 The Case, $n > 2$

Let $\Delta^2 \equiv 0$, $h = 1$, $n > 2$, and consider systematic sampling with a random start. The integral over the range $(0, n)$ can be separated into three integrals, $(0, t + 1/2)$, $(t + 1/2, n - \frac{3}{2} + t)$, $(n - \frac{3}{2} + t, n)$. The middle of these can, under the assumption $\Delta^2 \equiv 0$, be represented by the Centric Quadrature formula of degree 1,

$$Q^C = \sum_{i=2}^{n-1} y(t_i) ,$$

which is simply a sum of formulae of the type, Q_1^c .

Also, a quadrature formula of degree 1 for the sum of the two end integrals, in terms of the two end observations, can be constructed by the method outlined in paragraph 4.1.3.

$$W_1 = \frac{1}{t_1 - t_n} \left\{ \int_0^a (\tau - t_n) d\tau + \int_b^n (\tau - t_n) d\tau \right\} ,$$

$$W_n = \frac{1}{t_n - t_1} \left\{ \int_0^a (\tau - t_1) d\tau + \int_b^n (\tau - t_1) d\tau \right\} ,$$

where $a = t + 1/2$

$b = n - 3/2 + t$

$t_1 = t$

$t_n = n - 1 + t$.

Thus,

$$\underline{W}(t) = \begin{bmatrix} 1 + \frac{n(t - 1/2)}{n - 1} \\ 1 \\ \vdots \\ 1 \\ 1 - \frac{n(t - 1/2)}{n - 1} \end{bmatrix} \qquad (4.31)$$

is a set of weights for a quadrature formula of degree 1 when t is selected at random. This will be recognized as the continuous analogue to Cochran's (1953, p. 172) end corrections. Also, the result of Paragraph 4.62 is obtained by setting $n = 2$.

It will be noted that other "end corrections" can be devised as quadrature formulae of degree 1 by the simple expedient of modifying the segments of the integral. Also, it will be noted that one can, in general, construct "end corrections" as quadrature formulae of higher degree, and that the process is a direct application of the general process of construction of quadrature formulae for systematic samples with a random start as developed in Paragraph 4.6.1.

4.6.4 Properties of $Q_1(t)$

Some properties of $\mathcal{E}_t[Q_1(t) - M]$, $Q_1(t)$ defined as in paragraph 4.6.3 are of interest.

$$\mathcal{E}_t\Big[Q_1(t) = \sum \Big[\mathcal{E}_{t_i}(W_i y(t_i)) - M_i\Big]$$

$$\text{where } \Big[\mathcal{E}_{t_i}(W_i y(t_i)) - M_i\Big] = 0, \quad i = 2, 3, \ldots, n - 1,$$
$$\neq 0, \quad i = 1, n.$$

Hence,

$$\mathcal{E}_t\left[Q_1(t) - M\right] = \left(\frac{n}{n-1}\right)\left\{\int_0^1 (t_1 - 1/2)y(t_1)dt_1 - \int_{n-1}^{n} (t_n - 1/2)y(t_n)dt_n\right\} . \tag{4.32}$$

An interesting interpretation of (4.32) is,

$$\mathcal{E}_t\left[Q_1(t) - M\right] = \left(\frac{n}{n-1}\right)\left\{ M_1\left[\phi_1(t) - \frac{1}{2}\right] - M_n\left[\phi_n(t) - \frac{1}{2}\right]\right\} , \tag{4.33}$$

where $\phi_1(t)$ is the mean of a random variable t_1, $0 < t_1 < 1$, with density function $y(t_1)/M_1$, and $\phi_n(t)$ is the mean of a random variable t_n, $n - 1 < t_n < n$, with density function $y(t_n)/M_n$. It is seen by inspection that $\mathcal{E}_t[Q_1(t) - M] = 0$ only for special cases, and that, in general, $Q_1(t)$ is a biased estimator. Another useful expression for the bias of $Q_1(t)$ follows the methods of Chapter 2, where again $t \sim U(0,1)$,

$$\mathcal{E}_t\left[Q_1(t) - M\right] = \left(\frac{n}{n-1}\right)\left[\underline{\Delta}_0' - \underline{\Delta}_{n-1}'\right]\mathcal{E}_t\begin{bmatrix} t(t - 1/2) \\ t(t-1)(t - 1/2) \\ \vdots \\ t^{[k]}(t - 1/2) \\ \vdots \end{bmatrix} \tag{4.34}$$

$$= \left(\frac{n}{n-1}\right)\left[\sum_{i=0}^{n-2}\Delta_i^2, \frac{1}{2!}\sum_{i=0}^{n-2}\Delta_i^3, \frac{1}{3!}\sum_{i=0}^{n-2}\Delta_i^4 \ldots\right]A\begin{bmatrix}(\frac{1}{4} - \frac{1}{3}) \\ (\frac{1}{6} - \frac{1}{4}) \\ (\frac{1}{8} - \frac{1}{5}) \\ \vdots\end{bmatrix}$$

$$\mathcal{E}\left[Q_1(t) - M \mid \Delta^4 \equiv 0\right] = -\left(\frac{n}{n-1}\right)\frac{\sum_{i=0}^{n-2}\Delta_i^2}{12} = \frac{-1}{12}\sum_{i=0}^{n-1}\Delta_i^2 + \frac{n}{24}\Delta^3$$

In comparison, using for estimator the quadrature formula of degree 1 originally designated as M_{sc}, the bias of the centric systematic sample is, from (1.18),

$$(M_{sc} - M|\Delta^4 \equiv 0) = \left[\frac{1}{2}\Sigma\Delta_i^2, \frac{1}{6}\Sigma\Delta_i^3\right]\begin{bmatrix} -\frac{1}{12} \\ \\ \frac{1}{8} \end{bmatrix}$$

$$= -\frac{1}{24}\sum_{i=0}^{n-1}\Delta_i^2 + \frac{n}{48}\Delta^3$$

$$= \frac{1}{2}\,\mathcal{E}\left[Q_1(t) - M|\Delta^4 \equiv 0\right].$$

Thus, bias-wise, the advantage of the "end corrected" systematic random sample estimator is open to question, as this bias is twice that of the centric systematic sample estimator, when $\Delta^4 \equiv 0$, and both estimators are quadrature formulae of degree 1. A general comparison of these two biases was not accomplished, and comparison of variances will be deferred to the chapter (5) dealing with general comparison of quadrature designs.

4.7 Remainder Terms in Quadrature Formulae

It is generally possible to evaluate the remainder of a quadrature formula. This subject is treated in some depth by Kunz (1957, paragraph 7.10 and 7.11) and by Daniell (1940). Some properties of remainder terms are summarized here, and terms are developed for the formulae of interest.

Define

$$RP(\tau) = \int_0^W P(\tau)d\tau - QP(\tau)$$

where Q is any quadrature formula, and $P(\tau)$ is a polynomial of arbitrary degree.

4.7.1 Pertinent Properties of Quadrature Remainders

1. A quadrature formula is said to be of degree m if

$$Rx^k = 0, \; k \leq m$$
$$\neq 0, \; k > m$$

2. A quadrature formula is said to be simplex, degree m, if, given that D^{m+1} is continuous in the range of integration,

$$RP = 0 \text{ implies } D_\xi^{m+1} = 0 ,$$

where ξ is in the range of integration.

3. The remainder term of a simplex formula, Q, of degree m, can be obtained by

$$RP = \frac{Rx^{m+1}D_\xi^{m+1}}{(m+1)!} \qquad (4.35)$$

4. The remainder term of a simplex formula, Q, of degree m, in which the abscissae are equally spaced, can be obtained by:

$$RP(\tau) = \frac{h^{m+2}D_\xi^{m+1}}{(m+1)!} \int_{u_1}^{u_2} u^{[m+1]}du \qquad (4.36)$$

5. The sign of a quadrature formula is said to be the sign of the coefficient, C, in the remainder term,

$$RP = CD_\xi^{m+1}$$

6. The sum of simplex formulae of degree m, all of which are of the same sign, is also a simplex formula of degree m.

7. The remainder of such a sum of simplex formulae is given by

$$R = \sum C_i D_\xi^{m+1} \qquad (4.37)$$

where the C_i are the coefficients of the remainder terms of the composite formulae, and ξ is in the range of the new formula.

4.7.2 Properties of the Formulae Here Used

i. The Newton-Cotes formulae are all simplex (Kunz, p. 143). Further, the degree of the Newton-Cotes formulae is either n - 1 or n, and is always odd.

ii. At least some of the Guassian formulae are simplex, Daniell (1940), although proof that all are simplex was not found. Further, Kopal (1955, p. 351) gives the remainder of the Gaussian Formula in the form (435), which follows the simplex property. All are of degree (2n - 1).

iii. Nothing was found regarding the simplex nature of Tchebycheff formulae. The remainder term given by Kopal (1955, p. 419), is also of the form (4.35). All are of degree n or n + 1, whichever is odd.

iv. It is proven that the Centric formulae are simplex for n even, but proof for n odd was not found. Degree is always odd, either n or n - 1, as in the Newton-Cotes formulae. It is assumed that all are simplex and that the remainder can be obtained by either (4.35) or (4.36).

4.7.3 Remainder of Interest

Remainders of interest are summarized in Table 4.5. A general comment on the remainder terms is in order. They do not provide comparison for any specific function or interval, but rather for some "average" over all intervals and functions. In a specific case, an "inferior" design, as evaluated by the remainder term, may actually be superior by virtue of the particular value of ξ fulfilling the mean value theorem.

Table 4.5 First terms in the expansion of remainders of the basic quadrature formulae. These can also be interpreted as "average" remainders under the mean value theorem

Quadrature Formula	n	$\frac{w^3D^2}{2!10^2}$	$\frac{w^5D^4}{4!10^4}$	$\frac{w^7D^6}{6!10^4}$	$\frac{w^9D^8}{8!10^5}$	$\frac{w^{11}D^{10}}{10!10^6}$
Newton-Cotes:	2	-16.6667				
	3		-8.3333			
	4		-3.7037			
Q_n^N	5			-3.7202		
	6			-2.0952		
	7				-2.5536	
	8				-1.5772	
Centric:	1	8.3333				
	2	2.0833				
	3		3.2407			
	4		1.6764			
Q_n^C	5			2.6546		
	6			1.5946		
	7				2.5875	
	8				.42311	
Tchebycheff:	2		5.5556			
	3		2.0833			
	4			2.6455		
Q_n^T	5			1.3434		
	6				1.9841	
	7				1.1293	
	9					1.0767
Gaussian:	2		5.5556			
	3			3.5714		
Q_n^G	4				2.2676	
	5					1.4316

4.7.4 Relationships for Repeated Sampling

The following relationships are useful in converting the remainders of basic formulae into general remainder terms for repeated application of the same formula along an interval.

Let R_k be the remainder term for the basic formula of k observations and R_n be the combined remainder term, such that

$$R_n/D_\xi^{m+1} = \sum R_{ki}/D_{\xi i}^{m+1} \quad ,$$

$$\text{where } R_{ki} = C_1 h^{m+2} D_{\xi i}^{m+1} \quad ,$$

which can be rewritten,

$$R_{ki} = C_2 \overset{*}{w}^{m+2} D_{\xi i}^{m+1} \quad .$$

Then, if $h = \overset{*}{w}/k$, $C_2 = C_1/k^{m+2}$.

and if $h = \overset{*}{w}/k\text{-}1$, as in the Newton-Cotes formulae, then

$$C_2 = C_1/(k-1)^{m+2} \quad .$$

It follows that if $h = w/n$, where $\bar{w} = n\overset{*}{w}/k$ is the length of the new composite interval,

$$R_n/D_\xi^{m+1} = \sum C_1 h^{m+2}$$

$$= \frac{n}{k} C_1 h^{m+2}$$

$$= (\frac{1}{k})(\frac{1}{n^{m+1}}) \; C_1 w^{m+2} \qquad (4.38)$$

$$= (\frac{k}{n})^{m+1} \; C_2 w^{m+2} \qquad (4.39)$$

Similarly, if $h = w/n-1$, where $w = \frac{n-1}{k-1}\overset{*}{w}$, as in Newton-Cotes formulae, then one need only substitute (n-1) and (k-1) for n and k in formula (4.38) and (4.39).

4.7.5 Remainders in Terms of Δ

By the general relatioship between D^i and Δ^i, one can write the expression for the remainder of a formula of degree m,

$$R_m = \frac{h^{m+2} D_\xi^{m+1} RX^{m+1}}{(m+1)!} \tag{4.40}$$

(where RX^{m+1} is defined over the standardized interval, i.e., $h = 1$) and letting $\Delta^{m+2} \equiv 0$, as

$$R_m = \frac{h\ \Delta^{m+1} RX^{m+1}}{(m+1)!} \tag{4.41}$$

Note here that is necessary to require $\Delta^{m+2} \equiv 0$. Otherwise, ξ satisfying the mean value, Δ_ξ^{m+1}, may lie outside the interval of interest.

4.7.6 The Truncated Difference Expansion of R

An identical result is also obtained when one expands R in terms of differences and truncates the expansion. To illustrate, recall that

$$M = h\left\{\sum_{i=0}^{n-2} y_i + \frac{1}{2}\sum_{i=0}^{n-2}\Delta_i - \frac{1}{12}\sum_{i=0}^{n-2}\Delta_i^2 + \frac{1}{24}\sum_{i=0}^{n-2}\Delta_i^3 - \frac{19}{(24)(30)}\sum_{i=0}^{n-2}\Delta_i^4 + \dots\right\}$$

$$\text{where the } k^{th} \text{ term is } \frac{1}{k!}\sum_{\ell=1}^{k} s(k,\ell)\int_0^1 t^\ell dt\ .$$

Then, successive application of the prismoidal formulae, Q_2^N, to the $n - 1$ panels of a sample of n observation yields a quadrature formula of the form,

$$Q_{2,n}^N \doteq h\left\{\frac{1}{2}y_o + y_1 + y_2 + \dots + y_{n-2} + \frac{1}{2}y_{n-1}\right\},$$

which can be expanded in terms of differences,

$$Q_{2,n}^N = \left\{\sum_{i=o}^{n-2} y_i + \frac{1}{2}\sum_{i=o}^{n-2}\Delta_i\right\} . \tag{4.42}$$

Thus, $R(Q_{2,n}^N) = h\left\{-\frac{1}{12}\Sigma\Delta_i^2 + \frac{1}{24}\Sigma\Delta_i^3 - \dots\right\}$. (4.43)

Then, let $\Delta^3 \equiv 0$,

so that $R(Q_{2,n}^N) = \frac{-h}{12}\sum_{i=o}^{n-2}\Delta_i^2 = \frac{-h(n-1)}{12}\Delta^2$.

Recall (Paragraph 4.7.4) that

$$R(Q_{2,n}^N) = \sum_{i=1}^{n-1} R_{2,i}^N = \frac{-h^3(n-1)}{12} D_\xi^2$$

$$= \frac{-h(n-1)}{12}\Delta^2 \text{ if } \Delta^3 \equiv 0.$$

4.7.7 An Example: Successive Application of Simpson's Rule

$$Q_{3,n}^N = \frac{h}{3}\left\{y_0 + 4y_1 + 2y_2 + \dots + 4y_{n-2} + y_{n-1}\right\} \tag{4.44}$$

$$= h\left\{\Sigma y_i + \frac{1}{2}\Sigma\Delta_i - \frac{1}{12}\Sigma\Delta_i^2 + \frac{1}{24}\Sigma\Delta_i^3 - \frac{1}{48}\Sigma\Delta_i^4 + \frac{1}{90}\Sigma\Delta_i^5 - \dots\right\} .$$

$$R(Q_{3,n}^N) = - h\sum_{i=o}^{n-2}\left\{\frac{1}{48}\left(\frac{19}{15} - 1\right)\Delta_i^4 - \frac{1}{10}\left(\frac{3}{16} - \frac{1}{9}\right)\Delta_i^5 + \dots\right\}$$

$$= - h\sum_{i=o}^{n-2}\left\{\frac{1}{180}\Delta_i^4 - \frac{11}{1440}\Delta_i^5 + \dots\right\} .$$

Then, let $^5 \equiv 0$, so that

$$R(Q^N_{3,n}) = \frac{-h(n-1)}{180}\Delta^4_i ,$$

which is identical to the expression obtained from Paragraph 4.7.4,

$$R(Q^N_{3,n}) = \sum_{j=1}^{(n-1)/2} R^N_{3j} = -\frac{h^5(n-1)}{2(90)} D^4_\xi$$

$$= \frac{-h(n-1)}{180}\Delta^4 \text{ if } \Delta^5 \equiv 0.$$

4.7.8 Second Example

Consider the difference expansion of $R(Q^C_{3,n})$, the remainder of the third degree Centric formula, applied successively to sets of 3 observations, where n is the total number of observations.

$$Q^C_3 = \frac{h}{8}\left\{9y_{h/2} + 6y_{3h/2} + 9y_{5h/2}\right\}$$

$$= h\left\{\sum_{i=o}^{2} y_{h(i+1/2)} + \frac{1}{8}\Delta^2_{h/2}\right\} \quad (4.46)$$

where $\Delta^2_o = \frac{1}{3}\left\{\sum_{i=o}^{2}\left[\Delta^2_i - \Delta^3_i + \frac{2}{3}\Delta^4_i - \frac{1}{3}\Delta^5_i\right] + \frac{1}{3}\Delta^6_1\right\}$.

Expanding and collecting terms in (4.46), and subtracting from the expansion of M, one obtains,

$$R(Q^C_3) = \frac{-h}{640}\sum_{i=o}^{2}\left\{7\Delta^4_i + \frac{21}{2}\Delta^5_i + \dots\right\} \quad (4.47)$$

and

$$R(Q^C_{3,n}) = \frac{-h}{640}\sum_{i=o}^{n-1}\left\{7\Delta^4_i + \frac{21}{2}\Delta_i \quad \dots\right\}. \quad (4.48)$$

It is of interest that, from (4.46),

$$Q_3^C = Q_{1,3}^C + \frac{h}{8}\Delta_{h/2}^2 ,$$

so that $R(Q_3^C) = R(Q_{1,3}^C) - \frac{h}{8}\Delta_{h/2}^2$.

4.8 Derivation of Quadrature Formulae by Least Squares.

It is readily apparent that one can obtain an estimate of the area under a curve of specified degree by integrating the fitted least squares polynomial over the interval of interest. Further, if this procedure is applied to the observations taken in accordance with a basic quadrature design, then the result will be identical to the results of the quadrature formulae, by the uniqueness of the power series expansion.

4.8.1 General Considerations

Consider the application of the least squares quadrature to n observations, where n is greater than the degree of the quadrature k. Then, if $y(\tau)$ can be represented exactly over the interval by a polynomial of degree k, and there is no measurement error, then k^{th} degree quadratures applied to consecutive sets of observations (repeated applications) will yield a result identical to that of a k^{th} degree least squares polynomial over the interval.

If measurement error is present, and the degree of the function is k, then each of the sub-estimates, $\hat{M}_i$, will be an integral of an individual estimate of the polynomial. If the polynomial is truly constant over the entire interval, then the least squares polynomial fitted over all n observations constitutes a "good" average of the possible estimates.

If the polynomial changes between the sections of the interval to which the basic formulae are applied, then the above result does not hold, in general. In fact, appreciable changes in the polynomial necessitate stratification at the points of change, whether least squares or quadrature methods are used. The replicated application of quadratures in the next chapter fit this situation.

4.8.2 Orthogonal Polynomials

Use of orthogonal polynomials in deriving least squares quadrature formulae when ordinates are equally spaced allows use of existing tables (e.g., Anderson and Houseman, 1942), with a great saving of labor. Let

$$\hat{M} = \int \hat{Y}(\underline{X})dx$$

$$= \int \underline{X}'dx\underline{\beta}$$

where $Y = \underline{X}'\underline{\beta} + \epsilon$

$$\underline{Y} = \underset{\sim}{X}'\underline{\beta} + \underline{\epsilon}$$

$$= \underset{\sim}{\xi}'\underline{\beta}^* + \underline{\epsilon}$$

where $\underset{\sim}{\xi}' = \underset{\sim}{X}'\underset{\sim}{\Omega}\lambda$

and $\underset{\sim}{\Omega\lambda}$ is the orthogonalizing transformation.

Then, $\hat{\underline{\beta}}^* = (\underset{\sim}{\xi}\,\underset{\sim}{\xi}')^{-1}\underset{\sim}{\xi}\,\underline{Y}$

These entities can be defined variously, though for the present application, use is faciliated by defining the X variable to have mean zero, so that, from Anderson and Houseman (1942),

$$\underset{\sim}{O} = \begin{bmatrix} 1 & 0 & -(\frac{n^2-1}{12}) & 0 & \frac{3(n^2-1)(n^2-9)}{560} & 0 & \cdots \\ & 1 & 0 & -(\frac{3n^2-7}{20}) & 0 & \frac{15n^4-230n^2-407}{1008} & \\ & & 1 & 0 & -(\frac{3n^2-13}{14}) & 0 & \\ & & & 1 & 0 & \frac{-5(n^2-7)}{18} & \\ & & & & 1 & 0 & \\ & & & & & 1 & \\ & & & & & & \ddots \end{bmatrix}$$

and $$\underset{\sim}{\lambda} = \begin{bmatrix} 1 & & & & \\ & \lambda_1 & & & \\ & & \lambda_2 & & \\ & & & \lambda_3 & \\ & & & & \ddots \end{bmatrix}$$

and the necessary $\underset{\sim}{\lambda}$ and $\underset{\sim}{\xi}$ are tabulated.

Then, $\hat{M} = \int \underline{\xi}'\hat{\underline{\beta}}^* d\,\xi$

$$= \left[\int \underline{X}' dx \right] \left\{ \underset{\sim}{O}\,\underset{\sim}{\lambda}\,(\underset{\sim}{\xi}\,\underset{\sim}{\xi}')^{-1}\underset{\sim}{\xi} \right\} \underline{Y} \tag{4.49}$$

In (4.49), the term in curly brackets is fixed by n and the degree of the formulae, and the term in square brackets is determined by the interval of integration. To illustrate, consider the term in curly brackets when $n = 3$:

$$\begin{bmatrix} 1 & 0 & -8/12 \\ & 1 & 0 \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ & & 3 \end{bmatrix} \begin{bmatrix} 1/3 & & \\ & 1/2 & \\ & & 1/6 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & -2 \\ & 1 & 0 \\ & & 3 \end{bmatrix} \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ -1/2 & 0 & 1/2 \\ 1/6 & -2/6 & 1/6 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -1/2 & 0 & -1/2 \\ 1/2 & -1 & 1/2 \end{bmatrix} .$$

Then, in (4.49), the term in square brackets, centric sample, is

$$\int_{-3/2}^{3/2} \left[1, X, X^2\right] dX = \left[3, 0, 9/4\right] .$$

Hence,
$$Q_3^C = \left[3, 0, 9/4\right] \begin{bmatrix} 0 & 1 & 0 \\ -1/2 & 0 & -1/2 \\ 1/2 & -1 & 1/2 \end{bmatrix} \underline{Y}$$

$$= \left[\frac{9}{8}, \frac{6}{8}, \frac{9}{8}\right] \underline{Y}, \text{ as before.}$$

Similarly, to obtain Q_3^N, one integrates,

$$\int_{-1}^{1} \left[1, X, X^2\right] dx = \left[2, 0, \frac{2}{3}\right] ,$$

from which,
$$Q_3^N = \left[2, 0, \frac{2}{3}\right] \begin{bmatrix} 0 & 1 & 0 \\ -1/2 & 0 & -1/2 \\ 1/2 & -1 & 1/2 \end{bmatrix} \underline{Y}$$

$$= \left[1/3, 4/3, 1/3\right] \underline{Y}, \text{ also as before.}$$

4.8.3 Cases of Particular Interest

The orthogonal polynomial method is of particular interest in obtaining quadrature formulae in special cases of equally spaced ordinates, but non-symmetrically placed limits of integration. In addition, the procedure is quite useful when one desires a quadrature of, say, k^{th} degree, fit to n equally spaced observations. To illustrate, consider a third degree Centric quadrature fit to 5 observations, so that the term in curly brackets is, (since the considered integral is symmetrical, the cubic terms can be omitted!):

$$\begin{bmatrix} 1 & 0 & -2 \\ & 1 & 0 \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1/5 & & \\ & 1/10 & \\ & & 1/14 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ 2 & -1 & -2 & -1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1/5 & 0 & -1/7 \\ 0 & 1/10 & 0 \\ 0 & 0 & 1/14 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ 2 & -1 & -2 & -1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} -3/35 & 12/35 & 17/35 & 12/35 & -3/35 \\ -1/5 & -1/10 & 0 & 1/10 & 1/5 \\ 1/7 & -1/14 & -1/7 & -1/14 & 1/7 \end{bmatrix}$$

and the term in square brackets is

$$\int_{-5/2}^{5/2} \left[1, X, X^2\right] dX = \left[5, 0, \frac{125}{12}\right] ,$$

Thus,

$$\hat{M} = \frac{1}{98}\left[-44, 169, 240, 169, -44 \right] \underline{Y} . \qquad (4.50)$$

Chapter 5. COMPARISON OF QUADRATURE DESIGNS AND THE CONVENTIONAL ESTIMATORS OF SYSTEMATIC SAMPLING

5.1 Introduction

Several difficulties are encountered in making a general comparison of variance (mean square error) of the considered formulae. If one compares those basic formulae of fixed number of observations, then the degree varies, and the R^2 are not directly comparable. If one compares formulae of fixed degree, then the number of observations, and cost, varies. Also, designs requiring unequally spaced abscissae may be considered more costly than those of equal spacing. Lastly, the case of two-stage sampling complicates the comparison of variance of a fixed quadrature design with that of conventional systematic random sampling.

5.1.1 The General Expression of Mean Square Error of a Quadrature Design

As before, the model is

$$y_i = y(\tau_i) + \epsilon_i$$

where $y(\tau)$ is a continuous function of τ and $\mathcal{E}\epsilon = 0$, $\mathcal{E}\epsilon^2 = \sigma^2$, and $\mathcal{E}\epsilon_i\epsilon_j = 0$, all $i \neq j$. Then Q is any quadrature formula of degree m,

$$Q = \sum_{i=1}^{n} W_i y_i = M - R_m + \sum_{i=1}^{n} W_i \epsilon_i \,.$$

It follows that

$$VQ = \mathcal{E}_\epsilon (Q - M)^2 = R_m^2 + \sigma^2 \sum W_i^2 \,, \tag{5.1}$$

in the case of a fixed design, and

$$VQ(t) = \mathcal{E}_t \mathcal{E}_\epsilon \left[Q(t) - M \right]^2 = \mathcal{E}_t R_m^2(t) + \sigma^2 \mathcal{E}_t \sum W_i^2(t) \qquad (5.2)$$

in the case of a randomized design. For brevity, the ϵ's will be dropped from most of the subsequent development, but should be "understood" to be present.

5.1.2 "Types" of Comparisons

In order to put the first part of the problem in perspective, consider comparisons under the following conditions:

i. $D^{m+1} = 0$, but $D^m \neq 0$, where m is the degree of the formulae being compared.

ii. $D^{m+1} \neq 0$, but $D^{m+2} = 0$. In this case, the bias term (-R) involves the product of a known constant and (D^{m+1}), which is also a constant over the interval.

iii. $D^{m+2} \neq 0$. Here one can make "average" comparisons, in the sense of the mean value, D_ξ^{m+1} , or specific comparisons, but not general ones.

Under (i), the variance of formulae of degree m involves only the term σ^2, so that comparisons are readily made among all such formulae. (The additional problem of unequal numbers of observations is discussed later.)

Note, however, that all formulae of degree m' > m are also comparable to those of degree m, as they also involve only the terms in σ^2. Thus, given that $D^{m+1} = 0$, selections of the appropriate formula should be from all of degree m and greater.

Under (ii), comparison of formulae of degree $m^* > m$ is observed in the preceding paragraph. Consider comparison of two formulae of degree m and m + 1, having V_1 and V_2, respectively.

$$V_1 \leq V_2$$

$$\text{if} \quad R_{m1} + s_1\sigma^2 \leq \sigma^2 .$$

Now, under (ii), $R_m = CD^{m+1}$, so that one can make an exact comparison if one can specify the ratio D^{m+1}/σ^2, as follows:

$$V_1 \leq V_2$$

if $\quad CD^{m+1}/\sigma^2 + a_1 \leq a_2$.

Such comparisons will be considered in terms of specific formulae in a later section.

Under (iii), it has been shown that one can develop the general series expansion of the remainder terms. Such an expression may be of interest, but its utility may be questioned. To illustrate, let two formulae of degree m be compared, and let,

$$R_{m1} = \underline{\ell}_1' \, \Sigma\underline{\Delta}^{(m+1)}$$

$$\text{and} \quad R_{m2} = \underline{\ell}_2' \, \Sigma\underline{\Delta}^{(m+1)}$$

where $\underline{\ell}_1$ and $\underline{\ell}_2$ are vectors of constants

$$\text{and} \quad \Sigma\underline{\Delta}^{(m+1)} = \begin{bmatrix} \Sigma\Delta^{m+1} \\ \Sigma\Delta^{m+2} \\ \Sigma\Delta^{m+3} \\ \vdots \end{bmatrix} .$$

Then, for any given $\underline{\ell}_1$ and $\underline{\ell}_2$ and a given $\Sigma\underline{\Delta}^{(m+1)}$, one can determine whether R_1 or R_2 is smaller. However, a general comparison is impossible since for a fixed $\underline{\ell}_1$ and $\underline{\ell}_2$, the choice between R_1 and R_2 will depend on the choice of $\underline{\Sigma\Delta}^{(n+1)}$, *i.e.*, on the specified function and interval of interest.

However, it would seem generally admissable to choose, in the first place, a design of sufficient degree to account for all terms in the expansion known or assumed to be of appreciable magnitude. In general, it will be considered that the primary problem posed by (iii) is determination of the minimum degree which will allow acceptable approximation, so that all comparisons, as such, will be of type (ii).

5.1.3 The Relative Value of Two Formulae of Degree m

The relative "value" of two formulae of degree m can be directly evaluated in terms of mean square error (V) if the two formulae involve the same "cost." Cost differential may occur if one sampling scheme is more difficult to execute than another or if the numbers of observations are different for the two formulae. It would be impossible to define a generally applicable cost function for, say Tchebycheff quadratures as compared to Centric, as such a function involves a subjective element peculiar to a particular application. For this reason, no attention is here given to this problem. The problem of unequal numbers of observations is more generally defined, and is considered. Comparisons can be made independent of the cost function in n by the expedient of evaluating composite formulae, such that two composites constructed from basic formulae of unequal numbers of observations, have the same total number of observations.

5.2 Comparison of Quadrature Designs

5.2.1 Comparison of R^2

In comparison of R^2 of two composite formulae of degree m, each of the type, $\overset{*}{w} = kw/n$, (see Paragraph 4.7.4), where C_{11} is the appropriate constant defined by C_1 in formula (4.38) for the remainder in the numerator, and C_{12} is the comparable constant for the remainder in the denominator, and where the k's are also defined as in (4.38),

$$\frac{R_1^2}{R_2^2} = \left\{\frac{k_2 C_{11}}{k_1 C_{12}}\right\}^2 \text{, constant for all n,} \tag{5.4}$$

so that one or the other of the composite formulae has a constant advantage so far as the Bias term is concerned. This can then be treated as a characteristic advantage for that basic formula.

However, if one of the formulae being compared, (say the first) is a Newton-Cotes, having the relationship, $\overset{*}{w} = (k-1)\, w/(n-1)$,

then
$$\frac{R_{N-C}^2}{R_2^2} = \left\{\frac{(k_2)(n-1)^{m+1} C_{11}}{(k_1-1)(n)^{m+1} C_{12}}\right\}^2 . \tag{5.5}$$

Thus, for two specific formulae, with specified C_{11}, C_{12}, k_1 and k_2, one can solve (5.5) for n', such that the Newton-Cotes formula is superior for all $n > n'$, and inferior for all $n < n'$. It is noted that if $\left(\frac{C_{11}\, k_2}{C_{12}(k_1-1)}\right)^2 < 1$, then the Newton-Cotes formula is always inferior.

A more satisfactory comparison of Newton-Cotes formulae to other formulae can be made from the coefficients of R^2 in Table 5.1. For any n, it is possible to directly compare formulae by multiplying the tabled Numerical Coefficient of the Newton-Cotes formulae by $\left(\frac{n}{n-1}\right)^{2(m+1)}$, where $(m+1)$ is the exponent of D in the constant term of R.

Note that comparison of two Newton-Cotes R's also yields a constant,

$$\frac{R^2_{N-c_1}}{R^2_{N-c_2}} = \left(\frac{(k_2-1)c_{11}}{(k_1-1)c_{12}}\right)^2 . \tag{5.6}$$

5.2.2 Comparison of Coefficients of σ^2

In comparison of coefficients of σ^2, one can again consider the coefficient for composite formulae of equal n. Let C(n) be the coefficient of σ^2 in a composite formula of n observations, and C(k) the coefficient of the basic formula. Thus, if composite formulae are of the type such that any observation (ordinate) is used in only one application of the basic formula,

$$C(k) = \frac{\overset{*}{w}^2}{k^2}\sum_{i=1}^{k} w_i^2 \tag{5.7}$$

$$C(n) \quad \frac{w^2}{kn}\sum_{i=1}^{k} w_i^2$$

$$\text{and} \quad nC(n)/w^2 = kC(k)/\overset{*}{w}^2 . \tag{5.8}$$

That is, the coefficient of $w^2\sigma^2/n$ is the same as the coefficient of $\overset{*}{w}^2\sigma^2/k$. This relationship (5.8) holds for all formulae considered, except for the Newton-Cotes formulae, for which the following expression is appropriate

$$C(n)/_{h}2 = 2W_o^2 + \left(\frac{n-1}{k-1}\right)\sum_{i=1}^{k-2} W_i^2 + \frac{4(n-k)}{k-1}\, W_o^2$$

$$= \left(\frac{n-1}{k-1}\right)\sum_{i=o}^{k-1} W_i^2 + \frac{2(n-k)}{k-1}\, W_o^2 \quad .$$

Hence $$nC(n)/_{w}2 = \frac{n}{(n-1)(k-1)}\left\{\sum_{i=o}^{k-1} W_i^2 + \frac{2(n-k)}{n-1}\, W_o^2\right\} \tag{5.9}$$

It is seen, therefore, that the coefficient of $w^2\sigma^2/n$ is constant in n, for all permissible n, for any given formula except those of the Newton-Cotes type. The reader is referred to Table 5.1 for a summary of these terms.

5.2.3 Examples of Comparison of R^2

To illustrate the general procedure of comparison of remainders (or squared remainders) in series form, recall

i. from Paragraph 1.4.4, the remainder of the first degree formula, $M_{sc} = Q^C_{1,n}$

$$R(Q^C_{1,n}) = -h\left[\sum_{i=o}^{n-1} \Delta_i^! \, A(\underline{t} - \underline{\mu}_o)\right]_{t=\frac{1}{2}}$$

$$= h \sum_{i=o}^{n-1} \Delta_i^! \begin{bmatrix} 0 \\ 1/12 \\ -1/8 \\ 73/240 \\ \vdots \end{bmatrix} \tag{5.10}$$

ii. from Paragraph 4.7.6, the remainder of the first degree formula, $Q^N_{2,n}$

$$R(Q^N_{2,n}) \doteq h \sum_{i=o}^{n-2} \underline{\Delta}'_i \begin{bmatrix} 0 \\ -1/6 \\ 1/4 \\ -19/30 \\ \vdots \end{bmatrix} \qquad (5.11)$$

iii. from Paragraph 4.7.8, the remainder of the 3rd degree formula $Q^C_{3,n}$

$$R(Q^C_{3,n}) \doteq - h \sum_{i=o}^{n-1} \underline{\Delta}'_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 21/80 \\ 63/32 \\ \vdots \end{bmatrix} \qquad (5.12)$$

iv. from Paragraph 4.77, the remainder of the third degree formula, $Q^N_{3,n}$

$$R(Q^N_{3,n}) \doteq h \sum_{i=o}^{n-2} \underline{\Delta}'_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ -2/15 \\ 11/12 \\ \vdots \end{bmatrix} \qquad (5.13)$$

It would appear superficially that detailed compairson of the two Centric formulae, (5.10) and (5.12), with the Newton-Cotes formulae (5.11) and (5.13), is readily accomplished, and that one need only specify $\Sigma\Delta_i$. However, there are several difficulties, all due to the fact that there are n panels of integration, and terms in the summation, for the Centric formulae, but only n - 1 for the Newton-Cotes formulae. Thus, h = w/n for Centric and h = w/n-1 for Newton-Cotes. Further, the differences in the formulae are defined over the respective interval, h.

It is possible, by the methods developed earlier (Section 1.4) and variously used, to express $\Sigma\Delta_i$, as defined for, say, Newton-Cotes formulae in terms of $\Sigma\Delta_i$ defined for Centric formulae, by which devices one could make detailed comparisons of the formulae. However, such modifications are very involved, at best, and it would seem that the only practical comparison of Newton-Cotes and Centric quadratures of degree greater than 1 is limited to the first term in the remainder. As earlier indicated, this can be interpreted either as an exact comparison under the assumption that higher differences are zero or as an "average" comparison in the sense of conventional remainder terms evoking the mean value theorem. When modification is made for number of terms in the summation and the length of the panels of integration, (5.10), (5.11), (5.12) and (5.13) become:

$$R(Q^N_{2,n}|\Delta^3 \equiv 0) = -\frac{w}{12}\Delta^2 = \frac{-w^3}{12(n-1)^2} D^2 \tag{5.14}$$

$$R(Q^C_{1,n}|\Delta^3 \equiv 0) = \frac{w}{24}\Delta^2 = \frac{w^3}{24n^2} D^2 \tag{5.15}$$

$$R(Q^N_{3,n}|\Delta^5 \equiv 0) = -\frac{w}{180}\Delta^4 = \frac{-w^5}{180(n-1)^4} D^4 \tag{5.16}$$

$$R(Q^C_{3,n}|\Delta^5 \equiv 0) = \frac{-7w}{640}\Delta^4 = \frac{-7w^5}{640n^4} D^4 \tag{5.17}$$

It is seen, then, that in comparing formulae of this type, in which series expansion of the remainder term is relatively easily obtained, it is still expedient to resort to the conventional form of the remainder in order to make a proper comparison. Therefore, development of the series expansion of the remainders will not be necessary, and the method of Paragraph 4.7.4 is used to obtain the coefficients of R^2 in Table 5.1.

<u>5.2.4</u> <u>Discussion of Coefficients of σ^2 in Formulae of Interest</u>

Following the general considerations of 5.1.2 and the developments of 5.1.3, comparisons of the coefficient of $w^2\sigma^2/n$ are of interest. Reference is made to Table 5.1 and to equations (5.7) and (5.9). It is seen that many of the formulae have coefficients inflated no more than 15% over the minimum, 1. In particular one notices that Q_3^C, Q_4^C, Q_6^C, Q_3^G, and Q_4^G (in addition to the Tchebycheff formulae which have coefficients of 1, by definition) are <u>good</u> in this respect.

Also, it is noted that the Centric formulae are considerably better than the Newton-Cotes formulae for small n and 3^{rd} and 5^{th} degree formulae, but the 7^{th} degree Centric formula is appreciably poorer, relatively. Surprisingly, the Gaussian formulae hold up well for those considered.

<u>5.2.5</u> <u>Discussion of R^2 in Formulae of Interest</u>

With regard to the remainder terms, (Table 5.1), the Gaussian and Tchebycheff formulae hold up well, so that if there is no penalty associated with unequal abscissae, the Gaussian and Tchebycheff formulae are decidedly best, at least for formulae of degree 7 or less. In the formulae of evenly spaced abscissae, it is seen that no general superiority of one type or another exists.

5.2.6 General Comments on Comparison of Mean Square Error

From the two preceding paragraphs, and Table 5.1, it can be summarized that for a given ratio, D^{m+1}/σ^2, and a given cost function in n and in the configuration of abscissae, one can select the formulae of degree m which is "best." However, general superiorities do not exist unless one places restrictions on the cost function.

Further, the irregularity of the coefficients in Table 5.1, i.e., the absence of clear-cut trends, leads one to wonder if there exist certain formulae, higher than those considered, which will have desirable properties. This could be examined by a computer compilation of the pertinent coefficients of formulae to arbitrary degree, and such a compilation would appear desirable.

5.2.7 Placing Restrictions on D^m

It would be most helpful in choosing the "best" formula if it were possible to infer the magnitude of $\Sigma D_i^{m+1}/\sigma^2$ from a known (or assumed) $\Sigma D_i^m/\sigma^2$, or at least to be able to evaluate the relative magnitude of the successive terms in the expansion of R. However, this does not seem possible, short of specifying the function.

5.3 Some Specific Comparisons of Mean Square Error

Consider the conventional systematic sampling estimators, $M_{sc} = Q_1^C$, M_{sr} and also the end-corrected systematic sample with random start $Q_1(t)$.

5.3.1 The General Inferiority of $Q_1(t)$

It was shown earlier that:

1. Bias of M_{sc} was half that of $Q_1(t)$, when $\Delta^4 \equiv 0$.
2. Coefficient of σ^2/n in M_{sc} and M_{sr} is 1.
3. Coefficient of σ^2/n in variance of $Q_1(t)$ is $\left(1 + \frac{n}{6(n-1)^2}\right)$.

It is apparent that M_{sc} is always better than $Q_1(t)$, and that Q_2^N is better than $Q_1(t)$ for some n and some D^2/σ^2. Further, if $D^4 \equiv 0$, a number of higher degree formulae are better than $Q_1(t)$, at least for some n. In short, there doesn't seem much justification for the end-correction formula, $Q_1(t)$, although some of the other formulae of this type may have better characteristics.

The case of two stage sampling will be treated later (Section 5.4), and it will be demonstrated that in repeated sampling the randomized designs achieve superiority over the fixed designs when replication is sufficiently great and $\sigma^2 > 0$. This result may provide a justification for the end-corrected systematic random sample, as the end-corrections can be used to advantage if one has systematic random samples for a number of occasions, and then wishes to estimate the values for individual occasions.

5.3.2 Further Comparison of M_{sc} to M_{sr}

Comparison of M_{sr} with M_{sc}, $Q_1(t)$, and other higher order Quadrature formulae can be made by use of the results of Sections 2.7, 4.7

and the earlier part of Chapter 5. To illustrate another aspect of comparison of expanded error terms, in cases in which the coefficients of σ^2 are equal, consider,

$$V(M_{sr}|\Delta^3 \equiv 0) = \frac{w^2}{12}\left[\frac{(\Sigma\Delta_i)^2}{n^2} + \frac{(\Delta^2)}{60}\right] + w^2\sigma^2/n$$

$$V(M_{sc}|\Delta^3 \equiv 0) = \frac{5w^2}{48(60)}(\Delta^2)^2 + w^2\sigma^2/n$$

Then, $V(M_{sr}|\Delta^3 \equiv 0) < V(M_{sc}|\Delta^3 \equiv 0)$ as

$$\frac{(\Sigma\Delta_i)^2}{n^2(\Delta^2)^2} < 1/240\ , \quad \Delta^2 \neq 0, \tag{5.18}$$

and never if $\Delta^2 = 0$.

From (5.18) it is seen that the necessary condition (under $\Delta^3 \equiv 0$) for M_{sr} to be "better" than M_{sc} is for the constant second difference to be 16 times the absolute average first difference. This implies symmetry of the function about the mid-point of the interval, and constitutes a potential rule of thumb in application.

Similar comparisons between M_{sr} and M_{sc} are possible when higher differences are assumed $\neq 0$, but little will be accomplished. The regions of superiority are more complex, and there is little hope of finding working rules. Similarly, comparison of M_{sr} and higher degree quadrature formulae would be redundant, since the methods and ideas have been developed previously.

5.4 Comparisons under Two Stage Sampling

When the designs here considered are applied at the second stage of a two stage sampling scheme, the comparisons between randomized designs and fixed designs are affected, although comparisons among fixed designs are not. Since the only randomized design still under consideration is M_{sr}, and comparison among fixed designs has already been treated, the present treatment will be restricted to a comparison of M_{sr} and M_{sc} under two stage sampling.

5.4.1 The Two General Cases

Two "cases" are recognized in the two stage sampling problem.

i. Repeated sampling is over different "realizations" of the same function and interval. The $\Sigma\underline{\Delta}_{ij}$ are identical for all i; that is, constant for sampled occasions or locations (primary sampling units). Only the "measurement" error, ϵ, changes from realization to realization, and the ϵ are assumed independently and identically distributed, with mean zero.

ii. Repeated sampling is over different "realizations" of the same generating process, so that $\Sigma\underline{\Delta}_{ij}$ can be different on any distinct occasion or location (primary sampling unit). Measurement error also exists, subject to the restrictions under i.

5.4.2 Case i

Consider an Universe of r' PSU's, of which r are selected at random with equal probability. Then define

$$M_{sr.} = \frac{r'}{r} \sum_{i=1}^{r} M_{sri} \quad ,$$

where M_{sri} is the conventional systematic random sample estimator applied to the i^{th} PSU, where t, the random starting point, is selected independently for the sampled PSU's, and where,

$$V(M_{sri}) = \left(\frac{w_i}{n_i}\right)^2 \left\{ \Sigma\underline{\Delta}'_{ij} \mathcal{C} T_o^* \Sigma\underline{\Delta}_{ij} + n_i\sigma^2 \right\} .$$

Also, define,

$n = rn_i$, the total number of observations

$w = r'w_i$, the total "length" of the Universe.

Then, $V(M_{sr.}) = (\frac{r'}{r})^2 \sum_{i=1}^{r} V(M_{sri})$

$$= \frac{w^2}{n^2} \sum_{i=1}^{r} \left\{ \Sigma\underline{\Delta}'_{ij} \mathcal{C} T_o^* \Sigma\underline{\Delta}_{ij} + \frac{n\sigma^2}{r} \right\} .$$

Then, under the restriction of case i,

$$V(M_{sr.}) = (\frac{W}{n})^2 r \left\{ \Sigma\underline{\Delta}'_j \mathcal{C} T_o^* \Sigma\underline{\Delta}_j + \frac{n\sigma^2}{r} \right\} . \tag{5.19}$$

Similarly, define

$$M_{sc.} = \frac{r'}{r} \sum_{i=1}^{r} M_{sci} ,$$

where $V(M_{sci}) = \left(\frac{w_i}{n_i}\right)^2 \left\{ \Sigma\underline{\Delta}'_{ij} \underset{\sim}{C}^* \Sigma\underline{\Delta}_{ij} + n_i\sigma^2 \right\} .$

Then under the restriction of Case i,

$$V(M_{sc.}) = (\frac{w}{n})^2 r^2 \left\{ \Sigma\underline{\Delta}'_{ij} \underset{\sim}{C}^* \Sigma\underline{\Delta}_{ij} + \frac{n\sigma^2}{r} \right\} . \tag{5.20}$$

Thus, if one assumes that $\Delta^3 \equiv 0$, $\Delta^2 \neq 0$, (referring to the identities 1.14, 1.17) then $V(M_{sr}) < V(M_{sc})$ when

$$r > 48\left\{\frac{\Sigma\Delta_{ij}}{n_i\Delta^2}\right\}^2 + 4/5 \quad .$$

That is, the superiority of M_{sc} is reduced as the number of sampled PSU's increases, and for sufficiently large r, M_{sr} becomes superior. Again, however, $\Sigma\Delta_{ij}/\Delta^2$ must be specified for an explicit evaluation, and choice of design.

5.4.3 Case ii

Consider realizations of a process such that a polynomial of the same degree is appropriate for the realization characterizing each PSU, but such that coefficients are subject to change. Then if one samples each PSU in accordance with schemes yielding, say M_{sc} and M_{sr}, then the centric systematic sample will have the same over-all variance as if the error were simple measurement error, and the function were the sum of the realizations. That is, when $r = r'$,

$$\overset{*}{V}(M_{sc.}) = \frac{w^2}{n^2}\, \Sigma\, \Sigma\underline{\Delta}'_{ij}\underset{\sim}{C}^*\Sigma\, \Sigma\underline{\Delta}_{ij} + w^2\sigma^2/n \;, \qquad (5.21)$$

whereas the error of M_{sr}, when an independent random start is chosen for each PSU, is

$$\overset{*}{V}(M_{sr.}) = \frac{w^2}{n^2}\sum_{i=1}^{r}\left[\; \Sigma\underline{\Delta}'_{ij}\mathcal{E}T_o^*\Sigma\underline{\Delta}_{ij} \;\right] + w^2\sigma^2/n \;. \qquad (5.22)$$

(Expression (5.22) follows directly from (2.20), as the sampling scheme reduces to stratified sampling, with systematic random sampling within strata.)

Formulae (5.21) and (5.22) can be compared, letting $\Delta^3 \equiv 0$, $\Delta^2 \neq 0$, and specifying that Δ^2 is not necessarily the same for all realizations.

Then, $\overset{*}{V}(M_{sr.}) < \overset{*}{V}(M_{sc.})$ as

$$\frac{\Sigma(\Sigma\Delta_{ij})^2}{n_i^2} + \frac{\Sigma(\Delta_i^2)^2}{30} < \frac{(\Sigma\Delta_i^2)^2}{24} \quad . \tag{5.23}$$

Thus, the variance of the Δ^2's and the $\bar{\Delta}$'s, in addition to the relative magnitude of $\Sigma\Delta_i/\Delta^2$, influence the choice of sampling schemes, in case ii. (The term $\bar{\Delta}$ is here used to signify $\frac{1}{n}\Sigma\Delta_i$.) The larger the among realization variance of Δ^2 and $\bar{\Delta}$, the larger the value of r required to make M_{sr} superior to M_{sc}. As in Case i, it is of greater interest to consider a sample of PSU's than to include all in the sample. The two-stage sample variances of the two schemes under Case ii are developed below.

$$V(M_{sr.}) = \mathcal{E}\left\{ \frac{r'}{r}\sum_{i\epsilon s} M_{sri} - M \right\}^2$$

$$= \left(\frac{r'}{r}\right)^2 \left[r^2\, \mathcal{E}\left\{\frac{1}{r}\sum_{i\epsilon s} M_i - M/r'\right\}^2 + \mathcal{E}\left\{\sum_{i\epsilon s}(M_{sri} - M_i)\right\}^2\right] \tag{5.24}$$

since expectation of cross products is zero. The last term is the expression for $V(M_{sr})$ given by (5.22), and the first is the variance of the estimated mean of the M_i, based on the r sample M_i .

That is,

$$V(M_{sr.}) \doteq (r')^2 V_r(\bar{M}) + \left(\frac{r'}{r}\right) \overset{*}{V}(M_{sr.}) \tag{5.25}$$

where $\overset{*}{V}(M_{sr.})$ is defined by (5.22).

However, the comparable expression for $V(M_{sc.})$ involves the cross product term,

$$(\frac{r'}{r})^2 \mathcal{E} \Big[\sum_{i\epsilon s} M_i - \frac{rM}{r'} \Big] \Big[\sum_{i\epsilon s} M_{sci} - \sum_{i\epsilon s} M_i \Big] ,$$

which may be interpreted as the covariance of the sum of the sample M_i and the bias of the Centric systematic estimator over a sample of r realizations. Also, in the case of the centric designs,

$$(\frac{r'}{r})^2 \mathcal{E} \Big[\sum_{i\epsilon s} (M_{sci} - M_i) \Big]^2$$

$$= (\frac{r'}{r}) \overset{*}{V}(M_{sc.}) - \frac{(r')^2(r'-r)}{r} \mathcal{E}(M_{sci} - M_i)(M_{scj} - M_j), \; i \neq j ,$$

the second term of which is the covariance among the realization biases.

Thus,

$$V(M_{sc.}) = (r')^2 V_r(\bar{M}) + (\frac{r'}{r}) \overset{*}{V}(M_{sc.}) + \Big[(\frac{r'}{r})^2 \mathcal{E} \Big\{ \sum M_i - \frac{rM}{r'} \Big\} \Big\{ \sum M_{sci} - \sum M_i \Big\} \Big]$$

$$- \Big[\frac{(r')^2(r'-r)}{r} \mathcal{E}(M_{sci} - M_i)(M_{scj} - M_j) \Big], \; j \neq i. \qquad (5.26)$$

Since both of the covariance terms in (5.26) can generally be expected to be positive, no general effect of sampling of realizations on the relative merits of the random and centric designs can be inferred. Again, all comparisons require specification of the functions.

5.4.4 Generalization of Two-Stage Results

It is apparent that the results of the preceding section can be generalized to cover all quadrature designs under two-stage sampling. In particular, generalization follows the form of (5.26), with appropriate simplifications in accordance with the nature of the particular design:

$$V_r(Q) = \frac{w^2}{r'rn^2}\left\{\sum_{i=1}^{r'}\sum_{j=1}^{n}\underline{\Delta}'_{ij}\underset{\sim}{Q}^*\sum_{i=1}^{r'}\sum_{j=1}^{n}\underline{\Delta}_{ij}\right\} + Cw^2\sigma^2/nr$$

$$+ (r')^2V_r(\bar{M}) + (\frac{r'}{r})^2\,\mathcal{E}\left[\sum_{i\epsilon s}M_i - \frac{rM}{r'}\right]\left[\sum_{i\epsilon s}(Q_i - M_i)\right]$$

$$- \frac{(r')^2(r' - r)}{r}\,\mathcal{E}(Q_i - M_i)(Q_j - M_j)\ , \qquad (5.27)$$

where $\underset{\sim}{Q}^*$ is the appropriate matrix for the square of the difference expansion of (Q - M) and C is a function of the particular design, as described in Table 5.1. It was earlier shown that truncation of this expression to the first non-zero term is equivalent to utilizing the conventional remainder term.

It is anticipated that (5.27) will be of value in designing two stage samples. Modifications in design at the first stage may be justified when restrictions can be placed on the M_i. For example, certain knowledge of the M_i might lead to a systematic selection of PSU's. Such designs are not considered here, though it is apparent that the present results are directly applicable under stratification at the first stage.

Table 5.1 Components of mean square error, $R^2 + Cw^2\sigma^2/n$, of Quadrature Formulae. The general term refers to repeated application of the basic formulae along an interval. R^2 for the basic formula is easily obtained from the table of remainders, Table 4.5.

	Coefficient of $w^2\sigma^2/n$			R^2	
Formula[a]	Basic	General	limit $n \to \infty$	Coefficient	Common Term
N(2)	1	$n(2n-3)/2(n-1)^2$	1	2.7778[b]	$10^{-2}\left(\frac{w^3D^2}{2!n^2}\right)^2$
C(1)	1	1		.6944	
$Q_1(t)$	4/3	$1+n/6(n-1)^2$	1		
N(3)	1.5000	$2n(5n-6)/9(n-1)^2$	1.1111	1.7778[b]	$10^{-2}\left(\frac{w^5D^4}{4!n^4}\right)^2$
N(4)	1.2500	$3n(11n-14)/32(n-1)^2$	1.0313	9.0000[b]	
C(3)	1.0313	1.0313		6.8904	
C(4)	1.0069	1.0069		18.4177	
G(2)	1	1		.7901	
T(3)	1	1		2.8476	
N(5)	1.4137	$n(1.1793n-1.3728)/(n-1)^2$	1.1793	2.3219[b]	$\left(\frac{w^7D^6}{6!n^6}\right)^2$
N(6)	1.2277	$n(1.0666n-1.2842)/(n-1)^2$	1.0666	10.7174[b]	
C(5)	1.2541	1.2541		17.2043	
C(6)	1.0609	1.0609		55.3501	
G(3)	1.0556	1.0556		.06778	
T(4)	1	1		1.1742	
T(5)	1	1		4.4061	
N(7)	1.7075	$n(1.4922n-1.6637)/(n-1)^2$	1.4922	1.8396[b]	$10^{-3}\left(\frac{w^9D^8}{8!n^8}\right)^2$
N(8)	1.2883	$n(1.1537n-1.3388)/(n-1)^2$	1.1537	8.2669[b]	
C(7)	2.9766	2.9766		22.2500	
C(8)	1.4168	1.4168		5.0390	
G(4)	1.0926	1.0926		.00208	
T(6)	1	1		1.1106	
T(7)	1	1		4.2383	
G(5)	1.1176	1.1176			
G(6)	1.1353	1.1353			

[a] Formulae designations are of the form $Q_n^N = N(n)$, where n refers to the number of observations in the basic formula, with exception of $Q_1(t)$, in which case the subscript refers to the degree.

[b] Substitute (n-1) for n in the common term. With the exception of those indicated by b , all R^2's having the same common term can be compared directly, for all n, by comparison of the coefficient of R^2. When b appears by the coefficient, it must (for purposes of comparison) be multiplied by $\left(\frac{n}{n-1}\right)^{2(m+1)}$, where (m+1) is the power of n in the common term.

Chapter 6. ESTIMATION OF VARIANCE OF ESTIMATE OF QUADRATURE DESIGNS

6.1 Introduction

Estimators of the variance (mean square error) of simple quadrature designs (M_{sr}, M_{sc} and others) were considered in Chapter 3. In the present chapter, estimators of variance of higher degree formulae are investigated. Results are also extended to the case of two stage sampling.

6.1.1 Orientation

In Chapter 3, it was demonstrated that the difference expansions of the squared error of the conventional estimators of systematic sampling could be estimated by linear combinations of squares of linear combinations of differences. Further, it was pointed out that such a procedure was of dubious value, since the estimator of mean ordinate should be modified to account for all terms of known significant magnitude in the remainder. That is, information about the remainder is really information about the mean ordinate, and should be used as such if possible, rather than as information about error of estimate.

It is readily apparent that similar estimators of difference expansions of variance of quadrature formulae can be constructed, if all of the information has not been used in the quadrature formulae. Several cases are considered in the following paragraphs in which such information remains available. Except where indicated to the contrary, it is assumed that the basic formulae contain no unused information.

6.1.2 Successive Application of Quadratures Along an Inverval

This is the case treated in Chapter 3. Here, interest is in quadrature formulae of higher degree. As before, the available information lies in contrasts between the individual quadratures comprising the whole. As use of these contrasts in estimating variance requires assumptions regarding the "degree" of the underlying polynomial, the use of supplementary observations will be useful.

6.1.3 Application of Quadratures, Case i, Paragraph 5.4.1

This is one of the two cases of two-stage sampling that were considered in Chapter 5. Contrasts between PSU's yield information about σ^2, only. As will be seen later, this information is quite useful, but its use is dependent on the assumption that the polynomial and the interval is the same. Here supplementary observations are necessary to obtain any information about R.

6.1.4 Application of Quadratures, Case ii, Paragraph 5.4.1

This is a generalization of the preceding paragraph, in which R is no longer constant over PSU's. Here, contrasts between PSU's yield information about σ^2 and the variation in R, but not separately. Again, it is necessary to use supplementary observations in order to obtain any information about the magnitude of R.

6.1.5 Application of Systematic Random Sample

The systematic random sample, with conventional estimator, M_{sr}, provides an exception to the statements of the three preceding paragraphs regarding limitation of the contrasts in estimation of mean square error. If the "realizations" (PSU's) sampled are from an infinite population, the contrasts provide an unbiased estimator of the variance of the over-all estimate. However, if this population is finite, then supplementary observations are necessary in order to construct an appropriate variance estimator.

6.1.6 Application of Least Squares Quadratures

Least squares quadratures also provide an exception to Paragraphs 6.1.2, 6.1.3 and 6.1.4, as information about the variance is contained within the PSU's. As earlier indicated, the least squares quadrature is also valuable as an "average" of the successive fitted polynomials of the case in Paragraph 6.1.2 and may generally be better than that case (although this was not proven) so that estimation of variance of a single least squares quadrature is of interest.

From the results of Chapter 5, it appears likely that the degree of the least squares quadrature which provides the minimum mean square error is lower than the degree of the fitted polynomial which, in the conventional treatment of polynomials, provides the minimal residual mean square. Thus, it must be considered that R, and hence r_i, the contribution of bias to the difference between y_i and $\hat{y}_i$ (Paragraph 6.2.1), will not be negligible for the optimal degree with regards to quadrature.

A moderate amount of effort, with little success, was expended on the problem of obtaining an estimator of variance of the least squares quadrature from the residual mean square. Results are reported in Section 6.3. It appears that direct use of the residual mean square is not too damaging.

6.2 Supplementary Observations

Supplementary observations, $y_i = y(\tau_i) + \epsilon_i$, taken at random values of τ within the interval of interest, provide information about the error of quadrature in the form of

$$e_i = y_i - \hat{y}_i$$

where $\hat{y}_i$ is the value of the fitted polynomial implied by the quadrature formula at the point τ_i. Now in the calculus of finite differences, $\hat{y}_i$ can be written down in terms of various interpolation formulae. However, the results of Section 4.8 provide a useful framework in familiar notation that does not require development of additional formulae. In use of the notation of orthogonal transformations, recall that the transformation is general, and that the results apply equally to all quadrature designs.

In the subsequent development, it is specified that

$$\begin{aligned} e_i &= y(\tau_i) - \hat{y}_i + \epsilon_i \\ &= [y(\tau_i) - \mathcal{E}(\hat{y}_i)] - [\hat{y}_i - \mathcal{E}(\hat{y}_i)] + \epsilon_i \\ &= r_i - z_i + \epsilon_i \ . \end{aligned}$$

Restrictions on this general expression will be made in the various paragraphs.

As in Section 4.8, the variable X is used for the argument of y in development of general results. The primary justifications for this change are that re-definition of vectors is simplified, and the range of X is conveniently manipulated.

6.2.1 Estimation of R

Consider the estimation of R by use of supplementary observations, when it is assumed that measurement error is absent. Then,

$$e_i = r_i$$

and

$$R = M - \hat{M}$$

$$= h \int_0^n r(x)dx$$

where $r_i = r(x_i)$.

That is, $r(x)$ is the signed difference between the fitted polynomial and the true function at the point x, $0 \leq x \leq n$. Now $r(x)$ is observable, and if values of $r(x)$ are taken at k randomly chosen values of X, $X = U(0,n)$, then

$$\hat{R} = nh\bar{r} = \frac{nh}{k}\sum_{i=1}^{k} r(x_i) \qquad (6.1)$$

is an unbiased estimator of R.

6.2.2 The Estimation of R^2

It will be recognized that

$$(nh)^2 \mathcal{E} r^2 \geq R^2 \qquad (6.2)$$

and that $(nh)^2 \mathcal{E} r^2 > R^2$ when $R > 0$, but that $\mathcal{E} r^2 \neq 0$ even though $R = 0$, unless $y(x)$ is of the same degree as $\hat{y}(x)$. Further recall that many formulae are defined so that $R \equiv 0$ even when $y(x)$ is a polynomial of (specified) higher degree than $\hat{y}(x)$, so that the inequality (6.2) may be great indeed.

Thus, it is desirable that an estimator of R^2 be found which will, in effect, reflect as much as possible of the "hidden precision" due to the form of the design. Consider

$$\widehat{R_k^2} = \left[\frac{nh}{k} \sum_{i=1}^{k} r_i \right]^2 = (\hat{R})^2 . \tag{6.3}$$

Now $\widehat{R_k^2}$ is consistent, so that for large k it is a good estimator of R^2, even though it is biased. Consider, then, a modification proposed by Quenouille (1956),

$$\overset{*}{R}{}_k^2 = k \widehat{R_k^2} - \frac{(k-1)}{k} \sum_{i=1}^{k} (\widehat{R_{k-1}^2})_i , \tag{6.4}$$

where $$(\widehat{R_{k-1}^2})_i = \left[\frac{nh}{k-1} \sum_{\substack{j \neq i \\ =1}}^{k} r_j \right]^2 .$$

Now the Quenouille estimator is appropriate if $(\widehat{R^2} - R^2)$ is expandable in a power series in $1/k$, in which case $\overset{*}{R}{}^2$ has a bias to order $1/k^2$, compared to the bias of $\widehat{R^2}$ to order $1/k$ (Quenouille, 1956).

6.2.3 The Estimationof $V(\hat{M})$

Let the function be a polynomial of the same degree $(n - 1)$ as the polynomial specified by the formula, so that $r(x) \equiv 0$, but specify that measurement error is present. Recall, in this case,

$$V(\hat{M}) = h^2 \sum_{i=1}^{n} W_i^2 \sigma^2 = h^2 \underline{W}' \underline{W} \sigma^2$$

where the W_i are the weights of the formula.

Then, the difference between the supplementary observation, $y(x)$ and the fitted polynomial, at X, is

$$e = y - \hat{y}$$

where, in the notation of Section 4.8, one can write

$$\hat{y} = \underline{x}'(\underset{\sim}{X}\underset{\sim}{X}')^{-1}\underset{\sim}{X}\underline{Y}$$

$$= \underline{x}'\underset{\sim}{Q}\underset{\sim}{\lambda}\,(\underset{\sim}{\xi}\underset{\sim}{\xi}')^{-1}\,\underset{\sim}{\xi}\,\underline{Y}$$

where $\underline{Y}$ is the vector of observations used in the formula and $\underline{x}'$ is the vector $[1, x, x^2, \ldots, x^{n-1}]$ defined for the value of X chosen for the supplementary observation.

Then $\mathcal{E}_\epsilon e = 0$

$$\mathcal{E}_\epsilon e^2 = \{1 + \underline{x}'\,\underset{\sim}{\psi}\underline{x}\}\sigma^2$$

where $\underset{\sim}{\psi} = \underset{\sim}{Q}\underset{\sim}{\lambda}\,(\underset{\sim}{\xi}\underset{\sim}{\xi}')^{-1}\,\underset{\sim}{\lambda}\underset{\sim}{Q} = (\underset{\sim}{X}\,\underset{\sim}{X}')^{-1}$.

Hence $s^2 = \{1 + \underline{x}'\psi\underline{x}\}^{-1}(y - \hat{y})^2$ (6.5)

is an unbiased estimator of σ^2 .

Further, consider k random values of e. Then

$$\mathcal{E}_\epsilon \sum_{i=1}^{k} e_i^2 = \left\{k + \sum(\underline{x}_i'\,\underset{\sim}{\xi}\,\underline{x}_i)\right\}\sigma^2 \qquad (6.6)$$

and

$$\mathcal{E}_\epsilon\left(\sum e_i\right)^2 = \{k + (\sum \underline{x}_i')\,\underset{\sim}{\psi}\,(\sum \underline{x}_i)\} \qquad (6.7)$$

so that unbiased estimators of σ^2 may be derived by solving (6.6) and (6.7) in the manner of (6.5). The estimator defined by (6.6) is appropriate in the case considered in the present paragraph, but both (6.6) and (6.7) will be useful in subsequent development.

6.2.4 The General Case

Let the fitted polynomial be of lower degree than the function and measurement error be present. The results of Paragraphs 6.2.2. and 6.2.3 are pertinent.

As before,
$$e_i = y_i - \hat{y}_i$$
$$= r_i + \epsilon_i + \left[\mathcal{E}_\epsilon \hat{y}_i - \hat{y}_i\right]$$
$$\mathcal{E}_\epsilon e_i^2 = r_i^2 + \{1 + \underset{\sim}{x}' \underset{\sim}{\psi} \underset{\sim}{x}\}\sigma^2$$

and
$$\mathcal{E}_\epsilon\left(\sum e_i\right)^2 = \left(\sum r_i\right)^2 + \{k + \left(\sum \underset{\sim}{x}_i'\right) \underset{\sim}{\psi} \left(\sum \underset{\sim}{x}_i\right)\}\sigma^2 .$$

Then define
$$\hat{V}_k = \left(\frac{nh}{k} \sum_{i=1}^{k} e_i\right)^2 \tag{6.8}$$
$$(\hat{V}_{k-1})_i = \left(\frac{nh}{k-1} \sum_{\substack{j \neq i \\ =1}}^{k} e_j\right)^2 \tag{6.9}$$

and after aragraph 6.2.2,
$$\overset{*}{V}_k = k\hat{V}_k - \frac{(k-1)}{k} \sum_{i=1}^{k} (\hat{V}_{k-1})_i . \tag{6.10}$$

Then,
$$\mathcal{E}_\epsilon \overset{*}{V}_k = \overset{*}{R}{}_k^2 + \left\{k + \left(\sum_{i=1}^{k} \underset{\sim}{x}_i'\right) \underset{\sim}{\psi} \left(\sum_{i=1}^{k} \underset{\sim}{x}_i\right)\right\} \frac{(n-h)^2\sigma^2}{k}$$
$$- \left\{(k-1) + \frac{1}{k}\sum_{i=1}^{k}\left[\left(\sum_{\substack{j \neq i \\ =1}}^{k} \underset{\sim}{x}_j'\right) \underset{\sim}{\psi} \left(\sum_{\substack{j \neq i \\ =1}}^{k} \underset{\sim}{x}_j\right)\right]\right\} \frac{n^2h^2\sigma^2}{(k-1)} ,$$
$$= \overset{*}{R}{}_k^2 + \left\{\sum_{i=1}^{k} \sum_{j \neq i} (\underset{\sim}{x}_i' \underset{\sim}{\psi} \underset{\sim}{x}_j)\right\} \frac{n^2h^2\sigma^2}{k(k-1)} . \tag{6.11}$$

Then, since the k values of x are selected at random, one can write

$$\mathcal{E}_{x,\epsilon}\overset{*}{V}_k = \mathcal{E}_x\overset{*2}{R}_k + (\mathcal{E}\underline{x}')\ \underset{\sim}{\psi}\ (\mathcal{E}\underline{x})n^2h^2\sigma^2 .$$

Recall that

$$V(Q) = R^2 + h^2\underline{W}'\underline{W}\sigma^2$$

$$= R^2 + (\mathcal{E}\underline{x}')\ \underset{\sim}{\psi}\ (\mathcal{E}\underline{x})n^2h^2\sigma^2$$

so that, in expectation over x, $\overset{*}{V}_k$ would seem satisfactory. However, the attempt to construct estimators of the form of (6.5), i.e., in which the correct coefficients are obtained, conditional on $\underline{x}$, was unsuccessful. The problem, then, must be considered only partially solved, since a conditional estimator is patently desirable.

6.2.5 An Alternate Form

The term, $(\frac{nh}{k}\sum_{i=1}^{k} e_i)^2$, employed in the preceding paragraph and made up of observed deviations from the fitted curve, can be rewritten,

$$(\frac{nh}{k}\sum_{i=1}^{k} e_i)^2 = (\frac{nh}{k})^2 \left\{ \sum y_i - (\sum \underline{x}_i)\ \underset{\sim}{\psi}\ \hat{\underline{\beta}} \right\}^2$$

where $\sum y_i$ is the sum of the supplementary observations,

$\sum \underline{x}_i$ is the sum of the vectors defined by the chosen supplementary observations,

and $\hat{\underline{\beta}} = \underset{\sim}{\psi}\underset{\sim}{X}\underline{Y} = \underset{\sim}{Q}'\underset{\sim}{\lambda}(\underset{\sim}{\xi}\underset{\sim}{\xi}')^{-1}\underset{\sim}{\xi}\underline{Y}$, from the n observations comprising the formula.

6.2.6 Separate Estimation of Components of Mean Square Error

A useful device in problems of this sort is that of estimating the two components separately, or at least one of them independently of the other. It is apparent that when it is possible to sample so as to obtain clean estimates of σ^2, it is then possible to construct estimators of the form

$$\hat{V} \doteq \overset{*}{R}{}^2 + C\sigma^2$$

with any desired coefficient C, by linear combination of the estimator of σ^2 and the estimator of the preceding paragraphs. Unfortunately, it is seldom possible in the applications considered here to obtain clean estimates of σ^2. An exception is the situation noted in paragraph 6.1.3.

6.3 Differences Between Successive Quadratures Along an Interval

Although the differences between successive quadratures along an interval are considered of little general value, except perhaps in the cases discussed in Chapter 3, it is worthwhile to briefly examine one example in a quadrature formula of degree 1. Consider, for example, successive application of Simpson's rule along an interval

$$Q^N_{3,n} = \frac{h}{3}\sum_{i=o}^{n-3} x_i(y_i + 4y_{i+1} + y_{i+2})$$

where $x_i = 1$ if $i = 0, 2, 4, 6, \ldots$

$= 0$, otherwise

Thus, $$Q^N_{3,n} = \sum_{j=1}^{\frac{n-1}{2}} Q^N_{3j} \, . \qquad (6.12)$$

Let $$\delta^2(Q^N_{3,n}) = C \sum_{j=1}^{\frac{n-3}{2}} (Q^N_{3_{j+1}} - Q^N_{3_j})^2$$

$$= C \sum_{i=1}^{n-2} (1 - x_i) \left[4(y_{i+2} - y_i) + (y_{i+3} - y_{i-1})\right]^2 .$$

Then, choose C, such that when measurement error is present,

$$\mathcal{E}_\epsilon \delta^2(Q) = C\theta + \sigma^2 .$$

By inspection, C must be $\dfrac{1}{17(n-3)}$,

so that $$a\mathcal{E}_\epsilon \delta^2(Q_{3,n}) = \left[\frac{a\theta}{17(n-3)}\right] + a\sigma^2 .$$

Then, expanding θ in terms of $\underline{\Delta}$, one obtains,

$$\theta = \sum_{i=1}^{n-2} x_i \left[\Delta^2_{i+1} - \Delta^2_{i-1} + 6(\Delta_i + \Delta_{i+1})\right]^2 . \qquad (6.13)$$

Then (6.13) can be manipulated by the methods of Section 1.4, depending on the form desired.

6.4 Least Square Residuals

As indicated earlier, residuals from least squares polynomials were examined as a possible estimator of mean square error of quadrature. These residuals can be represented as

$$e_i = y_i - \hat{y}_i \quad ,$$

where the y_i are the observations used in the formula and the $\hat{y}_i$ the corresponding point on the fitted polynomial. It is well known that

$$\mathcal{E}_\epsilon \left(\frac{1}{n-k} \right) \sum_{i=1}^{n} e_i^2 = \frac{1}{n-k} \sum \overset{*}{r}{}_i^2 + \sigma^2 , \qquad (6.14)$$

where it is recalled that the $\overset{*}{r}_i$ are defined for the fixed observation points of the formula, and may be expressed

$$\overset{*}{r}_i = \mathcal{E}\hat{y}_i - \mathcal{E}y_i \quad .$$

The difference in the $\overset{*}{r}_i$ and the r_i of the preceding sections is immediately apparent: the present remainders are fixed, and those of the supplementary observations are random. As $r(x)$ should be smaller in magnitude in the neighborhood of the points of fit, it appears that

$$\text{average} \quad \overset{*}{r}{}_i^2 < \mathcal{E}r_i^2 \ .$$

However, it was not determined that

$$\frac{1}{n-k} \sum_{i=1}^{n} \overset{*}{r}{}_i^2 < \mathcal{E}r_i^2 \ ,$$

or even that

$$\frac{1}{n-k} \sum \overset{*}{r}{}_i^2 > \left(\frac{1}{nh}\right)^2 R^2 \ ,$$

although the latter would seem likely. Moderate effort to solve these questions analytically was unsuccessful, and it is suggested that empirical examination of the latter two inequalities would be quite valuable.

In spite of these difficulties, it is quite apparent that the residual mean square is useful in estimating quadrature variance when the

r_i^{*2} are known to be dominated by σ^2 . In this case, the residual mean square can be used as an approximation to $\hat{\sigma}^2$, and in conjunction with supplementary observations to yield an estimator of $V(\hat{M})$.

6.5 Summary

The central theme of the present chapter is that supplementary observations are necessary in order to estimate quadrature errors in practical situations. Many questions, such as allocation in two stage sampling and the necessary number of supplementary observations have not been answered, and are left for future investigations. It can be remarked that this approach to the problem is of particular interest, because the designer is forced to explicitly allocate his resources to the levels of inference, estimates and measures of uncertainty. This allocation is implicit in all sampling schemes, but seldom as well defined as in the presently considered designs.

Chapter 7. SUMMARY AND CONCLUSIONS

7.1 Introduction

The general problem of sampling a continuous function over a specified closed interval is considered. The traditional approaches are: simple random sampling, stratified random sampling and systematic random sampling. The last is of particular interest because of practical considerations in field application. Traditionally, the parameter of interest is considered to be the mean ordinate; in the present treatment, the problem is oriented toward estimation of area under the curve, and numerical integration methods are investigated.

The current status of Systematic Sampling is summarized nicely in the words of Cochran (1953, p. 185),

> "Systematic samples are convenient to draw and to execute. In most of the studies they compared favorably in precision with stratified random samples. Their disadvantages are that they may give poor precision when unexpected periodicity is present and that no trustworthy method for estimating (variance) from the sample data is known."

That is, the gain in precision from systematic sampling is not reflected in the conventional variance estimators, and the procedure is occasionally dangerous if the investigator doesn't know his experimental material. In the present investigation, a moderately successful attempt was made to:

1) evaluate conventional estimators of mean ordinate and variance by methods of numerical analysis.

2) evaluate classical quadrature formulae as estimators in systematic sampling, and further exploit the interpretation of mean ordinate as area under the curve.

3) develop estimators of variance of quadrature by the methods of numerical analysis.

The experimental material was assumed to be a continuous function over a finite interval.

7.2 Estimators of Mean Ordinate (Area Under the Curve)

The estimators considered were the conventional mean observation, a number of classical quadrature formulae, and the integral of a least-squares polynomial. Some of the quadrature formulae utilize equally spaced abscissae, while others require unequal spacing. The quadrature methods apply also to observations taken at random over the interval, but this case was not treated except as a simple mean. (However, the application of least squares quadratures to random observations follows conventional theory.)

The formulae, $\hat{M}$, for estimating area under the curve are developed and evaluated in Chapters 2 and 4, and in Tables 4.1, 4.3 and 4.4 are presented quadrature formulae for small n. Considerable attention was given to the problem of choosing a formula (or design) from several eligible ones. In general, this choice can be made only if the function and interval are explicitly specified. The attempt to define general classes of functions such that one or another formula is best for a given class was unsuccessful. However, when classes of functions were defined in terms of restrictions on the difference expansion (for example, letting $\Delta^3 \equiv 0$ thereby specifying a quadratic function), useful comparisons if the formulae were possible.

Some useful conclusions regarding the estimators of mean ordinate (area under the curve) are:

1) The unequally spaced quadrature designs (e.g., Gaussian and Tchebycheff) are generally superior to the equally spaced ones, and it is abundantly clear that these should be used when practicable (Table 5.1).

2) Of the equally spaced designs, the centric designs are generally best for small n and low degree, with the Newton-Cotes designs better for high degree and large n. Selection of the better design in a particular case requires specification of the relative magnitude of σ^2.

3) The integral of the least squares polynomial is considered superior to repeated application of a fixed formula along an interval of a function, degree of formulae remaining constant. The least squares quadratures are also quite practical, as existing tables of orthogonal polynomials are readily accessible (e.g., Anderson and Houseman, 1942) and the methods are more familiar to statisticians than are the methods of finite differences.

4) It is well known (e.g., Cochran, 1953, p. 170) that in sampling a function with a linear trend (i.e., $\Delta^2 \equiv 0$), the

systematic random sample is inferior to the stratified random sample (both with conventional estimator). In Paragraph 2.7.3 it is shown that the superiority of the stratified sample is always reduced when $\Delta^2 \neq 0$, $\Delta^3 \equiv 0$, and that for sufficiently large n, systematic sampling is superior under these conditions.

5) The comparison of the systematic random sample with the centric systematic sample (both with conventional estimator) is of interest. Although M_{sr} is unbiased, and M_{sc} is not (in general), M_{sc} is unbiased for $\Delta^2 \equiv 0$, and is in this case superior to M_{sr}. Further, when $\Delta^2 \neq 0$, and $\Delta^3 \equiv 0$ (i.e., when the function is quadratic), the constant second difference must be 16 times the absolute average first difference for M_{sr} to be superior to M_{sc} (Paragraph 5.3.2). This condition implies symmetry of the function about the mid-point of the interval [o, w].

6) When two-stage sampling is employed, with simple random selection of r PSU's (occasions or intervals), and application of systematic sampling to a continuous function at the second stage, further comparisons of systematic random and systematic centric sampling are of interest (Paragraphs 5.4.2 and 5.4.3). It is shown that the superiority of M_{sc} over M_{sr} is reduced as the number of sampled PSU's increases, and is increased with the increased variation among the functions characterizing the PSU's.

7) The end-correction estimator for systematic random sampling, $Q_1(t)$, is developed as a first degree quadrature formula in Paragraph 4.6.3. It is shown (Paragraph 4.6.4) that the bias of $Q_1(t)$ is exactly double that of M_{sc}, under the restriction $\Delta^4 \equiv 0$. Further, it is apparent from Table 5.1 that, under this restriction, a number of higher degree formulae are better than $Q_1(t)$, at least for some n. Thus, there seems little justification for use of $Q_1(t)$ in the one-stage sampling case (Paragraph 5.3.1).

However, in the two stage and stratified cases, when systematic random sampling is advantageous at the second stage (or within strata) by virtue of the accuracy of the over-all estimator, one may still be interested in computing estimates for individual strata or PSU's. In this case, the formula $Q_1(t)$ will be useful, although it is likely that the more general least squares estimator will be more accurate.

7.3 Estimators of Variance (Mean Square Error)

It is demonstrated in Chapter 3 that unbiased estimators of the truncated difference expansion of mean square error of quadrature formulae can be constructed in the case of repeated application of formulae along an interval. It was also pointed out that such estimators involve information about the mean ordinate. One of the more important questions for further investigation can then be formulated: how much of the information about mean ordinate should be used in the estimator of variance?

However, in those situations in which estimators of variance of quadrature can be constructed from the differences between successive quadratures, the least squares quadratures are preferred [see 3), Paragraph 7.2], so that there would seem to be little opportunity to use the variance estimator based on quadrature differences. Exception might be made in the case of simple systematic sample estimators [e.g., Formula 3.19 and below, 1)].

The more important sources of information about mean square error, in the one stage sampling case, seem to be a) residuals from least squares fit, and b) supplementary observations. The investigation of these two sources must be considered inadequate for completely satisfactory use of the information although usable results were obtained (Chapter 6). Again, both of these sources provide information about the mean ordinate, and, again, the question arises regarding the amount of this information that should go into measurement of accuracy.

Some useful results are:

1. In the case of a systematic random sample, and estimator M_{sr}, an unbiased estimator of $V(M_{sr})|\Delta^2 = 0)$ is given by:
   a. the mean square successive difference, δ^2, <u>if $n = 6$</u> (Paragraph 3.3.1).
   b. the mean square successive difference of lag ℓ, δ_ℓ^2, if $n = 6\ell^2$, (Paragraph 3.3.2).
   c. $\delta_*^2 = (\frac{n-6}{18})\delta_2^2 + (\frac{24-n}{18})\delta_1^2$, <u>for general n</u> (Paragraph 3.3.2).
2. The squared mean alternate difference, E, (Paragraph 3.3.3) is shown to be always positively biased.

3. All of the conventional estimators considered are related, and are generated by linear combinations of squares of linear combinations of sample differences.
4. An estimator, $\overset{*}{V}$, of variance of quadrature is developed (Paragraph 6.2.4) from supplementary observations. This estimator appears to have useful properties.
5. An attempt to evaluate the use of least squares residuals was unsuccessful, but these residuals are surely useful when it is known that the r_i^{*2} are dominated by σ^2 (Section 6.4).

7.4 Suggestions for Further Investigation

7.4.1 Empirical Studies

Several empirical studies are suggested for problems encountered that do not appear analytically tractable. These are:

1. A study of least square residuals when the degree of the fitted curve is less than the true function. (Section 6.4).
2. A study of "characterizing" real populations (functions) in terms of physical features, such that the characterization can be translated into statements about the terms in the power series expansion of the function.

7.4.2 Examination of Quadrature Formulae

It was suggested in Paragraph 5.2.6 that the irregularity of the patterns among coefficients of σ^2 and R^2 (in Table 5.1) suggests that there may be higher degree quadratures with very desirable properties. For this reason, a more complete compilation of formulae would appear warranted.

In addition, a general extension of quadrature formulae in the random systematic sample would be desirable. In particular, an extension of the least squares quadratures for equally spaced observations would be most valuable, as these would provide formulae for the random systematic sample, as well as for other designs requiring equally spaced observations.

7.4.3 Decision as to Degree of Formula

Further investigation is needed with regard to the question of degree of formula required in a particular case. This question is related to the second problem of Paragraph 7.4.1, and also to the further development of least squares quadratures.

7.4.4 Characteristics of Variance Estimators

The only characteristic of the variance estimator receiving attention in this dissertation is bias. Other features of the distribution of these estimators are of practical interest and should be investigated.

7.4.5 Allocation of Resource

Attention was called on several occasions to the necessity of utilizing, in the variance estimator, part of the information regarding area under the curve. The general problem of deciding how much such information should be diverted to estimators of precision is worthy of investigation.

7.4.6 Sampling in Two Dimensions

It is quite desirable that the results of this investigation be extended to two dimensions. The cubature formulae of numerical analysis provide a direct extension of the present results to the two-dimensional problem, in which mean ordinate is best interpreted as volume under a surface. It is apparent that this is also an extension of conventional theory of response surfaces, which theory also utilizes results of classical numerical analysis.

7.4.7 Development as a Problem in Time Series

It was mentioned in Section 1.1 that the present problem could well be considered a problem in Time Series, although the methods of Conventional Time Series were not employed in this dissertation. However, it seems likely that an application of the models of Time Series Analysis to the objective of estimation of area under the curve would be rewarding.

LIST OF REFERENCES

Aitken, A. C. 1939. Statistical Mathematics. Oliver and Boyd, Edinburgh.

Anderson, R. L. and E. E. Houseman 1942. Tables of orthogonal polynomial values extended to N = 104. Res. Bull. 297. Agric. Exp. Sta., Ames, Iowa.

Booth, A. D. 1955. Numerical Methods. Butterworths Scientific Publications, London.

Cochran, W. G. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. Annals Math. Stat., 17:164-177.

Cochran, W. G. 1953. Sampling Techniques, 1st edition. John Wiley and Sons, New York.

Daniell, P. J. 1940. Remainders in interpolation and quadrature formulae. Math. Gazette 24:238-244.

Das, A. C. 1950. Two dimensional systematic sampling and associated stratified and random sampling. Sankhya 10:95-108.

Guest, P. G. 1961. Numerical Methods of Curve Fitting. Cambridge Univ. Press.

Haynes, J. D. 1948. An empirical investigation of sampling methods for an area. M. S. Thesis, North Carolina State College, Raleigh.

Kopal, Z. 1955. Numerical Analysis. John Wiley and Sons, New York.

Kunz, K. S. 1957. Numerical Analysis. McGraw-Hill Book Company, New York.

Madow, W. G. 1949. On the theory of systematic sampling, II. Annals of Math. Stat. 20:333-354.

Madow, W. G. and L. Madow 1944. On the theory of systematic sampling, I. Annals of Math. Stat. 15:1-24.

Milne, A. 1959. The centric systematic area-sample treated as a random sample. Biometrics 15:270-297.

Moran, P. A. P. 1950. Numerical integration by systematic sampling. Proc. Camb. Philos. Soc. 46:111-115.

Moran, P. A. P. 1957. Approximate relations between series and integrals. Math. Tables and Aids in Computing. 34-37.

von Neuman, J., R. H. Kent, H. R. Bellinson and B. I. Hart. 1941. The mean square successive difference. Annals of Math. Stat. 12:153-162.

Osborne, J. G. 1942. Sampling errors of systematic and random surveys of cover-type areas. Jour. Amer. Stat. Assoc. 37:256-264.

Parzen, E. 1961. An approach to time series analysis. Annals of Math Stat. 32:951-989.

Quenouille, M. H. 1949. Problems in plane sampling. Annals of Math Stat. 20:355-375.

Quenouille, M. H. 1956. Notes on bias in estimation. Biometrika 43: 353-360.

Riordan, J. 1958. An Introduction to Combinatorial Analysis. John Wiley and Sons, Inc., New York.

Simmons, G. F. 1963. Introduction to Topology and Modern Analysis. McGraw-Hill Book Company, Inc., New York

Whitaker, E. and G. Robinson, 1944. The Calculus of Observations, 4th Ed. Blackie and Son, ltd., London.

Yates, F. 1949. Systematic sampling. Phil. Trans. Roy. Soc. 241(A): 345-377.

APPENDIX

Values of B(n): Refer to Section 4.5

1. $$B(2) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

2. $$B(3) = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & -0 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}\begin{bmatrix} 2 & -3 & 1 \\ 0 & -2 & 1 \\ 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -3/2 & 1/2 \\ 0 & 2 & -1 \\ 0 & -1/2 & 1/2 \end{bmatrix}$$

3. $$B(4) = \begin{bmatrix} -1/6 & & & \\ & 1/2 & & \\ & & -1/2 & \\ & & & 1/6 \end{bmatrix}\begin{bmatrix} -6 & 11 & -6 & 1 \\ 0 & 6 & -5 & 1 \\ 0 & 3 & -4 & 1 \\ 0 & 2 & -3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -11/6 & 1 & -1/6 \\ 0 & 3 & -5/2 & 1/2 \\ 0 & -3/2 & 2 & -1/2 \\ 0 & 1/3 & -1/2 & 1/6 \end{bmatrix}$$

4. $$B(5) = \begin{bmatrix} 1/24 & & & & \\ & -1/6 & & & \\ & & 1/4 & & \\ & & & -1/6 & \\ & & & & 1/24 \end{bmatrix}\begin{bmatrix} 24 & -50 & 35 & -10 & 1 \\ 0 & -24 & 26 & -9 & 1 \\ 0 & -12 & 19 & -8 & 1 \\ 0 & -8 & 14 & -7 & 1 \\ 0 & -6 & 11 & -6 & 1 \end{bmatrix}$$

5. $$B(6) = \begin{bmatrix} -1/120 & & & & & \\ & 1/24 & & & & \\ & & -1/12 & & & \\ & & & 1/12 & & \\ & & & & -1/24 & \\ & & & & & 1/120 \end{bmatrix}\begin{bmatrix} -120 & 274 & -225 & 85 & -15 & 1 \\ 0 & 120 & -154 & 71 & -14 & 1 \\ 0 & 60 & -107 & 59 & -13 & 1 \\ 0 & 40 & -78 & 49 & -12 & 1 \\ 0 & 30 & -61 & 41 & -11 & 1 \\ 0 & 24 & -50 & 35 & -10 & 1 \end{bmatrix}$$

6. $$B(7) = \begin{bmatrix} 1/720 & & & & & & \\ & -1/120 & & & & & \\ & & 1/48 & & & & \\ & & & -1/36 & & & \\ & & & & 1/48 & & \\ & & & & & -1/120 & \\ & & & & & & 1/720 \end{bmatrix} . Z(7)$$

where

$$Z(7) = \begin{bmatrix} 720 & -1764 & 1624 & -735 & 175 & -21 & 1 \\ 0 & -720 & 1044 & -580 & 155 & -20 & 1 \\ 0 & -360 & 702 & -461 & 137 & -19 & 1 \\ 0 & -240 & 508 & -372 & 121 & -18 & 1 \\ 0 & -180 & 396 & -307 & 107 & -17 & 1 \\ 0 & -144 & 324 & -260 & 95 & -16 & 1 \\ 0 & -120 & 274 & -225 & 85 & -15 & 1 \end{bmatrix}$$

7. $$B(8) = \begin{bmatrix} -1 & & & & & & & \\ & 7 & & & & & & \\ & & -21 & & & & & \\ & & & 35 & & & & \\ & & & & -35 & & & \\ & & & & & 21 & & \\ & & & & & & -7 & \\ & & & & & & & 1 \end{bmatrix} \cdot \left(\frac{1}{5040}\right) Z\ (8)$$

where

$$
Z(8) = \begin{bmatrix}
-5040 & 13068 & -13132 & 6769 & -1960 & 322 & -28 & 1 \\
0 & 5040 & -8028 & 5104 & -1665 & 295 & -27 & 1 \\
0 & 2520 & -5274 & 3929 & -1420 & 270 & -26 & 1 \\
0 & 1680 & -3796 & 3112 & -1219 & 247 & -25 & 1 \\
0 & 1260 & -2952 & 2545 & -1056 & 226 & -24 & 1 \\
0 & 1008 & -2412 & 2144 & -925 & 207 & -23 & 1 \\
0 & 840 & -2038 & 1849 & -820 & 190 & -22 & 1 \\
0 & 720 & -1764 & 1624 & -735 & 175 & -21 & 1
\end{bmatrix}
$$

NORTH CAROLINA STATE
OF THE UNIVERSITY OF NORTH CAROLINA
AT RALEIGH

INSTITUTE OF STATISTICS

(Mimeo Series available for distribution at cost)

344. Roberts, Charles D. An asymptotically optimal sequential design for comparing several experimental categories with a standard or control. 1963.
345. Novick, M. R. A Bayesian indifference procedure. 1963.
346. Johnson, N. L. Cumulative sum control charts for the folded normal distribution. 1963.
347. Potthoff, Richard F. On testing for independence of unbiased coin tosses lumped in groups too small to use χ^2.
348. Novick, M. R. A Bayesian approach to the analysis of data for clinical trials. 1963.
349. Sethuraman, J. Some limit distributions connected with fixed interval analysis. 1963.
350. Sethuraman, J. Fixed interval analysis and fractile analysis. 1963.
351. Potthoff, Richard F. On the Johnson-Neyman technique and some extensions thereof. 1963.
352. Smith, Walter L. On the elementary renewal theorem for non-identically distributed variables. 1963.
353. Naor, P. and Yadin, M. Queueing systems with a removable service stations. 1963.
354. Page, E. S. On Monte Carlo methods in congestion problems—I. Searching for an optimum in discrete situations. February, 1963.
355. Page, E. S. On Monte Carlo methods in congestion problems—II. Simulation of queueing systems. February, 1963.
356. Page, E. S. Controlling the standard deviation by cusums and warning lines.
357. Page, E. S. A note on assignment problems. March, 1963.
358. Bose, R. C. Strongly regular graphs, partial geometries and partially balanced designs. March, 1963.
359. Bose, R. C. and J. N. Srivastava. On a bound useful in the theory of factorial designs and error correcting codes. 1963.
360. Mudholkar, G. S. Some contributions to the theory of univariate and multivariate statistical analysis. 1963.
361. Johnson, N. L. Quota Fulfilment in finite populations. 1963.
362. O'Fallon, Judith R. Studies in sampling with probabilities depending on size. 1963.
363. Avi-Itzhak, B. and P. Naor. Multi-purpose service stations in queueing problems. 1963.
364. Brosh, I. and P. Naor. On optimal disciplines in priority queueing. 1963.
365. Yniguez, A. D'B., R. J. Monroe, and T. D. Wallace. Some estimation problems of a non-linear supply model implied by a minimum decision rule. 1963.
366. Dall'Aglio, G. Present value of a renewal process. 1963.
367. Kotz, S. and J. W. Adams. On a distribution of sum of identically distributed correlated gamma variables. 1963.
368. Sethuraman, J. On the probability of large deviations of families of sample means. 1963.
369. Ray, S. N. Some sequential Bayes procedures for comparing two binomial parameters when observations are taken in pairs. 1963.
370. Shimi, I. N. Inventory problems concerning compound products. 1963.
371. Heathcote, C. R. A second renewal theorem for nonidentically distributed variables and its application to the theory of queues. 1963.
372. Portman, R. M., H. L. Lucas, R. C. Elston and B. G. Greenberg. Estimation of time, age, and cohort effects. 1963.
373. Bose, R. C. and J. N. Srivastava. Analysis of irregular factorial fractions. 1963.
374. Potthoff, R. F. Some Scheffe-Type tests for some Behrens-Fisher type regression problems. 1963.