1973

Model-free tolerancing

Thomas C. Pingel
Lehigh University

MODEL FREE TOLERANCING

by

Thomas C. Pingel

A Thesis

Presented to the Graduate Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Industrial Engineering

Lehigh University

1973

CERTIFICATE OF APPROVAL

This thesis is accepted and approved in partial fulfillment of the requirements for the degree of Master of Science.

April 17, 1973

Date

Sutton Monro

Professor in Charge

Chairman of the Department
of Industrial Engineering

ACKNOWLEDGEMENTS

The author wishes to express appreciation to:

Professor S. Monro, who served as his faculty advisor, for his advice and guidance during the preparation and writing of this thesis.

Mr. E. O. Goll and Drs. K. E. Larson and C. R. Dudgeon, Western Electric Engineering Research Center, for their technical guidance and encouragement in this endeavor.

Ms. S. Scheidemantel and Miss P. Renzo for their excellent typing.

His wife, Jeanne, without whose love, understanding, and encouragement this thesis would have been impossible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The subject of empirical optimization as applied to complex manufacturing processes is considered. As opposed to the usual empirical techniques of natural variation, evolutionary optimization, and designed experiments which depend upon the development of a mathematical model to describe the behavior of the process, the procedure developed here "optimizes" the process without such a model, which is to say it is "model free." The technique is freely adapted from the Monte Carlo variance reducing method of importance sampling.

The procedure is evaluated by simulating conditions which might be encountered in an actual production environment. Various optimizing criteria are compared and statistical behaviors are interpreted to answer the questions of when there is enough data and how good the results are.

Since the model free procedure considers the data to be a random sample of some sort and thus requires a reconstruction of its underlying population, a generalized technique based upon the deltaspline transformation is selected and evaluated for quantitative application.

The results of the study indicate that a model free approach to empirical optimization under conditions likely to be encountered in a manufacturing environment is feasible.

1.0 Introduction

The subject of this thesis is one of optimization, specifically the selection of optimal tolerances in a manufacturing process. Optimization is, by definition, a worthy goal and thus has received considerable and continuing attention over the years. This can be attested by the numerous techniques which have been developed and continue to be developed for achieving it. The topic of manufacturing tolerancing has also received considerable attention, often with respect to this worthy goal.

Tolerancing is a consideration applicable to all stages in the life cycle of a typical manufacturing process. However, "optimal" tolerancing generally is not. Consider the life cycle of such a manufacturing process. Once conceived the process is designed to meet its performance objectives in the best way known. The engineering considerations in this phase quite naturally are directed toward optimizing this performance. This includes specifications of optimal tolerances which, however determined, will affect the hardware and tooling selected for use in the final process. The next phase in this cycle typically is the construction of a pilot operation to verify the feasibility of the design. Often at this stage it is found that the design, while workable in theory, cannot meet, in practice, a minimal performance requirement and modifications are therefore made to the design. The goal of these modifications usually is not, however, the optimization of the performance, but rather just the attainment of the elusive minimum acceptable

level of performance. This may include respecification of tolerances. Once the modified design is shown capable of meeting performance requirements, the process is converted to full production status and other inadequacies may appear, thus resulting in further design changes. Again the usual goal for these changes is attainment of the minimum acceptable level of performance, and even further specification of tolerances may result. The end product is a process which is "adequate," in the sense that it meets its required performance levels, but often far from "optimal." The only optimization considerations were made in initial design and specifications, and this is likely to be the only time in its life cycle that the process will have been "optimized." After achieving full production status, the process is monitored either continuously or periodically to assure that it maintains the minimum acceptable level of performance.

Seldom is any real effort made to significantly improve upon that level. The reason for this may be that the effort required to achieve significant improvements is just not available. A team of development specialists generally accompany the process from its conception until it has achieved full production status. At this point these specialists are generally released for use in other areas. The monitoring function is then taken over by a team of production specialists whose interest and energy is spent in maintaining the necessary level of performance. The inherent complexities of today's manufacturing processes seldom leave time to the specialist for anything else--including the search for improved performance.

This is not to say that improvement in the process performance level is not desirable. Typically a measure of this performance is based on a cost/profit margin and an obvious source of improvement in performance level is then the reduction or elimination of scrap loss. This is true since scrap loss represents an increase in cost without comparable increases in profit. It is not likely that any process could operate in such a fashion that there was no scrap loss, but if the total loss is significant, say 10 to 25% of all production costs as is often the case with complex processes resulting in products such as integrated circuit chips, some effort in the direction of seeking optimal, or at least improved, conditions or tolerances can easily be justified.

Efforts aimed toward such improvement typically fall under the classification of empirical optimization analyses. Empirical optimization, as opposed to but not independent of, analytical optimization is so called because the results are based upon observation and experimentation. Several such techniques have been developed for use in a manufacturing environment. Among these are the classical designed experiment approach[19,6,14,16,10], evolutionary optimization (E.V.O.P.) [2,3,10], and the method of natural variation.[10] These and similar empirical optimization techniques most often utilize a hill climbing or gradient type of analysis in seeking the optimal process state. The major distinctions between the techniques is in how the data is collected and what is done with the gradient information. Common to all the techniques, however, is the derivation, based on the data,

of some mathematical model which can explain the variation in the observations. This model is then used to calculate the gradient, or direction, in which a process shift is likely to produce some improvement in the process level of performance. Such a model is most commonly obtained by regressing the data using some least squares curve fitting technique. The confidence put in the model is then most often dependent not only on its statistical "significance" but also its intuitive appeal.

Statisticians will generally agree that such model building qualifies in some respects as an art in that the analyst's success in deriving an accurate model depends upon a great deal more than his ability to apply cook-book procedures. Next to any theoretical understanding of the relationships between variables, experience in similar endeavors is the analyst's greatest resource. The experienced model builder has his own bag of tricks for looking at the data and uncovering reasonable relations even without a theoretical foundation.

However, even the most experienced model builder will not always be able to uncover or develop a model which can adequately explain the behavior observed in the data. Many times this may reflect on the quality of the data, but the likelihood that the model builder will fail even with good data exists and this likelihood grows rapidly as the number of variables being modeled increases. If n is this number of variables, the number of relations the model builder must consider between the variables is given by

no. of relations = $2^n - 1$ where n = no. of variables 1-1

and clearly the more relations to consider, the more difficult it becomes to model. Thus more and more expertise is required of the analyst as the dimensions of his data expand. This situation then impacts upon application of the usual empirical optimization techniques which become useless without models.

Consider now the manufacture of a complex product such as an integrated circuit chip which requires as many as 300 to 400 sequential operations to produce. Many of these operations are subject to tolerancing specifications for various product parameters and as many as 40 to 50 of these parameters may be thought to relate significantly to the quality of the end product. Though it is likely many of the parameters are highly correlated so that application of data reduction techniques may be employed to identify but a handful of the most significant parameters, it is unlikely that fewer than four, or a minimum of three parameters will remain in this handful, particularly for such a complex process. Even for just a three parameter model, however, seven relations among parameters exist: three taking the parameters one at a time; three two at a time; and one three at a time. Unless some theory is available to suggest what functional forms to use for these relations, the analyst will have to resort to his "bag of tricks" to determine just how the parameters do relate in accounting for the behavior observed in the end product.

The most useful tools for empirically uncovering these functional forms are, perhaps, plots of the data from which the analyst

can draw on past experience. But this is clearly an inadequate approach for considering three or more parameters together. A frequent standby is the hit or miss application of a number of standard forms: linear; power; log; or exponentiation. But there is no reason to suspect that these or any combination of these forms will yield a model which adequately describes the behavior observed in the data. Experience with actual integrated circuit data bears this possibility out.

It seems that a desirable state of affairs would be the availability of an empirical optimization technique which could relieve this burden of uncertainty related to model building from the analyst without sacrificing information as to the possible location of the best or optimal process state. It is in search of such a technique that this thesis was undertaken. The next section provides a formal statement of the problem considered.

1.1 Statement of the Problem

1.1.1 Mathematical

Consider a complex manufacturing process in which the end product has some series of final responses: $y_1, y_2, \ldots, y_n$. Let Y be some composite response, based on the simple responses, y_i, such that Y can be used as a criterion for measuring the process performance.

If the process parameters $x_1, x_2, \ldots, x_n$ explain the simple responses, y_i, then these same parameters will also explain Y so that the composite response Y exists as some function, h, of $x_1, x_2, \ldots, x_n$, or

$$Y = h(x_1, x_2, \ldots, x_n) = h(\underset{\rightarrow}{x}) \qquad 1\text{-}2$$

Assume the process parameters to be random variables whose distributions can be controlled with respect to their nominal operational levels so that the multivariate probability density function, f, for $\underset{\rightarrow}{x}$ is known if the respective operating levels $\mu_1, \mu_2, \ldots, \mu_n = \underset{\rightarrow}{\mu}$ are specified so that

$$\text{multivariate probability density of } \underset{\rightarrow}{x} \text{ at } \underset{\rightarrow}{\mu} \text{ is } f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu}) \qquad 1\text{-}3$$

Then since the process parameters are random variables, the composite response Y is also and it makes sense to talk about its expected value with respect to $f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})$. Thus

$$E(Y) = \int h(\underset{\rightarrow}{x}) f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu}) d\underset{\rightarrow}{x} = E(Y;\underset{\rightarrow}{\mu}) \qquad 1\text{-}4$$

so that the expected value of Y is dependent upon the vector of operating levels, $\underset{\rightarrow}{\mu}$.

The optimal process state can then be defined as that operating vector, $\underset{\rightarrow}{\mu}^*$, which maximizes 1-4, thus providing the maximum expected process performance level. The problem is to estimate $\underset{\rightarrow}{\mu}^*$ under these conditions:

1. A historical collection of data from the process is available.
2. No experiments for obtaining more data may be run.
3. Assume that no model can be found that adequately explains the behavior observed in the historical

data.

4. No theoretical basis for any relations among parameters is known.

1.1.2 Practical

Consider the manufacture of integrated circuit chips for which the series of final responses might be any number of electrical measures such as voltages, resistivities, or capacitances. Realizing that in integrated circuit manufacture, the production unit is typically a wafer which at the end of the process is converted into as many as 700 integrated circuit chips, a composite response relating to process performance might be percent chip yield per processed wafer. This composite response is then dependent on the electrical responses measured for each chip on a wafer. Assume that some set of in-process parameters has been identified as "significant" in relating to the electrical responses observed in the chips. Such parameters might be wafer thickness, surface resistivity of the wafer, etc. Then the composite response, percent chip yield per processed wafer, is some function of that set of significant in-process parameters. If these parameters are random variables--a reasonable assumption--then the composite response will be too.

The question of optimization, when related to manufacturing processes, generally implies an existing ability to control the process. Even if the process were not controllable, however, an optimization analysis could provide valuable information as to the benefits which might result from installation of control capabilities.

Since optimization is the question, the process parameters are not only assumed to be random variables, but random variables whose distributions can be controlled. Ideally if all parameters could be controlled to have zero variance, optimizing the process would be a simple matter of maximizing the response with respect to the in-process parameters. Such conditions of zero variance, however, are unattainable in a production environment and a more realistic situation would be one in which an "optimal" control capability for each of the parameters could be specified. This "optimal" control capability would represent the minimum variance that one could hope to achieve under normal process operating conditions. Thus after analytically, experimentally, or arbitrarily specifying these "optimal" capabilities, and since the shape of the distribution of the parameters is known, any specification of nominal process operating levels will determine the multivariate probability density function of the in-process parameters.

Since the percent yield per processed wafer is a random variable, one can talk about its expected value with respect to the optimally controlled process parameters. Also since the distribution of the optimally controlled process parameters depends upon the nominal operating levels, it follows that the expected percent yield per processed wafer does also so that optimizing the process can be considered as finding those nominals which maximize the expected percent yield per processed wafer. The analyst who attempts to optimize such a process is likely to encounter these situations:

1. Production personnel whose primary responsibility is meeting production quotas are not likely to submit to the analyst's request for many, if any, experimental production lots. They are reluctant to deviate from the proven norm.
2. A collection of recent and accurate historical data is available.
3. Attempts to model older sets of data from this process have failed, and the analyst cannot expect to have any better success at modeling with this new data.

What can the analyst do, under these conditions, to estimate the optimal process state?

1.2 Objective

The objective of this thesis, in light of these discussions, becomes the survey of the suitability of available techniques if any for dealing with this problem, and then if necessary the development and evaluation of an empirical optimization technique which relieves the burden of model building from the analyst's shoulders in seeking the best or most favorable operating state of a manufacturing process. This optimal state, $\underset{\rightarrow}{\mu}^*$, together with the optimal control capability of the process will then permit ready derivation of the "optimal" tolerances to be specified for future product manufacture.

2.0 A Model Free Empirical Optimization Technique

As Hammersley and Handscomb put it, "Every Monte Carlo computation that leads to quantitative results may be regarded as estimating the value of a multiple integral."[12] Recalling the mathematical statement of the problem in Chapter 1, this thesis considers the question of process optimization basically as one of evaluating such a multiple integral (equation 1-4). Unfortunately, as will be seen in this chapter, every evaluation of a multiple integral is not necessarily suited to a traditional Monte Carlo computation. Nevertheless, after disposing of currently available methods for evaluating and optimizing such an integral, a Monte Carlo inspired technique will be developed which does promise to optimize, or maximize, the expected process response under the restrictions and constraints of the stated problem.

2.1 Background

Assume for the moment that the response function, $h(\underset{\rightarrow}{x})$, is either known or has been "adequately" modeled from the historical data. If the resulting integrand of equation 1-4 permits analytic evaluation of the expected response, then classical optimization techniques may be applied to identify $\underset{\rightarrow}{\mu}^*$, the most favorable process operating condition. However not only are the functions h and f seldom simple enough to permit ready integration or evaluation of 1-4, but as discussed earlier, an "adequate" model may not be available and a more versatile technique is desired. Optimization aside, the evaluation of 1-4 then suggests a Monte Carlo estimation.

Monte Carlo estimations or methods are essentially techniques for

the experimental determination of an unknown quantity. Any number of such methods have been devised and published. They range from the simpler hit-or-miss and crude Monte Carlo to the more sophisticated variance reducing techniques such as stratified sampling, control variates, symmetrization of integrand, and importance sampling. These latter variance reducing techniques have resulted from needs for increased accuracy in the Monte Carlo estimate without increasing the number of observations to be gathered. Such techniques employ any prior knowledge of the properties of the involved functions for obtaining improved estimates. This reflects a constant theme throughout Monte Carlo texts: Whenever theory can, in part or whole, replace experiment, the uncertainty of the evaluation is decreased, hence accuracy increased.

However, the traditional application of any of the Monte Carlo methods is generally not feasible in the production environment because of the key word "experiment." Recall that production personnel are concerned primarily with meeting output quotas. Thus any intervention by an analyst to gather experimental data will only jeopardize the immediate concerns of these persons. Since the term Monte Carlo itself implies the conduct of an experiment of some sort in order to evaluate the multiple integral, the second condition in the mathematical statement of the problem seems to eliminate a Monte Carlo evaluation of 1-4. Even if a small experiment were permitted, the more commonly applied Monte Carlo applications would provide no means of optimization since several samples or experiments would be required

for this purpose.

To the author's knowledge, no technique has been published which attempts to optimize an expression such as 1-4 under all of the stated conditions. Every available technique requires either that experimental data be gathered, or that the historical data be modeled, or both. The only known work on this problem under such conditions has been done by Evans.[9] He proposes a modified Monte Carlo approach, a summary of which is given below:

1. Let N be the number of data points and n be the number of process parameters.
2. Arbitrarily partition Euclidean n-space into K mutually exclusive and exhaustive hyper-rectangular subregions such that N of these contain exactly one data point.
3. Rewrite the expected process response as given in 1-4 as

$$E(Y:\underset{\rightarrow}{\mu}) = \sum_{i=1}^{N} \int_{I_i} hf + \sum_{j=1}^{K-N} \int_{J_j} hf \qquad 2\text{-}1$$

where I_i is the subregion containing the i^{th} observation and J_j is one of the (K-N) subregions containing no data point.

4. Estimate each of $\int_{I_i} hf$ from the single known point contained in the subregion I_i and bound the error on the integration. Both $\int_{I_i} hf$ and this error are dependent on the partitioning.

5. Estimate $E(Y;\underset{\rightarrow}{\mu})$ by assuming the occupied cells to be "random," in some sense:

$$\hat{E}(Y;\underset{\rightarrow}{\mu}) = \frac{\sum_{i=1}^{N} \int_{I_i} hf}{\sum_{i=1}^{N} \int_{I_i} f} \qquad 2\text{-}2$$

6. Bound $\hat{E}(Y;\underset{\rightarrow}{\mu})$ with respect to the integration errors in the $\int_{I_i} hf$'s.

7. Estimate Var $(\hat{E}(Y;\underset{\rightarrow}{\mu}))$ which is also dependent on the partitioning.

8. "Balance" the two sources of error: the integration error and the estimate of the variance on $\hat{E}(Y;\underset{\rightarrow}{\mu})$. "Balance" means minimize the total error by modifying the partitioning of n-space and repeating these steps.

This sequence is executed for each of a list of candidate process optima and the $\underset{\rightarrow}{\mu}$ providing the largest "balanced" estimate of $E(Y;\underset{\rightarrow}{\mu})$ is selected as "best." Evans provides no numerical example, and this author tried applying the concept to a sample problem himself. The results indicated that the technique is unworkable. Evans' premise is that the two sources of error, integration and variance, each argue for contrary subregion sizes: small subregions for small integration error and large subregions for small variances. The "balancing" above represents the best compromise of these opposed requirements. The major flaw in his technique, which appears irreconcilable, is his "exact" estimator for the variance of $E(Y;\underset{\rightarrow}{\mu})$. Its form is unwieldy and for this reason is not presented here.

However, it clearly is a function not only of the number and size of the occupied subregions, but also on the number of unoccupied subregions in n-space. This dependence on the number of unoccupied subregions is such that by increasing the number of unoccupied cells, the variance estimate will eventually go, and stay, negative. This can be done without changing $\hat{E}(Y;\underset{\rightarrow}{\mu})$. Also the behavior of the variance estimate was found to be unpredictable as the shapes and sizes of subregions were changed.

The reason for such behavior, and the reason that the technique appears irreconcilable, is that trying to estimate the variance of an expression like 2-2, which is really just a weighted average of the observed responses, is not unlike trying to estimate the variance of an unknown population from but a single observation which is, of course, impossible. The weighted average of 2-2 represents but a single observation from an unknown population of dimension $n \cdot N$. Thus Evans' technique of balancing errors is doomed to failure unless an estimator of $E(Y:\underset{\rightarrow}{\mu})$ whose variance <u>can</u> be estimated is determined.

With this result and the conclusion that traditional Monte Carlo applications can not be employed to solve the problem addressed by this thesis, it appears, perhaps, that Monte Carlo is not the avenue to pursue. Before completely abandoning a Monte Carlo approach, recall that the reason traditional Monte Carlo methods are not applicable is that they require an experiment of some sort for evaluation of an integral such as 1-4. Even of more concern was the implication that for an optimization analysis via Monte Carlo, numerous samples

or experiments need be run. This is because the more commonly known Monte Carlo methods would require the analyst to sample from the distribution $f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})$, and to thus optimize, samples would be taken for as many $\underset{\rightarrow}{\mu}$ as possible. One dissimilar and often misunderstood Monte Carlo method exists in the literature which estimates an integral of the type 1-4 by sampling not from $f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})$ but rather from some seemingly unrelated distribution. This is the technique known as importance sampling. Its characteristics can be exploited to deal with the problem addressed by this thesis, i.e., a model free estimation of the optimal process state based on a collection of historical data.

2.2 Importance Sampling

Consider evaluation of the integral given here:

$$X = \int \theta(\underset{\rightarrow}{x})\,\phi(\underset{\rightarrow}{x})\,d\underset{\rightarrow}{x} \qquad 2\text{-}3$$

Let $\gamma(x)$ be any function such that

1. $\gamma(\underset{\rightarrow}{x}) \geq 0$ 2-4.1
2. $\int \gamma(\underset{\rightarrow}{x})\,d\underset{\rightarrow}{x} = 1$ 2-4.2
3. If $\gamma(\underset{\rightarrow}{x}) = 0$, then $\phi(\underset{\rightarrow}{x}) = 0$ 2-4.3

Then $\gamma(\underset{\rightarrow}{x})$ is a probability density function and 2-3 can be rewritten as

$$X = \int \left[\frac{\theta(\underset{\rightarrow}{x})\phi(\underset{\rightarrow}{x})}{\gamma(\underset{\rightarrow}{x})}\right]\gamma(\underset{\rightarrow}{x})\,d\underset{\rightarrow}{x} \qquad 2\text{-}5$$

Then instead of the usual sampling from $\phi(\underset{\rightarrow}{x})$ to estimate X, the analyst samples rather from the distribution defined by $\gamma(\underset{\rightarrow}{x})$ and obtains as the importance sampling estimate, I, of X:

$$I = \sum_{i=1}^{N} \left[\frac{\theta(\underset{\rightarrow}{x}^{(i)})\phi(\underset{\rightarrow}{x}^{(i)})}{\gamma(\underset{\rightarrow}{x}^{(i)})} \right] \Big/ N; \qquad 2\text{-}6$$

$$\left\{ \underset{\rightarrow}{x}^{(1)}, \underset{\rightarrow}{x}^{(2)}, \ldots, \underset{\rightarrow}{x}^{(N)} \right\} \text{ , a random sample from } \gamma(\underset{\rightarrow}{x})$$

As for other Monte Carlo estimators, the variance on I is calculable as

$$\mathrm{Var}(I) = \mathrm{Var}\left[\frac{\sum_{i=1}^{N} \frac{\theta(\underset{\rightarrow}{x})\phi(\underset{\rightarrow}{x})}{\gamma(\underset{\rightarrow}{x})}}{N} \right] = \frac{1}{N} \mathrm{Var}\left[\frac{\theta(\underset{\rightarrow}{x})\phi(\underset{\rightarrow}{x})}{\gamma(\underset{\rightarrow}{x})} \right] \qquad 2\text{-}7$$

The statistic I is unbiased and, for proper choice of $\gamma(\underset{\rightarrow}{x})$, will have a smaller variance than the crude Monte Carlo estimate obtained simply by sampling $\phi(\underset{\rightarrow}{x})$. Clearly if $\gamma(\underset{\rightarrow}{x})$ is chosen not only to satisfy conditions 2-4, but also so that $\gamma(\underset{\rightarrow}{x})$ "mimics" the product $\theta(\underset{\rightarrow}{x})\phi(\underset{\rightarrow}{x})$, then $\mathrm{Var}\left[\frac{\theta(\underset{\rightarrow}{x})\phi(\underset{\rightarrow}{x})}{\gamma(\underset{\rightarrow}{x})} \right]$ will be less than $\mathrm{Var}\,(\theta(\underset{\rightarrow}{x}))$ which is proportional to the variance of the crude Monte Carlo estimate. The importance sampling estimate of X will thus be more accurate than the crude Monte Carlo estimate. This is the variance reducing Monte Carlo technique of importance sampling.

While importance sampling was developed as a powerful variance reducing technique, its characteristics give it a versatility not inherent in other Monte Carlo methods. This is due to the analyst's ability to sample from an apparently unrelated distribution. While

it has been suggested how careful selection of the sampling distribution, $\gamma(\underset{\rightarrow}{x})$, can significantly reduce the uncertainty of the crude Monte Carlo estimate, it is just as true that careless selection of a sampling distribution, $\gamma(\underset{\rightarrow}{x})$, will produce an importance sampling estimate with far greater uncertainty than that of the crude estimate. Consequently there must be any number of sampling distributions which will have an intermediate effect on the variance of the importance sampling estimate. With this in mind, attention is re-directed to the integral which is the subject of this thesis, equation 1-4.

Let $g(\underset{\rightarrow}{x})$ be some abritrary probability density function whose effect on the variance of the importance sampling estimate of $E(Y;\underset{\rightarrow}{\mu})$ is unknown. Assume that $g(\underset{\rightarrow}{x})$ is never zero when $f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})$ is not zero. Then 1-4 can be rewritten as

$$E(Y;\underset{\rightarrow}{\mu}) = \int h(\underset{\rightarrow}{x}) f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu}) d\underset{\rightarrow}{x} = \int \left[\frac{h(\underset{\rightarrow}{x}) f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})}{g(\underset{\rightarrow}{x})} \right] g(\underset{\rightarrow}{x}) d\underset{\rightarrow}{x} \qquad 2\text{-}8$$

Such an expression allows the analyst to estimate the expected process response at any number of operating state vectors, $\underset{\rightarrow}{\mu}$, while each time sampling but from the distribution defined by $g(\underset{\rightarrow}{x})$. In fact, a single process sample consisting of the observations $\underset{\rightarrow}{x}^{(1)}, \underset{\rightarrow}{x}^{(2)}, \ldots, \underset{\rightarrow}{x}^{(N)}$ can be used to estimate the behavior of $E(Y;\underset{\rightarrow}{\mu})$ for any or all $\underset{\rightarrow}{\mu}$'s in a particular region of interest. The importance sampling estimate of $E(Y;\underset{\rightarrow}{\mu})$ becomes

$$\hat{E}(Y;\underset{\rightarrow}{\mu}) = \sum_{i=1}^{N}\left[\frac{h(\underset{\rightarrow}{x}^{(i)})f(\underset{\rightarrow}{x}^{(i)};\underset{\rightarrow}{\mu})}{g(\underset{\rightarrow}{x}^{(i)})}\right] \Big/ N = I(\underset{\rightarrow}{\mu}) \qquad 2\text{-}9$$

A measure of the uncertainty of such an estimate is essential. From 2-7, it follows that

$$\mathrm{Var}(I(\underset{\rightarrow}{\mu})) = \frac{1}{N}\mathrm{Var}\left[\frac{h(\underset{\rightarrow}{x})f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})}{g(\underset{\rightarrow}{x})}\right] \qquad 2\text{-}10$$

But $\mathrm{Var}\left[\frac{h(\underset{\rightarrow}{x})f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})}{g(x)}\right]$ is estimable from the sample observations as

$$\hat{\mathrm{Var}}\left[\frac{h(\underset{\rightarrow}{x})f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})}{g(x)}\right] =$$

$$\frac{\sum_{i=1}^{N}\left[\frac{h(\underset{\rightarrow}{x}^{(i)})f(\underset{\rightarrow}{x}^{(i)};\underset{\rightarrow}{\mu})}{g(\underset{\rightarrow}{x}^{(i)})}\right]^2 - N\left[\frac{h(\underset{\rightarrow}{x}^{(i)})f(\underset{\rightarrow}{x}^{(i)};\underset{\rightarrow}{\mu})}{g(\underset{\rightarrow}{x}^{(i)})}\right]^2}{N-1} \qquad 2\text{-}11$$

so that

$$\hat{\mathrm{Var}}(I(\underset{\rightarrow}{\mu})) = \frac{1}{N}\hat{\mathrm{Var}}\left[\frac{h(\underset{\rightarrow}{x})f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})}{g(\underset{\rightarrow}{x})}\right] = S_I^2(\underset{\rightarrow}{\mu}) \qquad 2\text{-}12$$

The expression in 2-9 is unbiased for any $\underset{\rightarrow}{\mu}$, and it follows that a good estimate of the most favorable operating condition, $\underset{\rightarrow}{\mu}^*$, might

be that vector, $\hat{\underset{\rightarrow}{\mu}}^*$, which maximizes 2-9. The confidence in the estimate $I(\hat{\underset{\rightarrow}{\mu}}^*)$ could then be determined from $S_I^2(\hat{\underset{\rightarrow}{\mu}}^*)$. Perhaps a better estimate of the optimal state $\underset{\rightarrow}{\mu}^*$ would be obtained for a criterion more general than simply maximizing 2-9. Such a criterion might be to estimate $\underset{\rightarrow}{\mu}^*$ as that vector, $\hat{\underset{\rightarrow}{\mu}}^*$, which maximizes

$$I(\underset{\rightarrow}{\mu}) - k\sqrt{S_I^2(\underset{\rightarrow}{\mu})};\ k\geq 0 \qquad 2\text{-}13$$

thus incorporating the uncertainty of the estimate, $I(\underset{\rightarrow}{\mu})$, into the result. Such an expression is equivalent to the low limit on a one-sided confidence interval for $E(Y;\underset{\rightarrow}{\mu})$ with the degree of confidence dependent not only on the magnitude of k but also the distribution of the statistic, $I(\underset{\rightarrow}{\mu})$. Thus it appears possible to estimate $\underset{\rightarrow}{\mu}^*$ without modeling or regressing the sampled values, and "optimization" is achieved with but a single sample. The actual maximization of 2-13 might be performed analytically if the multivariate probability density function, $f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})$, is not too complicated. Whatever the nature of $f(\underset{\rightarrow}{x};\underset{\rightarrow}{\mu})$, $\hat{\underset{\rightarrow}{\mu}}^*$ could be found via classic numerical search methods.

Some comment on the confidence interval defined by 2-13 is warranted. With the degree of confidence dependent not only on k but also the distribution of $I(\underset{\rightarrow}{\mu})$, it should be clear that as $\underset{\rightarrow}{\mu}$ is varied, the shape of the distribution of $I(\underset{\rightarrow}{\mu})$ also changes so that even the most carefully selected k will not provide the same degree of confidence everywhere. Thus setting $k=k_o$ will not assure the analyst of uniform confidence intervals at all $\underset{\rightarrow}{\mu}$. This can be expressed as

$$\text{Prob}\left[E(Y;\underset{\rightarrow}{\mu}) > I(\underset{\rightarrow}{\mu}) - k_o\sqrt{S_I^2(\underset{\rightarrow}{\mu})}\right] \neq \text{Prob}\left[E(Y;\underset{\rightarrow}{\mu}') > I(\underset{\rightarrow}{\mu}') - k_o\sqrt{S_I^2(\underset{\rightarrow}{\mu}')}\right] \qquad 2\text{-}14$$

$$\text{if } \underset{\rightarrow}{\mu} \neq \underset{\rightarrow}{\mu}'$$

The consequences of such a situation can only be surmised at this time, and the author's guess is that the situation should cause no great concern. Large k will still give high, and small k low degrees of confidence and the change in confidence associated with k_o as $\underset{\rightarrow}{\mu}$ is varied through a "reasonable" region of interest is likely to be of no significance. Such a "reasonable" region might be those areas where the sample values are not too sparse. If this is the case, then maximizing 2-13 for $k=k_o$ within that "reasonable" region is not unlike minimizing possible regret.

This concept of utilizing the observations of a single sample in different calculations as has been done here by considering different $\underset{\rightarrow}{\mu}$ is not novel. Such an idea was suggested by Clark in specific reference to importance sampling[4] and appears as the central point in Evans' multiplex sampling which employs both control variate and importance sampling concepts.[7,8] However, neither the suggestions by Clark, nor the techniques of Evans, nor the development thus far in this thesis is capable of dealing with the problem as stated in section 1.1. There still is an urgent need for the analyst to obtain that one experimental sample taken from the distribution $g(\underset{\rightarrow}{x})$.

Recall that the constraints of the problem restrict the analyst to but a single collection of historical data and do not permit his further experimentation. Why not consider this historical collection a kind of "random sample" and let that data represent the "results" of some past "experiment?" If a population can be hypothesized or constructed from which the historical data, or "sample values," might comprise a representative sample, then the techniques developed above can be applied and an estimate of the optimal process state made. Considering the nature of historical data, however, finding such a population may be a difficult task. These difficulties, however, are not of immediate concern. What is important at this point is that once such a population is found, the historical data which was gathered in the usual production environment without disrupting normal operations provides the analyst with the necessary "experimental" data for employing the optimizing criterion of 2-13 without further experimentation. Thus the analyst is able to estimate the optimal process state from historical data without further experimentation and without modelling the data.

This completes development of the procedures for a model free empirical optimization techniques as required by the problem statement and objectives of this thesis. This procedure is summarized in the next section.

2.3 Analysis Procedure

The steps in the analysis procedure developed in preceding sections can be summarized as follows:

1. Collect as much historical process data as possible;
2. Hypothesize or construct a distribution, $g(\underset{\rightarrow}{x})$, for which the historical data constitute a representative random sample;
3. Identify the optimal process control capability, $f(\underset{\rightarrow}{x},\underset{\rightarrow}{\mu})$;
4. Select $k \geq 0$ for use in the optimization criterion, 2-13;
5. Estimate $\underset{\rightarrow}{\mu}^*$ by $\hat{\underset{\rightarrow}{\mu}}^*$ where $\hat{\underset{\rightarrow}{\mu}}^*$ maximizes 2-13;
6. Determine new "optimal" tolerances from $f(\underset{\rightarrow}{x};\hat{\underset{\rightarrow}{\mu}}^*)$;
7. Implement the recommended nominal operating state, $\hat{\underset{\rightarrow}{\mu}}^*$, and specify the new tolerances;
8. Begin collecting data again.

While the theory and procedure for this model free approach to empirical process optimization are now complete and appear rather straightforward, several questions are raised. First, how will the not so easy task of constructing the population $g(\underset{\rightarrow}{x})$ be done? This is vital to the procedure developed and is the subject of Chapter 5 in this thesis.

Also, given that this population is, in fact, recoverable, what kind of results can be expected from such an analysis? This is a situation in which the analyst takes and analyzes historical data as if it were the result of some carefully designed experiment to utilize a powerful variance reducing technique. Quite the contrary. While importance sampling was introduced and adopted for a Monte Carlo

estimation of 1-4 because of its flexibility in allowing the analyst to perform a maximization from but a single sample, this flexibility really only allows the analyst to manipulate data with which he is "stuck" and not free to augment. Given the freedom to experiment, the analyst could select a $g(\underset{\rightarrow}{x})$ to reduce the variance of his estimate, but the effects of the constructed $g(\underset{\rightarrow}{x})$ may have quite the opposite effect.

Lastly, is the criterion 2-13 adequate, or is there some potentially better estimator of $\underset{\rightarrow}{\mu}^*$?

These last two questions are the subject of the study in Chapter 3.

3.0 A Study of the Model Free Empirical Optimization Technique

The intent of this study is to gain insight into what kind of results can be obtained by the analyst when the concepts of importance sampling as previously presented are applied to the question of identifying the most favorable operating state in a manufacturing process. For the sake of simplicity, and because results can easily be visualized, the study will focus on a process whose response Y can be characterized by a single process parameter x. It is hoped that proper interpretation of the results will permit ready generalization to the case of n process parameters.

To satisfy what will be seen to be a need for a large number of process samples, simulated data and responses will be used. Simulated data has the property that the mechanisms underlying its generation are known thus providing valuable information in the analysis of the results. Also simulated data allows the analyst to control or eliminate unwanted effects and concentrate on what are thought to be the dominant effects--a situation which is invaluable in developing and evaluating a new analysis technique.

3.1 Preliminary Investigation

The importance sampling estimate of the expected value of the response Y with respect to the process operating mean, $\underset{\rightarrow}{\mu}$, was given by 2-9 and is written here for the case of a single process parameter;

$$I(\mu) = \sum_{i=1}^{N} \left[\frac{h(x^{(i)}) f(x^{(i)};\mu)}{g(x^{(i)})} \right] \Big/ N = \sum_{i=1}^{N} \left[\frac{Y^{(i)} f(x^{(i)};\mu)}{g(x^{(i)})} \right] \Big/ N \qquad 3\text{-}1$$

an estimate of its uncertainty, as a result of 2-19, is

$$\hat{Var}(I(\mu)) = \frac{1}{N} \hat{Var} \frac{h(x)f(x;\mu)}{g(x)} = S_I^2(\mu) \qquad 3\text{-}2$$

where the degree to which $g(x)$ mimics the product $Y \cdot f(x;\mu)$ will determine the magnitude of $S_I^2(\mu)$.

Preliminary investigation of 3-1 and 3-2 exhibited quite large estimates of the uncertainty, and also frequent estimates of the expected process response larger than any of the observed values in the data. A result such as this could be unsettling. Consider the example mentioned earlier of integrated circuit manufacture for which the response Y was the percent chip yield per processed wafer. The use of criterion 2-13 for small values of k could result in an estimate, $\hat{\mu}^*$, of the optimal process state for which $I(\hat{\mu}^*)$ were 115% which is, of course, ridiculous (Figure 3.1). One can argue that it is not with $I(\hat{\mu}^*)$ that the analyst is concerned, but only with $\hat{\mu}^*$, the best estimate of where the optimal operating state is, and that equation 3-1 is only a vehicle for obtaining that estimate. However the analyst will be hard pressed to answer the production supervisor if asked: "What will my process response be with these new operating means?"; particularly if $I(\hat{\mu}^*)$ were 115%. Rather than reply

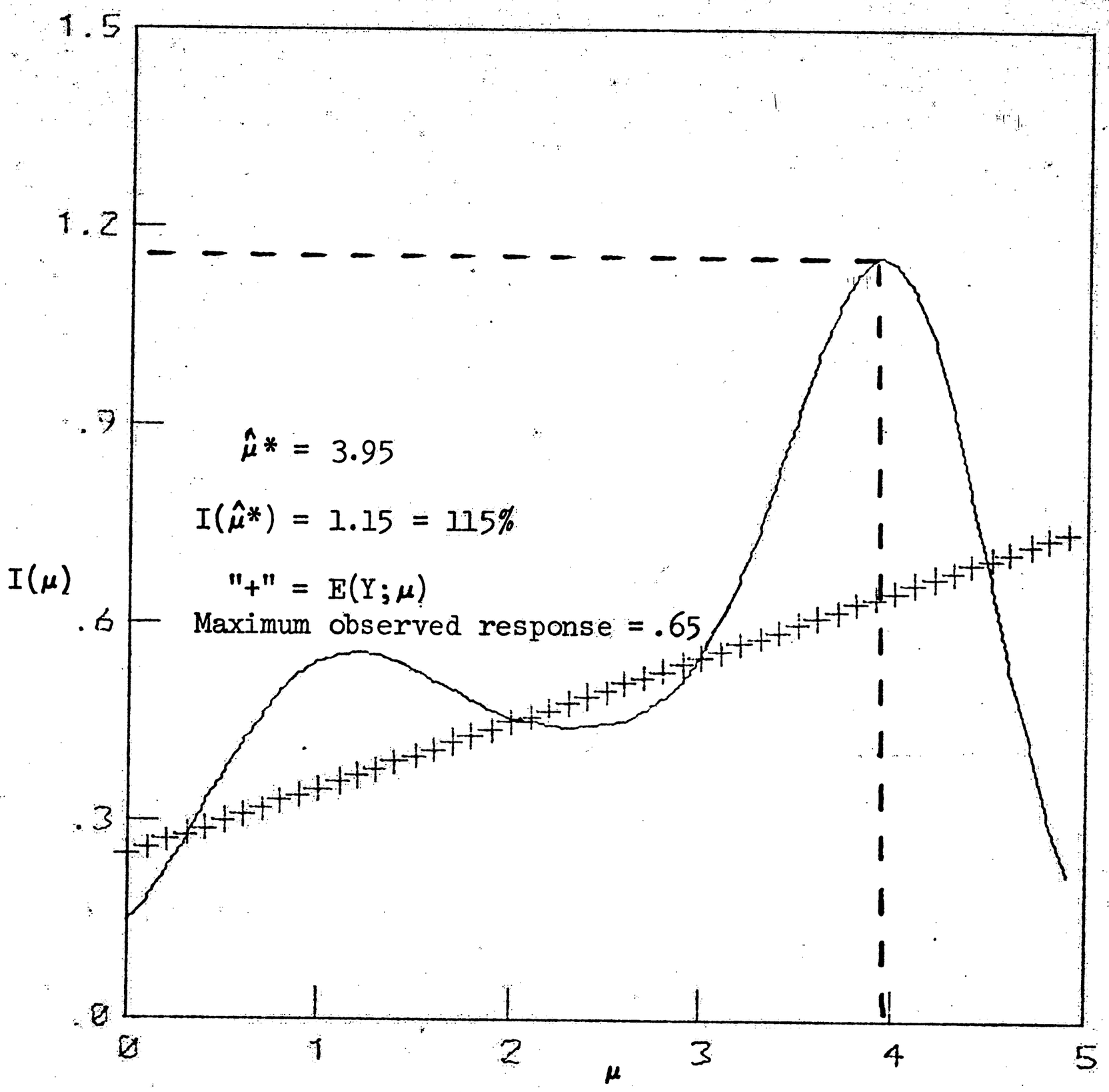


Figure 3.1 - Estimates of $E(Y;\mu)$ greater than maximum observed response

"115%" analyst might reply, appealing to equation 2-13, "I am about 90% certain that it will be at least y%." But because of possibly large values of $S_I^2(\mu)$, y% might be lower than the smallest yield ever observed in the process. "For this," the production supervisor would say, "you need a sophisticated analysis?"

To circumvent these possibilities, a new statistic, $W(\mu)$, which is based on importance sampling concepts is presented. It is this simple weighted average of the observed responses:

$$W(\mu) = \sum_{i=1}^{N}\left[\frac{Y^{(i)}f(x^{(i)};\mu)}{g(x^{(i)})}\right] \Bigg/ \sum_{i=1}^{N}\left[\frac{f(x^{(i)};\mu)}{g(x^{(i)})}\right] = E(Y;\mu) \qquad 3\text{-}3$$

where the functions f and g are as before. A similar statistic which was first suggested by Handscomb,[13] later studied by Powell and Swan,[17] and then generalized by Halton[11] appears in a Monte Carlo variance reducing technique known as "weighted uniform sampling." The statistic $W(\mu)$ as defined in 3-3 is actually derivable from these weighted uniform sampling concepts. Consider the integral of equation 1-4. It can be rewritten as

$$E(Y;\mu) = \int\left[\frac{h(x)f(x;\mu)}{f(x;\mu)}\right] f(x;\mu)dx \qquad 3\text{-}4$$

The weighted uniform sampling estimate of Powell and Swan then becomes

$$E_p(Y;\mu) = \frac{\sum_{i=1}^{N} h(x^{(i)})f(x^{(i)};\mu)}{\sum_{i=1}^{N} f(x^{(i)};\mu)} \qquad 3\text{-}5$$

where it is assumed the $x^{(i)}$ are uniformly distributed. Halton generalizes the technique so that the $x^{(i)}$ may be sampled from any distribution $g(x)$. The weighted estimate then becomes

$$E(Y;\mu) = \frac{\sum_{i=1}^{N} \left[\frac{h(x^{(i)} f(x^{(i)};\mu)}{g(x^{(i)})} \right]}{\sum_{i=1}^{N} \left[\frac{f(x^{(i)};\mu)}{g(x^{(i)})} \right]} \qquad 3\text{-}6$$

which is simply $W(\mu)$! Halton concludes that such a statistic is a biased estimate of $E(Y;\mu)$ converging to $E(Y;\mu)$ "in quadratic mean, in probability, and with probability 1."[11]

What this last statement means is that while $E(I(\mu)) = E(Y;\mu)$, it is not the case at least for any finite sample, that $E(W(\mu)) = E(Y;\mu)$. In fact, determination of its expected value for any finite sample is, presently, intractable. Also while $Var(I(\mu))$ was readily estimable, $Var(W(\mu))$ is not because of the nature of a weighted average. The results of neither Powell and Swan nor Halton indicate how the variance of $W(\mu)$ compares to $I(\mu)$, but it is hoped this study will reveal what can be expected. What, on the surface, appear to be the advantages of $W(\mu)$? First, it is a property of weighted averages that

$$\min_i Y^{(i)} \leq W(\mu) \leq \max_i Y^{(i)} \qquad 3\text{-}7$$

so that $W(\mu)$ is bounded by the minimum and maximum observed responses. Thus the analyst will not be presented with ridiculous estimates of

$E(Y; \mu^*)$ from an optimizing criterion based on $W(\mu)$. Also in preliminary investigations the statistic $W(\mu)$ appeared to yield better and more consistent estimates of μ^* than criteria based on 2-13 which employ $I(\mu)$. Thus as an alternative to those already presented, a challenger optimizing criterion is offered here. The new estimate, $\hat{\mu}^*$ of μ^* will be chosen so that

$$W(\hat{\mu}^*) = \max_{\mu} W(\mu) \qquad 3\text{-}8$$

Consequently, the study takes the direction of a comparative investigation of the behavior of the statistics $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$ and their ability to predict the optimal process state. The data will be simulated under what are hoped to representative process conditions.

3.2 Factors To Be Considered

In order to emulate representative process conditions it is first necessary to identify what factors may affect the behavior of the statistics $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$ and their predictive capabilities. From equations 3-1, 3-2, and 3-3 it is seen that this behavior is solely dependent upon:

1. $g(x)$, the probability density function of the distribution of values for the process parameter, x, as observed in the historical data;
2. $f(x;\mu)$, the probability density function for the distribution of values for the process parameters, x, when operating under optimal control at mean μ;

3. Y or h(x), the response variable and its functional form.

3.2.1 The Data and g(x)

Historical data as taken from a manufacturing process exhibits any number of peculiar traits. One of these might be the lack of statistical "niceness." The values recorded for any process parameter will rarely be representative of one single mechanism at work in the process. A typical set of data will be the result of engineers and production supervisors continual tinkering to realize marginal production increases or to correct an apparent deficiency. Consequently the data is likely to cover a wide range, but yet be clustered in two, three or more regions within that range. Thus frequency plots, or histograms, may indicate multiple modes, severe skewness, or any number of other shapes differing from a simple normal, lognormal, or "nice" unimodal symmetric distribution. Such behavior is depicted in Figure 3.2 showing actual integrated circuit data. In order that this study be made under conditions such as the these, the data to be analyzed will be simulated not from "nice" populations, but rather from either of a bimodal or skewed distribution. The actual populations selected for the study are

$$g_1(x) = .5\phi(1.5,.25) + .5\phi(3.0,.36) \qquad 3\text{-}9$$

VALUE	FREQUENCY	DISTRIBUTION
0.00	0	
0.02	0	
0.04	1	X
0.06	7	XXXXXXX
0.08	24	XXXXXXXXXXXXXXXXXXXXXXXX
0.10	20	XXXXXXXXXXXXXXXXXXXX
0.12	20	XXXXXXXXXXXXXXXXXXXX
0.14	16	XXXXXXXXXXXXXXXX
0.16	6	XXXXXX
0.18	1	X
0.20	4	XXXX
0.22	1	X
0.24	1	X
0.26	0	
0.28	2	XX
0.30	1	X
0.32	0	
0.34	0	
0.36	0	
0.38	1	X
0.40	0	

Mean------- 0.1195

Variance--- 0.0030

Figure 3.2 - Histogram of actual integrated circuit data

which is bimodal, reflecting the presence of two mechanisms at work during collection of the data with each contributing 50% of the observations, and

$$g_2(x) = .12\phi(1.5,.36) + .15\phi(2.0,.25)$$
$$+ .19\phi(2.5,.16) + .24\phi(3.0,.09) \qquad 3\text{-}10$$
$$+ .30\phi(3.5,.04)$$

which is skewed reflecting five mechanisms with increasing likelihood and decreasing variance. In each of these, $\phi(\mu,\sigma^2)$ is the probability density function for a normally distributed random variable with mean μ and variance σ^2.

Another question regarding the data is how much data the analyst will have to work with. For some processes, gathering data may be no chore and the analyst may, for all practical purposes, have as much data as he wants. For other processes, such as integrated circuit manufacture, the production cycle may be as long as a few months and the chances of obtaining many more than one hundred observations may be practically zero. Since the variance of the estimate of the expected response can be made as small as is necessary by simply increasing sample size (see equation 2-12), interest lies

not in the analysis of processes from which very large samples can easily be gathered, but rather in processes for which a collection of 100 observations can be considered a luxury.

Consequently the study focuses on the behavior of the statistics $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$ for simulated process samples of 100 observations taken either from $g_1(x)$ or $g_2(x)$. As an indication of how "representative" of $g_1(x)$ or $g_2(x)$ the sample is, the observations will be subjected to a two tailed Kolmogorov-Smirnov test to the .80 level of significance. Those samples failing the test will not be considered for analysis.

3.2.2 The Optimal Control Capability: $f(x;\mu)$

A reasonable assumption for any variable which can be controlled is that it will exhibit normal behavior when under control. Thus $f(x;\mu)$ is assumed to be the probability density function of a normally distributed random variable. One of the pre-analysis jobs the analyst has is determination of the variance of the control distribution. For this study, that determination will be made with respect to the widths of the functions $g_1(x)$ and $g_2(x)$. Clearly the variance on the controlled x will be less than the variability of the g(x)'s or the process would have been operating under better than optimal control while the data was gathered. It seems reasonable to assume that if the data were gathered over any length of time, the width or variance

of the controlled probability density function could be set at no more than half the width of the observed distribution of the data. For this reason, and with respect to the g(x)'s of equations 3-9 and 3-10, $f(x;\mu)$ will be the probability density function of a normally distributed random variable with variance no greater than .25, or

$$f(x;\mu) = \phi(x;\mu, \sigma_f^2);\ \sigma_f^2 \leq .25 \qquad 3\text{-}11$$

The study will concentrate on the case where $\sigma_f^2 = .25$. However effects on the statistics $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$ obtained by decreasing σ_f^2 toward zero will also be investigated.

3.2.3 The Process response, Y

The process response, Y, where

$$Y = h(x) \qquad 3\text{-}12$$

can be expected to take on any number of shapes in practice and the behavior of the statistics $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$ will depend not only on g(x) and $f(x;\mu)$, but also on the form of h(x). At the risk of oversimplification, the study will concern itself only with responses which are linear, or at worst quadratic, in form. In spite of all the potential forms, limiting the study to these forms is not without justification if the response is at all well behaved. Consider a process which has been operating in the vicinity of its optimal state. One would expect the observations to reflect a single peak such as the quadratic form provides. On the other hand if the processes

were operating in a region away from the optimal state when the data was collected, then the response quite possibly will not exhibit any peaks and a linear response might not be a bad approximation. Referring again to the integrated circuit example, the response simulated in the study will be percent chip yield per processed wafer so that $0. \leq Y \leq 1.0$.

It can be argued that with real data, one will never observe responses which are entirely free from either or both of measurement error and random fluctuations and that a better representation of the process response would be

$$Y = h(x) + \varepsilon(\sigma_\varepsilon^2) \qquad 3\text{-}13$$

where ε is a normally distributed random variable with mean zero and variance σ_ε^2. To better approximate these real conditions, the yields to be simulated in the study will inculde an error term. The actual functions which will be simulated are;

$$\text{quadratic: } Y = .80 - .08(x - 2.5)^2 + \varepsilon(\sigma_\varepsilon^2) \qquad 3\text{-}14$$

$$\text{linear, decreasing: } Y = .75 - .1x + \varepsilon(\sigma_\varepsilon^2) \qquad 3\text{-}15$$

$$\text{linear, increasing: } Y = .25 + .1x + \varepsilon(\sigma_\varepsilon^2) \qquad 3\text{-}16$$

The asymmetry of the considered distributions of data, i.e., $g_1(x)$ and $g_2(x)$, is responsible for the presence of two linear forms. The study will therefore consider the behavior of $W(\mu)$, $I(\mu)$,

$S_I^2(\mu)$ and the predictors of μ^* with respect to equations 3-14 - 3-16 and variances on the error component ranging from zero upwards.

3.3 Design of an Experiment

The investigation of the behavior of $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$ and their predictive behavior will be done in three phases. The conditions under which the data will be simulated are those discussed in the preceding section. To reiterate:

1. $g(x)$, the distribution of the process parameter, will be either bimodal or skewed (equation 3-9 and 3-10);
2. $f(x;\mu)$, the optimal control capability will be normal with variance σ_f^2. This variance will be ranged downward from 0.25; and
3. Y, the process response will be either quadratic or one of two linear forms. The response will contain a random component, ε, whose variance, σ_ε^2, will be considered from zero upward.

The first phase of the experiment investigates the behavior of $W(\mu)$, $I(\mu)$, $S_I^2(\mu)$, and the candidate optimization criteria as functions of the combinations of responses and distributions of the data. For this phase of the experiment σ_f^2 is held fixed at 0.25 and the random component of the response is held identically to zero for all sample responses. The two functional forms selected for $g(x)$ (equations 3-9 and 3-10) and the three response relations

(equations 3-14 - 3-16) give rise to a six cell qualitative factorial design:

DATA

RESPONSES	Bimodal: $g_1(x)$	Skewed: $g_2(x)$
quadratic		
linear, decreasing		
linear, increasing		

$\sigma_f^2 = .25$

$\varepsilon = 0$ for all observations

Figure 3.3

Each cell will consist of eight sample replications consisting of 100 observations each. Pictorial representation of the function forms used in each cell are presented in Figure 3.4. Of primary interest in each replication are the estimates of the optimal

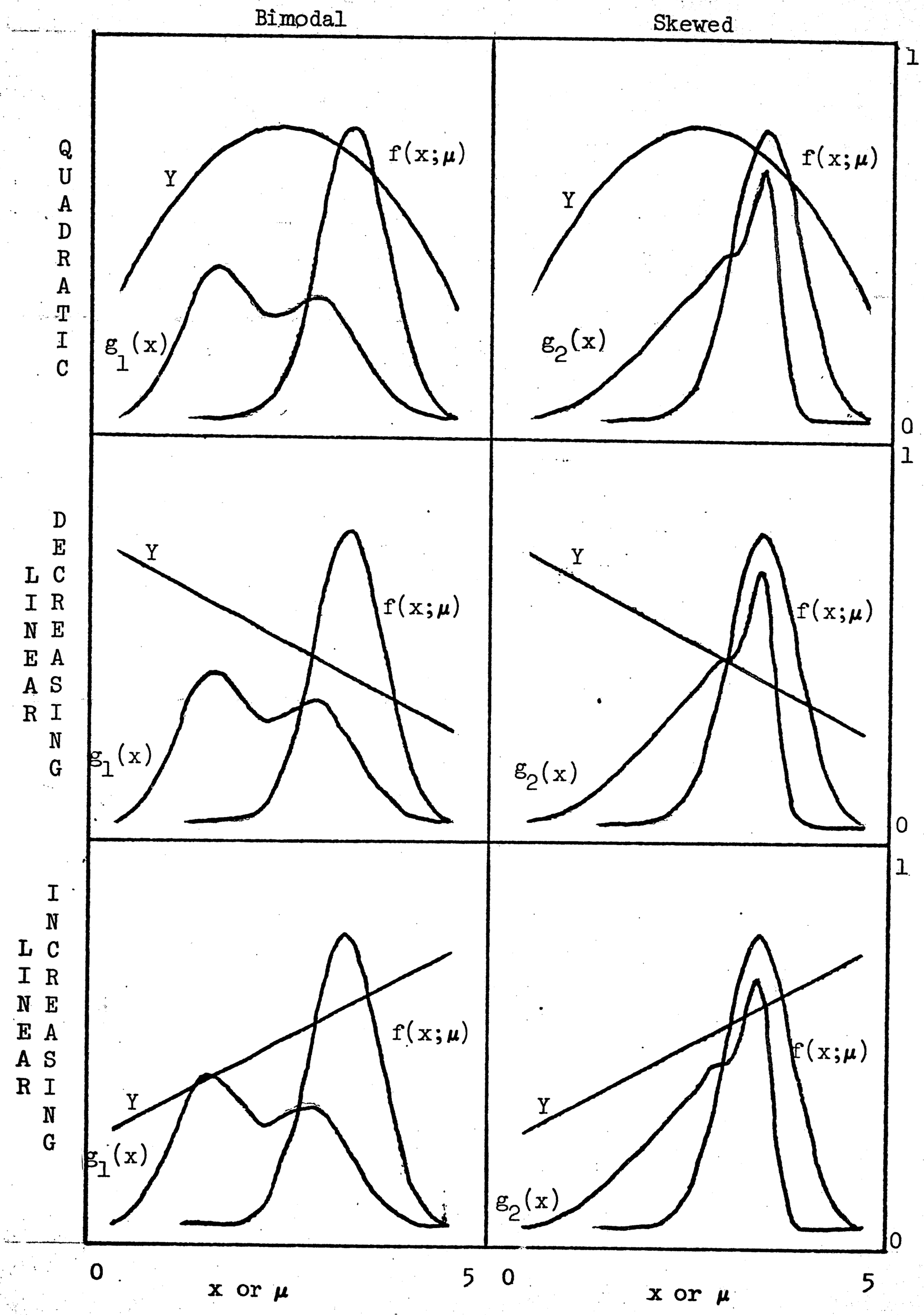


Figure 3.4 - Function forms for the experiment

operating state, μ^*. Five such estimates will be made for each replication: one based on $W(\mu)$ (criterion 3-8) and the last four based on criterion 2-13 for values of k equal to 0, 1.0, 1.5, and 2.0. Another output for each replication will be a graph showing the behavior of $W(\mu)$, $I(\mu)$, and $S_I(\mu)$ as functions of μ. In addition to the graphs and estimates of μ^* for each replication in the design, a "composite" graph will be made, one for each cell in the design. This "composite" graph will show $\overline{W}(\mu)$, $\overline{I}(\mu)$, $R(I(\mu))$, and $R(W(\mu))$ where $\overline{W}(\mu)$ and $\overline{I}(\mu)$ are the average values of $W(\mu)$ and $I(\mu)$ respectively for the eight replications in the cell and R is the observed range of the statistic for the eight replications. Although the variance on $W(\mu)$ is not directly estimable, its observed range in this composite graph should indicate a value relative to the variance on $I(\mu)$.

Because of the relative simplicity of the functions employed, the actual expected process response for any μ is calculable and will be shown on each graph. For the linear responses of the form $ax + b$,

$$E(I(\mu)) = E(Y;\mu) = \int h(x)f(x;\mu)dx \qquad 3\text{-}17.1$$

$$= \int (ax + b)f(x;\mu)dx \qquad 3\text{-}17.2$$

$$= a\int xf(x;\mu)dx + b\int f(x;\mu)dx \qquad 3\text{-}17.3$$

$$= a\mu + b = h(\mu) \qquad 3\text{-}17.4$$

so that for the linear decreasing and linear increasing responses of equations 3-15 - 3-16:

$$\text{Linear decreasing:} \quad E(I(\mu)) = .75 - .1\mu \qquad 3\text{-}18$$

$$\text{Linear increasing:} \quad E(I(\mu)) = .25 + .1\mu \qquad 3\text{-}19$$

For a quadratic form $ax^2 + bx + c$,

$$E(I(\mu)) = E(Y;\mu) = \int h(x)f(x;\mu)dx \qquad 3\text{-}20.1$$

$$= \int (ax^2 + bx + c)f(x;\mu)dx \qquad 3\text{-}20.2$$

$$= a\int x^2 f(x;\mu)dx + b\int xf(x;\mu)dx + c\int f(x;\mu)dx \qquad 3\text{-}20.3$$

$$= aE(x^2;\mu) + bE(x;\mu) + c \qquad 3\text{-}20.4$$

$$= a[E^2(x^2;\mu) + Var(x;\mu)] + b\mu + c \qquad 3\text{-}20.5$$

$$= a[\mu^2 + \sigma_f^2] + b\mu + c \qquad 3\text{-}20.6$$

$$= h(\mu) + a\sigma_f^2 \qquad 3\text{-}20.7$$

so that equation 3-14 yields for expected responses:

$$\text{Quadratic:} \quad E(I(\mu)) = .80 - .08(\mu-2.5)^2 - .08(.25) \qquad 3\text{-}21$$

Thus it will be possible not only to compare the stability of the statistics $I(\mu)$ and $W(\mu)$ relative to each other but to compare them to the actual expected response. The results of this phase are hoped to shed light on the predictive capabilities of the candidate optimizing criteria.

The second phase of the experiment investigates the effect of an increasing variance of the error component of the response on the behavior of the statistics $I(\mu)$ and $W(\mu)$ and the optimizing criteria. This effect will be sought by taking one replication from each of the six cells in the first phase of the experiment and adding random components to the observed responses for three levels of σ^2_ε. This leads to a three factor, 18 cell factorial design:

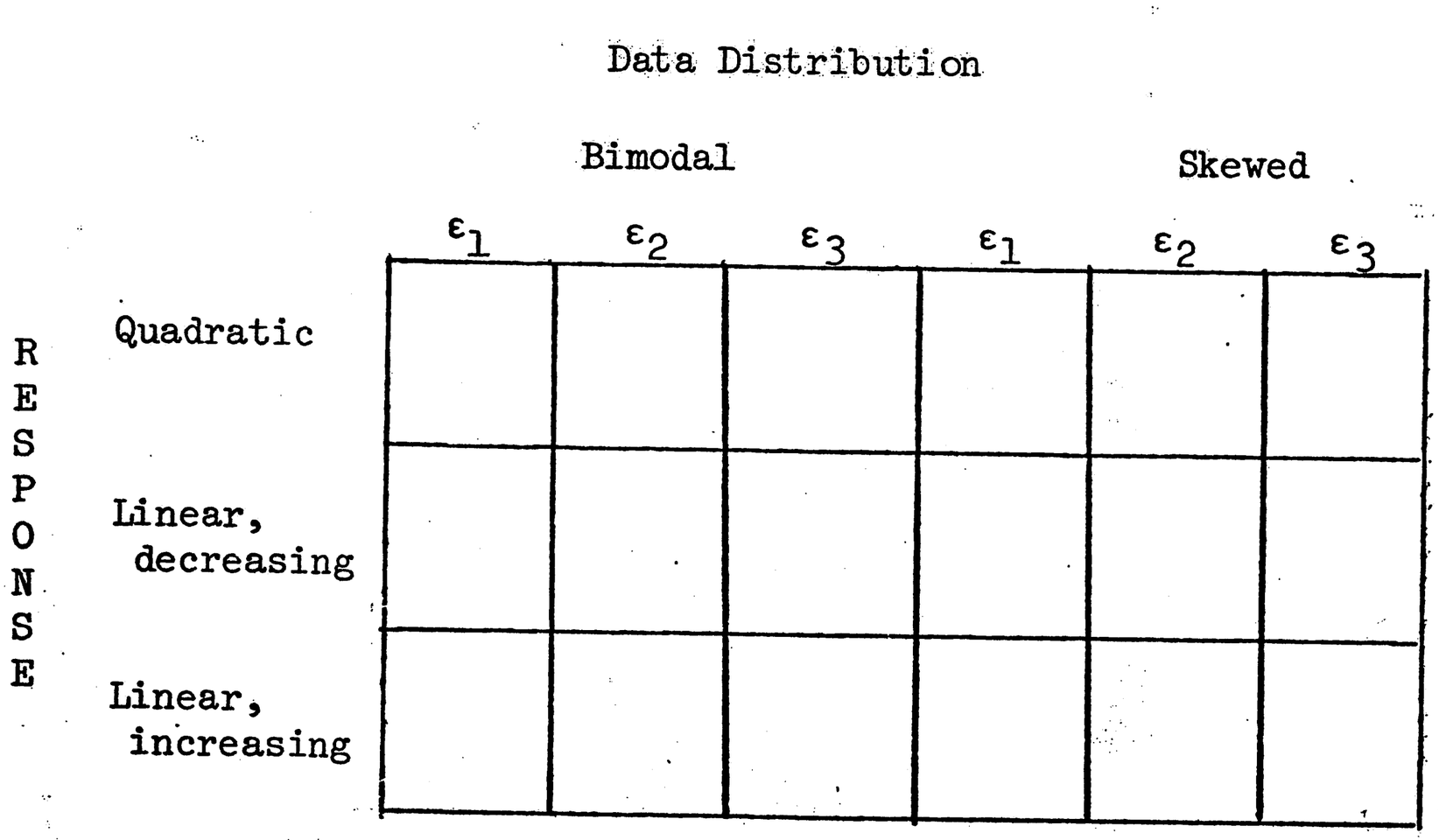


Figure 3.5

The replications selected from phase 1 of the experiment and used in this second phase will be those replications which simultaneously give the best visual fits of $I(\mu)$ and $W(\mu)$ to their theoretical behavior. Each of the 18 cells will contain three replications. Again the results of each replication will include not only the same five estimates of μ^* but graphs showing the effects of the error components. These graphs can be compared then to the error-free graphs of the statistics from phase 1.

The third and final phase of the experiment will determine the effect on the optimizing criteria and $I(\mu)$, $W(\mu)$, and $S_I^2(\mu)$ as a result of decreasing the variance, σ_f^2, of the control distribution $f(x;\mu)$. Here the results, in theory, are predictable. For example consider $W(\mu)$ as formulated in equation 3-3. Let $\mu = x^{(i')}$ for some i'; $1 \leq i' \leq N$; so that the mean of the control distribution is equal to one of the sample values of the process parameter. Then as σ_f^2 goes to zero, the ratios $f(x^{(i)};\mu)/g(x^{(i)})$ will be affected in the following manner:

for $i \neq i'$; $f(x^{(i)};\mu)$ may increase temporarily but in the limit will go to zero, and since $g(x^{(i)})$ remains unchanged as σ_f^2 goes to zero, the ratio thus goes to zero.

for $i = i'$: $f(x^{(i)};\mu)$ increases without bound as σ_f^2 tends to zero and since, again, $g(x^{(i)})$ remains unaffected, the ratio increases without bound.

As a result, $W(\mu)$ will approach $Y^{(i')}$ as a limit when σ_f^2 tends to zero since all the weight effectively is concentrated on the one observation $(x^{(i')};Y^{(i')})$. Now let μ be somewhere between two observed values of the process parameter, say $x^{(i')}$ and $x^{(i'')}$. Then as σ_f^2 goes to zero, all ratios go to zero. But the ratio of that $x^{(i)}$ (either $x^{(i')}$ or $x^{(i'')}$) which is closest to μ will tend to zero much slower than any of the others. Thus it will effectively receive all the weight and $W(\mu)$ will tend toward the response observed at that point. The limiting behavior of $W(\mu)$ will thus appear as discrete jumps from observed response to observed response with the shifts occurring midway between the observed $x^{(i)}$.

The behavior of $I(\mu)$ is also dependent on how the ratios $f(x^{(i)};\ )/g(x^{(i)})$ change in the formulation of $I(\mu)$ (equation 3-1) as σ_f^2 goes to zero. Again let $\mu = x^{(i')}$, some observed value of the process parameter x, and left σ_f^2 go to zero. Again, in the limit $f(x^{(i')};\mu)/g(x^{(i')})$ dominates all the other ratios so that the numerator of 3-1 very nearly becomes $Y^{(i')}f(x^{(i')};\mu)/g(x^{(i')})$, and $I(\mu) = Y^{(i')}f(x^{(i')};\mu)/Ng(x^{(i')})$ which increases without bounds as σ_f^2 nears zero. Now let μ be somewhere between two sample values of x. Then as σ_f^2 goes to zero, $f(x^{(i')};\mu)/g(x^{(i')})$ approaches zero for all i and

$$\lim_{\sigma_f^2 \to 0} I(\mu) = \lim_{\sigma_f^2 \to 0} \sum_{i=1}^{N} f(x^{(i)};\mu)Y^{(i)}/Ng(x^{(i)}) \qquad 3\text{-}22.1$$

$$= \frac{1}{N} \sum_{i=1}^{N} \lim Y^{(i)} f(x^{(i)}; \ ) / g(x^{(i)}) \qquad 3\text{-}22.2$$

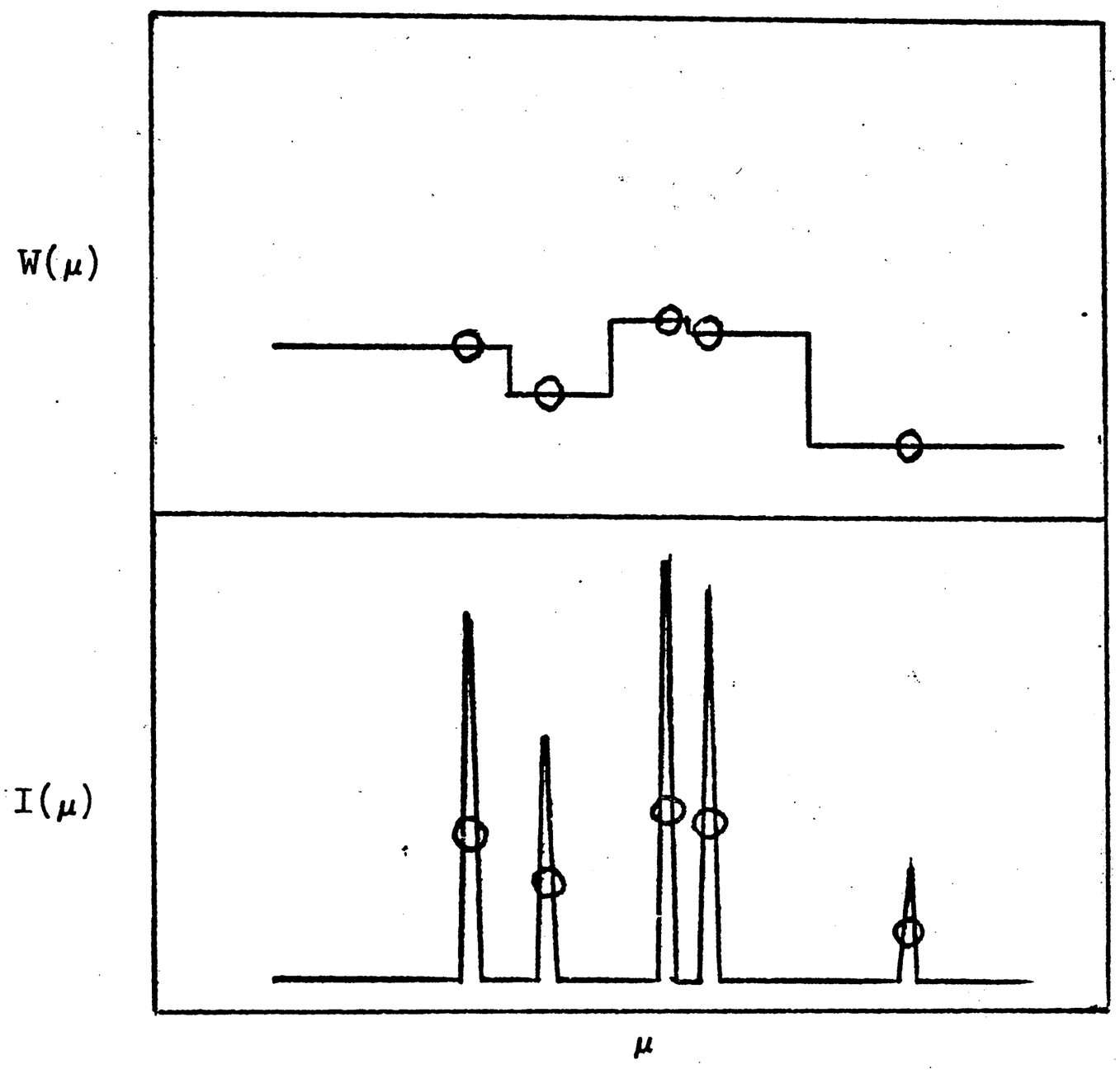
$$\sigma_f^2 \to 0$$

$$= 0 \qquad 3\text{-}22.3$$

The expected behavior of $I(\mu)$ for small σ_f^2, then, is a succession of spikes or rabbits occurring over the observed values of the process parameter. This behavior as well as that of $W(\mu)$ for small σ_f^2 is depicted in Figure 3.6.

As for the effect on the predictive ability of the candidate criteria of a small σ_f^2, the estimate provided by $W(\mu)$ will tend toward the maximum observed response and that of $I(\mu)$ toward the maximum ratio of $Y^{(i)}/g(x^{(i)})$. Clearly if the observed behavior of $I(\mu)$ and $W(\mu)$ in an actual application approximates that of Figure 3.6, then these estimates are useless.

These expected behaviors will be substantiated by taking one replication from phase 2 of the experiment and allowing σ_f^2 to reduce toward zero.



"o" = sample observation

Figure 3.6 - Expected behaviors of $W(\mu)$ and $I(\mu)$ for small σ_f^2

4.0 Analysis of Results

The subject of this chapter is the experiment described in Chapter 3. Its results will be presented and analyzed according to the phases as described there.

4.1 Analysis of Phase 1

Although the first phase of the experiment is factorial by design, it should be clear that the notion of conventional analysis of variance techniques does not apply to the reduction of the results. Aside from a summary of the predictions of the optimal process state by the candidate optimizing criteria, the primary goal of the experiment is a qualitative understanding of how the statistics $I(\mu)$, $W(\mu)$, and $S_I^2(\mu)$ behave under varied process conditions. Therefore, presentation of the results will be, for the most part, qualitative. To begin with, however, a quantitative summary of the estimates, $\hat{\mu}^*$, of the optimal process state for the eight replications of each cell is presented in Table 4-1. The region of interest for the mean, μ, of the optimal control distribution, $f(x;\mu)$, was the interval (0,5) for all replications of all cells. More need be, and will be, said about selection of this region later. The table shows the minimum, maximum, and average estimates of μ^* for each candidate criterion as well as the actual or theoretical optimum state for the region of interest. Also shown is the composite estimate of μ^* which differs from the average by being that value of μ which maximizes $\overline{W}(\mu)$ and $\overline{I(\mu)-k\sqrt{S_I^2(\mu)}}$; $k=0$, 1.0, 1.5, 2.0, as taken from the composite graph made for each cell. Lastly, the table shows for each cell the average value of μ for which

Bimodal						Skewed				
W	I	I-1.0 S_I	I-1.5 S_I	I-2.0 S_I		W	I	I-1.0 S_I	I-1.5 S_I	I-2.0 S_I
					QUADRATIC					
2.5	1.2	1.4	1.4	1.5	min	2.4	1.8	1.9	2.0	2.0
2.6	3.1	3.0	3.0	3.0	max	2.5	4.1	3.0	3.0	3.0
2.51	2.46	2.46	2.45	2.46	ave	2.46	3.12	2.85	2.65	2.72
2.50	2.65	2.65	2.65	2.65	comp	2.45	3.70	2.80	2.80	2.80
	$\mu^* = 2.50$		$\bar{\mu}_s = 1.60$				$\mu^* = 2.50$		$\bar{\mu}_s = 2.85$	
					LINEAR, DECREASING					
0.0	0.4	0.6	0.9	1.3	min	0.0	0.8	1.3	1.5	1.7
0.0	1.9	1.8	1.8	1.8	max	0.0	2.7	1.8	2.8	2.9
0.00	1.38	1.49	1.55	1.64	ave	0.00	1.30	1.82	2.24	2.45
0.00	1.30	1.60	1.60	1.65	comp	0.00	1.00	1.65	2.00	2.40
	$\mu^* = 0.0$		$\bar{\mu}_s = 2.80$				$\mu^* = 0.0$		$\bar{\mu}_s = 3.05$	
					LINEAR, INCREASING					
5.0	2.6	2.6	2.6	2.6	min	5.0	2.9	2.9	2.9	2.9
5.0	4.7	4.3	3.9	3.6	max	5.0	3.9	3.7	3.3	3.2
5.00	3.90	3.28	3.16	3.05	ave	5.00	3.32	3.16	3.07	3.01
5.00	4.00	3.30	3.15	3.05	comp	5.00	3.50	3.20	3.10	3.00
	$\mu^* = 5.0$		$\bar{\mu}_s = 1.60$				$\mu^* = 5.0$		$\bar{\mu}_s = 2.80$	

NOTE: Region of interest for μ: $0 \leq \mu \leq 5$; μ^* is the μ for which $E(Y;\mu)$ is maximized within the region of interest; $\bar{\mu}_s$ is the average μ for which $S_I^2(\mu)$ was minimized.

Table 4.1 - Summary of optimal process state estimates-- phase 1

the minimum value of $S_I^2(\mu)$ occurred. It is clear from the table that not only is $W(\mu)$ the most accurate predictor in all cases considered but that $W(\mu)$ is also the more stable. This trait is borne out by the composite graph for the cell representing a linear, decreasing response and bimodal distribution of the process parameter (Figure 4.1). This graph is typical of the composite graphs for other cells in that it shows, on the average, that $W(\mu)$ and $I(\mu)$ are both good estimators of the actual expected response. However, the observed ranges of the statistics indicate that $W(\mu)$ will be the more consistent, or stable estimator. Recall that while the variance of $I(\mu)$ was readily estimable, the variance on $W(\mu)$ was not. The fact that $R(W(\mu))$ is significantly less than $R(I(\mu))$ everywhere in this and other composite graphs permit the conclusion that, while still not estimable from a single sample, the variance on $W(\mu)$ is clearly bounded by the variance on $I(\mu)$.

A trend which is to be expected and does, in fact, appear in Table 4.1 is that as the k in the optimizing criterion, 2-13, increases, the estimates of the optimal state tend toward that value of μ which provided, on the average, the minimum value of $S_I^2(\mu)$. This should caution the analyst against large k since there is no reason for this point of minimum variance to be anywhere near the optimal process state, μ^*. Recalling the definition of the variance on $I(\mu)$, its value depends solely on the shape of $h(x)\cdot f(x;\mu)/g(x)$. The analyst should also be cautioned against small k, since examination of individual replications of the experiment shows occasionally wild behavior for the statistic $I(\mu)$. Figure 4.2 representing a replication from

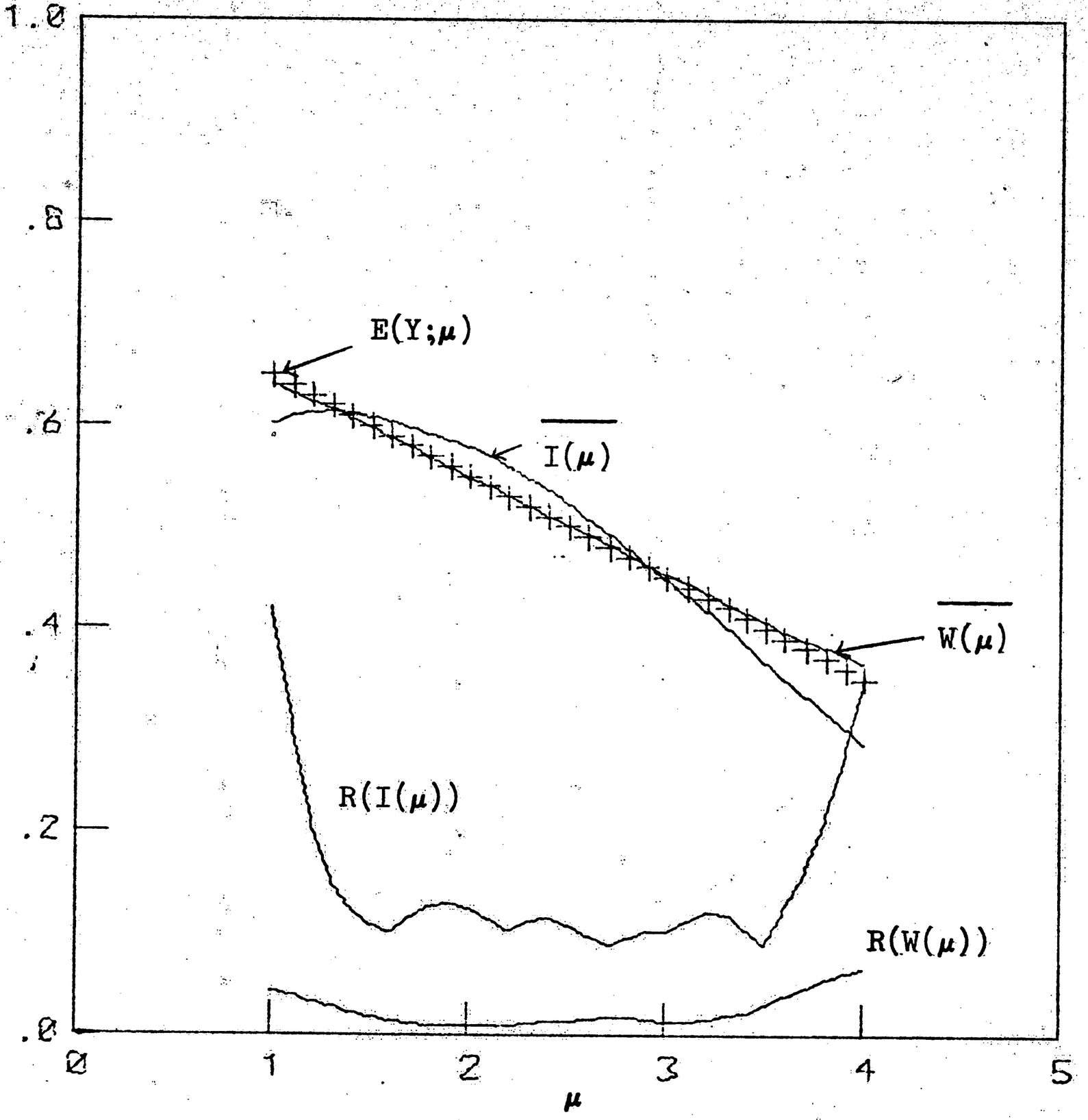
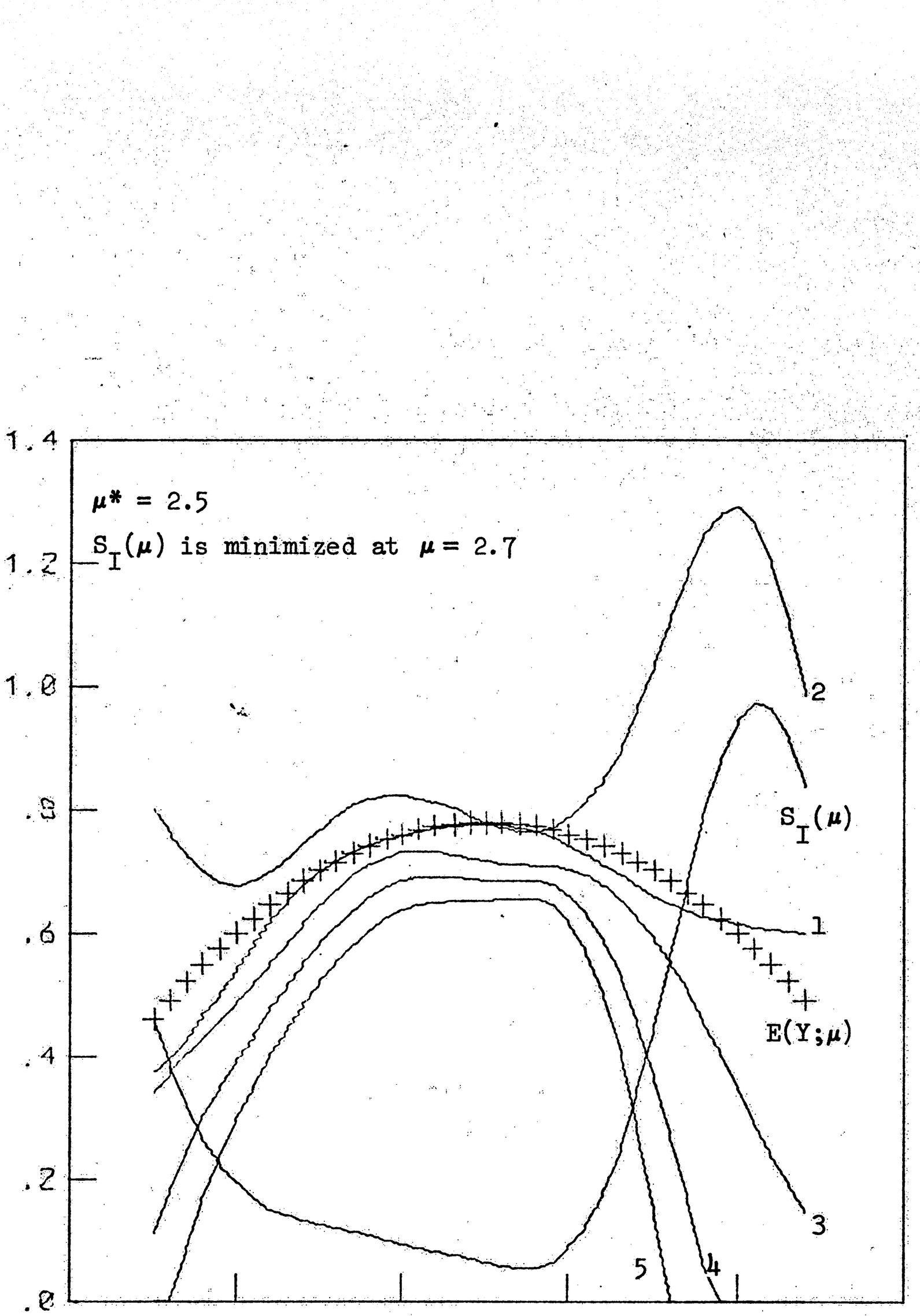


Figure 4.1 - Composite behavior of $\overline{I(\mu)}$ and $\overline{W(\mu)}$ for 8 replications



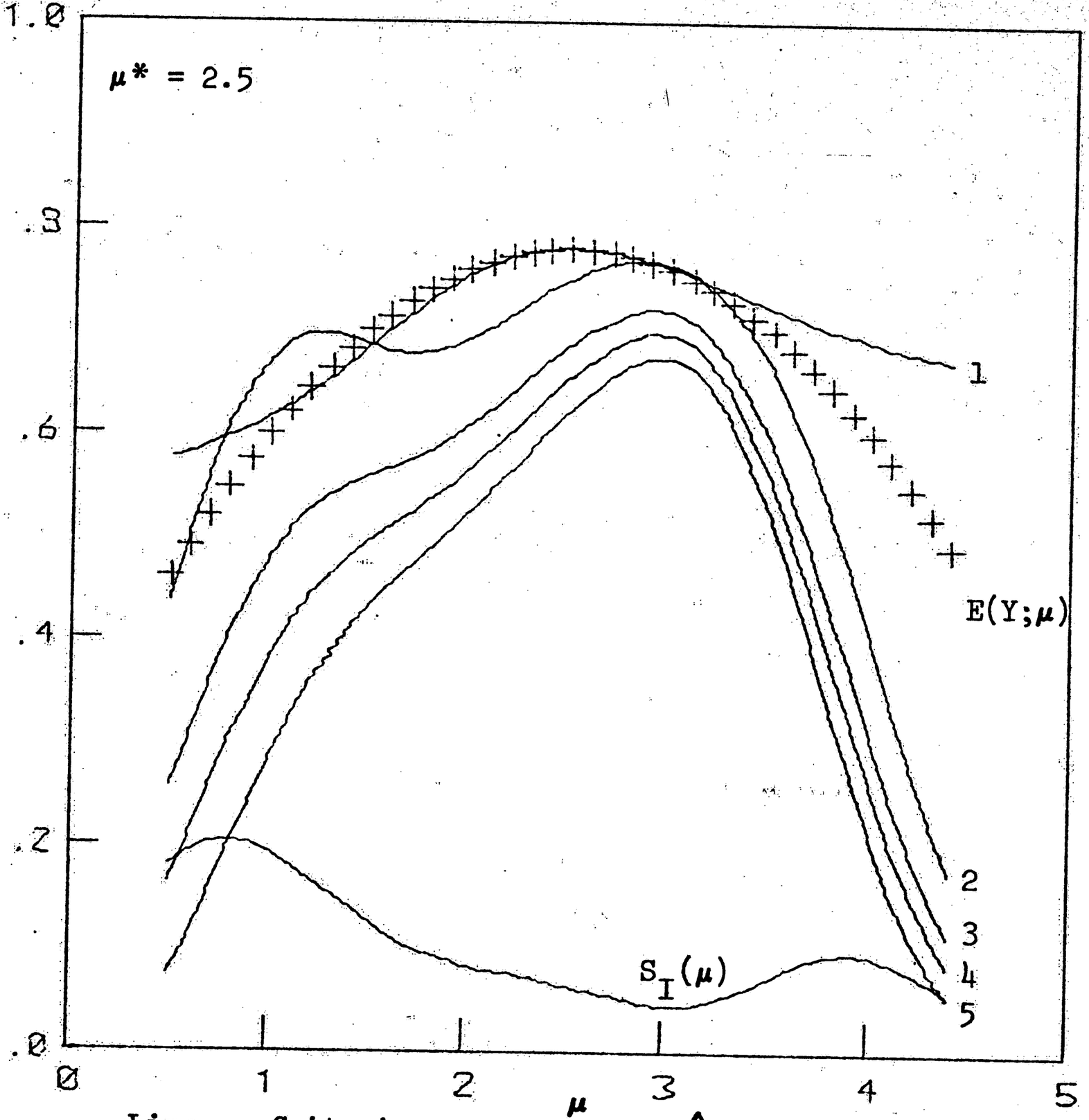
Line	Criterion	$\hat{\mu}^*$	Maximum
1	$W(\mu)$	2.5	.78
2	$I(\mu)$	4.0	1.29
3	$I(\mu) - 1.0S_I(\mu)$	2.1	.73
4	$I(\mu) - 1.5S_I(\mu)$	2.2	.69
5	$I(\mu) - 2.0S_I(\mu)$	2.7	.66

Figure 4.2 - Tendencies of criteria based on $I(\mu)$

the parabolic response, skewed data, cell of the design exhibits both of these effects. Notice the wild behavior of $I(\mu)$ which, using optimizing criterion 2-13 with k=0, results in an estimate of μ^* of 4.0. As k increases to 1.0 and then to 2.0 the estimates of μ^* converge rapidly toward the μ at which $S_I^2(\mu)$, or equivalently, $S_I(\mu)$, was minimized. Notice also that $W(\mu)$ remains unfooled and estimates μ^* to be 2.5, which is the actual optimum process state using this reponse. This wild behavior of $I(\mu)$ was, however, the exception; occurring in fewer than 10% of the replications. In defense of the statistic $I(\mu)$, the results from another replication in that same cell are presented in Figure 4.3. Although better behaved than before, $I(\mu)$ is still less stable and far less accurate than $W(\mu)$.

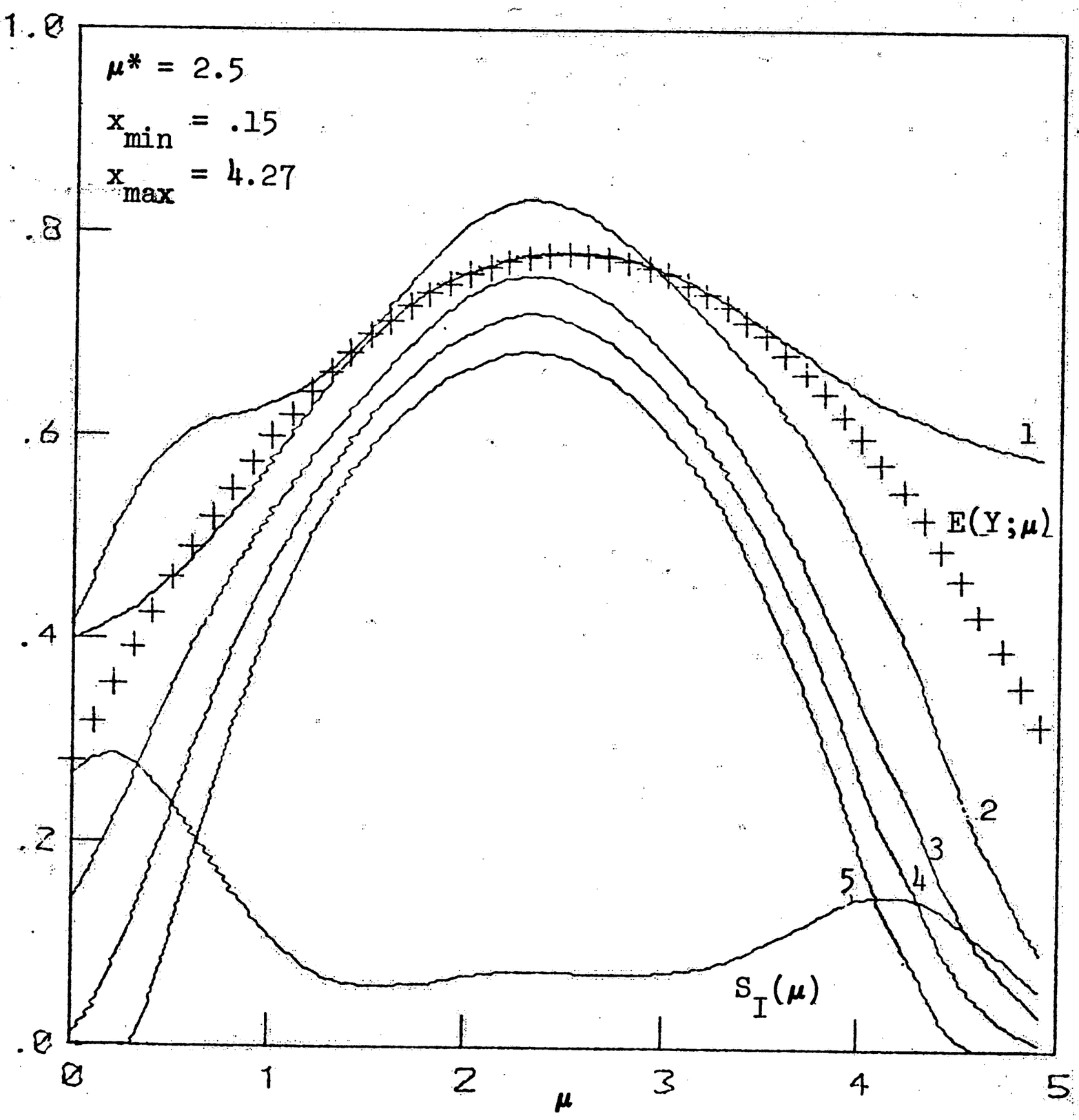
If the analyst were forced to use criterion 2-13 for estimating μ^*, the results presented thus far indicate that, while there is a need for k to be other than zero, it is not clear what other value should be used. The optimizing criterion based on $W(\mu)$ has no such ambiguity and since estimates based on this weighted average appear not only more accurate than those of criteria based on $I(\mu)$ and $S_I^2(\mu)$ but also significantly more stable, it offers the most promise as an optimizing criterion for use on "real" data.

With the intent now of establishing some reasonable rule for selection of a "suitable region of interest" the behavior of $I(\mu)$, $W(\mu)$, and $S_I^2(\mu)$ will be examined in detail, particularly for extreme μ. Figure 4.4 shows the results of one replication from the cell with parabolic response and bimodal data. This graph yields little



Line	Criterion	$\hat{\mu}^*$	Maximum
1	$W(\mu)$	2.5	.78
2	$I(\mu)$	2.8	.77
3	$I(\mu) - 1.0S_I(\mu)$	2.9	.70
4	$I(\mu) - 1.5S_I(\mu)$	2.9	.72
5	$I(\mu) - 2.0S_I(\mu)$	2.9	.68

Figure 4.3 - Comparison of optimizing criteria



Line	Criterion	$\hat{\mu}^*$	Maximum
1	$W(\mu)$	2.5	.78
2	$I(\mu)$	2.3	.84
3	$I(\mu) - 1.0S_I(\mu)$	2.3	.76
4	$I(\mu) - 1.5S_I(\mu)$	2.3	.72
5	$I(\mu) - 2.0S_I(\mu)$	2.3	.68

Figure 4.4 - Examination of the behavior of $I(\mu)$, $S_I(\mu)$, and $W(\mu)$

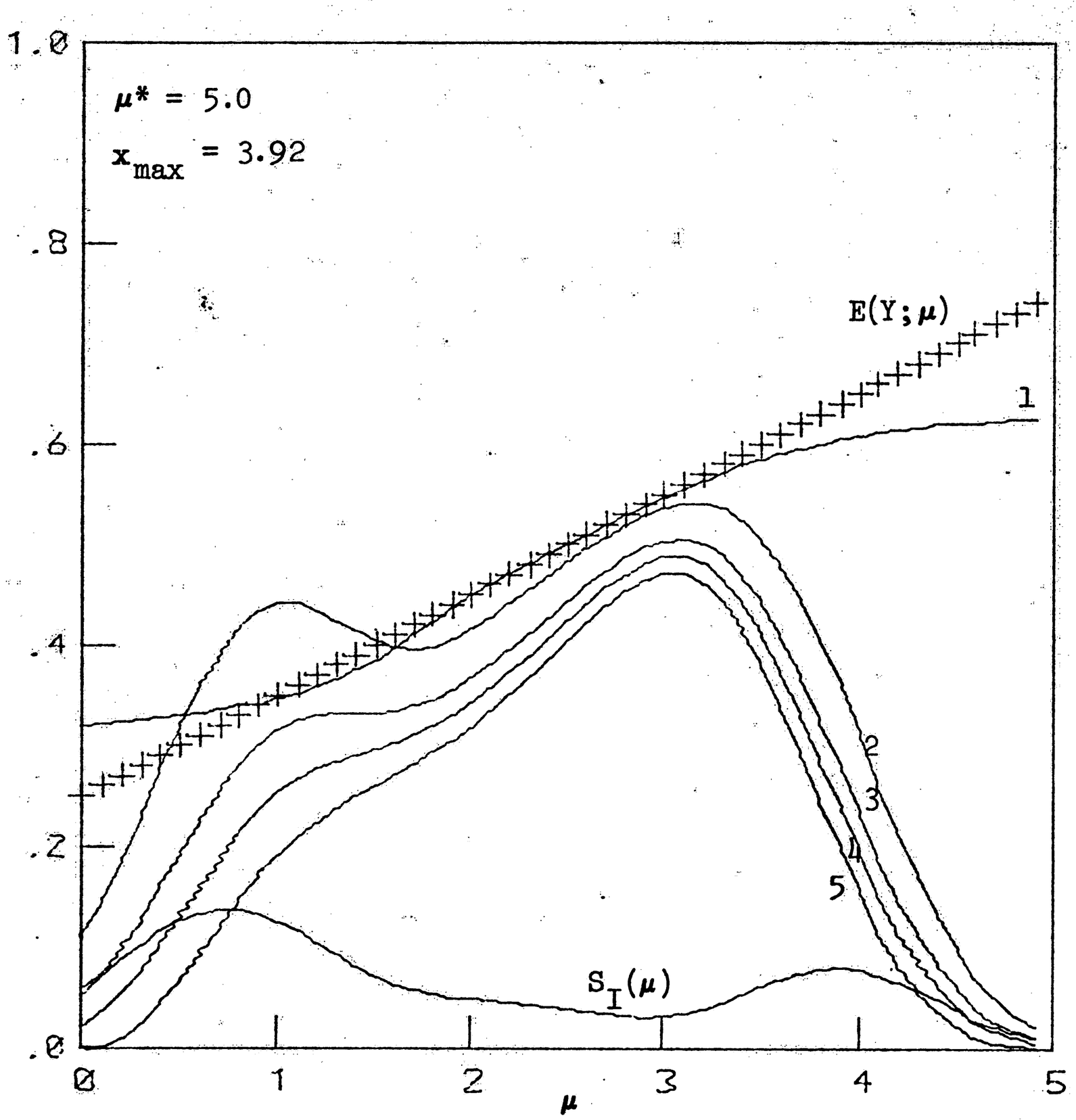
information as to the relative desirability of the optimizing criteria. It is noticed that as the center, μ, of the control distribution, $f(x;\mu)$, is moved toward either extreme of the arbitrarily selected region of interest, $S_I^2(\mu)$ increases before tailing off monotonically. It begins tailing off at x_{min} and x_{max} respectively. Here x_{min} is the smallest sample value of the process parameter x and x_{max} the largest. From its formulation (equation 2-11 and 2-12) the limiting behavior of $S_I^2(\mu)$ is as shown here:

$$\lim_{\substack{\mu \to \infty \\ \text{or} \\ \mu \to -\infty}} S_I^2(\mu) = 0 \qquad 4\text{-}1$$

Clearly once μ is outside the bounds of the data, the terms contributing to $S_I^2(\mu)$ can only get smaller. The increasing values of $S_I^2(\mu)$ as μ moves from the center of the distribution toward x_{min} or x_{max} are not unexpected and will, under rather general conditions, hold. This can be explained heuristically by considering placement of the control distribution at $\mu=4.0$. Only about 2% of the observations from the bimodal $g(x)$ will have values of the process paramter greater than 4.0 so that by placing $f(x;\mu)$ at $\mu=4.0$ and since $f(x;4.0)$ is symmetric about 4.0, half of the expected process response (equation 1-4)--i.e., $x \geq 4.0$, will be estimated by only some 2% of the observations while the other half, ($x<4.0$) will be based on 98% of the observations. For a sample size of 100, this means on the average 98 observations will contribute to one half of the integral and only 2 observations to the rest. One can intuitively expect inter-sample stability for the contribution to the estimate made by the 98 observations, but would also

expect a large variation from the half estimated from only two observations. One would naturally conclude that the variance of the estimates of the expected response at $\mu = 4.0$ would be greater than at a point where, say on the average, 60 observations contribute to one half the integral and 40 to the other. Similarly as μ moves out further into the tail of $g(x)$, since on the average even fewer sample values will contribute to one half of the estimate, that the variance of the estimates should continue to increase. The seemingly contradictory observed behavior of $S_I^2(\mu)$ in these extremes is strictly a function of the data and indicates a lack of information in the regions outside the limits of the data. This serves as a warning that attempts to extrapolate beyond the limits of the data could lead to meaningless and/or hazardous results. While the exact behavior of the variances on $I(\mu)$ and the statistic $S_I^2(\mu)$, of course, depend entirely on how well $g(x)$ mimics the product $Y \cdot f(x;\mu)$, this heuristic argument will hold for any but the worst behaved responses. Linear and quadratic responses such as are the subject of this study certainly are not such functions. The conclusion to be reached here is that the heretofore arbitrary region of interest should at least be confined to the limits of the data if results are to be meaningful.

Is this limitation on the region of interest sufficient? No. Consider now the skewed distribution, $g_2(x)$. Only 1% of the observations sampled from it will exceed 3.87. Figure 4.5 shows the results of one of the replications in the cell representing skewed data and a linear increasing response. X_{max} for this replication is 3.92.



Line	Criterion	$\hat{\mu}^*$	Maximum
1	$W(\mu)$	5.0	.64
2	$I(\mu)$	3.2	.54
3	$I(\mu) - 1.0S_I(\mu)$	3.1	.51
4	$I(\mu) - 1.5S_I(\mu)$	3.1	.49
5	$I(\mu) - 2.0S_I(\mu)$	3.0	.47

Figure 4.5 - Examination of the behaviors of $I(\mu)$, $S_I(\mu)$, and $W(\mu)$

Notice not only the tailing of $S_I(\mu)$ past x_{max} but also the rapid decline of $I(\mu)$, again indicating a lack of information beyond the bounds of the data. But $I(\mu)$ began its decay for values of μ considerably less than x_{max} indicating possibly a serious lack of information even within the bounds of the data. This suggests limiting the region of interest, perhaps, by some percentage point of the distribution of the data. Contrasting behavior from the skewed data can be observed by referring back to Figure 4.2. For that replication $x_{max} = 4.13$. Rather than exhibiting decay for μ's less than x_{max}, $I(\mu)$ "blows up" as μ nears x_{max}. This is the effect of a sample outlier (probability $(x > 4.0) < .00001$). Thus this indicates that estimates of the process response for μ's near x_{max} should be discounted and substantiates the need to further limit the region of interest.

One might suggest eliminating the outlier from analysis, and this would be an easy matter--for this example. But outlier detection is not always such a simple matter and a similar behavior could result from a small clustering of non-outliers. Agreed then that the region of interest should be restricted further than the bounds of the data, what criterion should be used to determine the limits? An arbitrary designation of letting μ vary through the middle 60, 70, or 80% of the observations might be sufficient and safe, but the author feels that because of the interaction of the two distributions $f(x;\mu)$ and $g(x)$ in calculation of the various statistics, a designation based on $g(x)$ alone--which the above presumes--is inadequate. Therefore as a criterion for selecting a suitable region of interest, the following procedure is proposed:

For the lower/upper limit on μ, use that point μ' at which $A_1\%$ of the control distribution, $f(x;\mu')$ lies within $A_2\%$ of the left hand/right hand tail of $g(x)$, the distribution of the observed values of the process parameter.

The value of such designations might be determined once reasonable values of A_1 and A_2 are established. For clues as to what might be "good" estimates of A_1 and A_2 the individual and composite behaviors of the statistics $I(\mu)$, $W(\mu)$, and $S_I^2(\mu)$ were carefully analyzed. $W(\mu)$ was considered first, however since it is bounded by the maximum and minimum observed responses its behavior was thought to be too stable to yield any useful information for determining "reasonable" values for A_1 and A_2. Likewise $I(\mu)$ was observed to be too unstable almost everywhere to produce cutoff points. The behavior of $S_I^2(\mu)$ appeared to offer the most promise.

It was observed that from cell to cell within the design of the experiment, the composite behavior of $S_I^2(\mu)$ exhibited abrupt shifts in levels when moving from the heart of the distribution of the data toward the extremes. As long as μ was somewhere in the center of $g(x)$ the range on $S_I^2(\mu)$ for the eight replications within a cell was less than 15% of its average value, but as μ moved out of the center the range would suddenly begin to increase until the range on $S_I^2(\mu)$ exceeded the average value by as much as two or three times (Figure 4.6). An attempt was then made to interpret this behavior in terms of identifying a reasonable A_1 and A_2. First, in each summary of $S_I^2(\mu)$ the points at which $S_I^2(\mu)$ "takes off" were noted. Two criteria

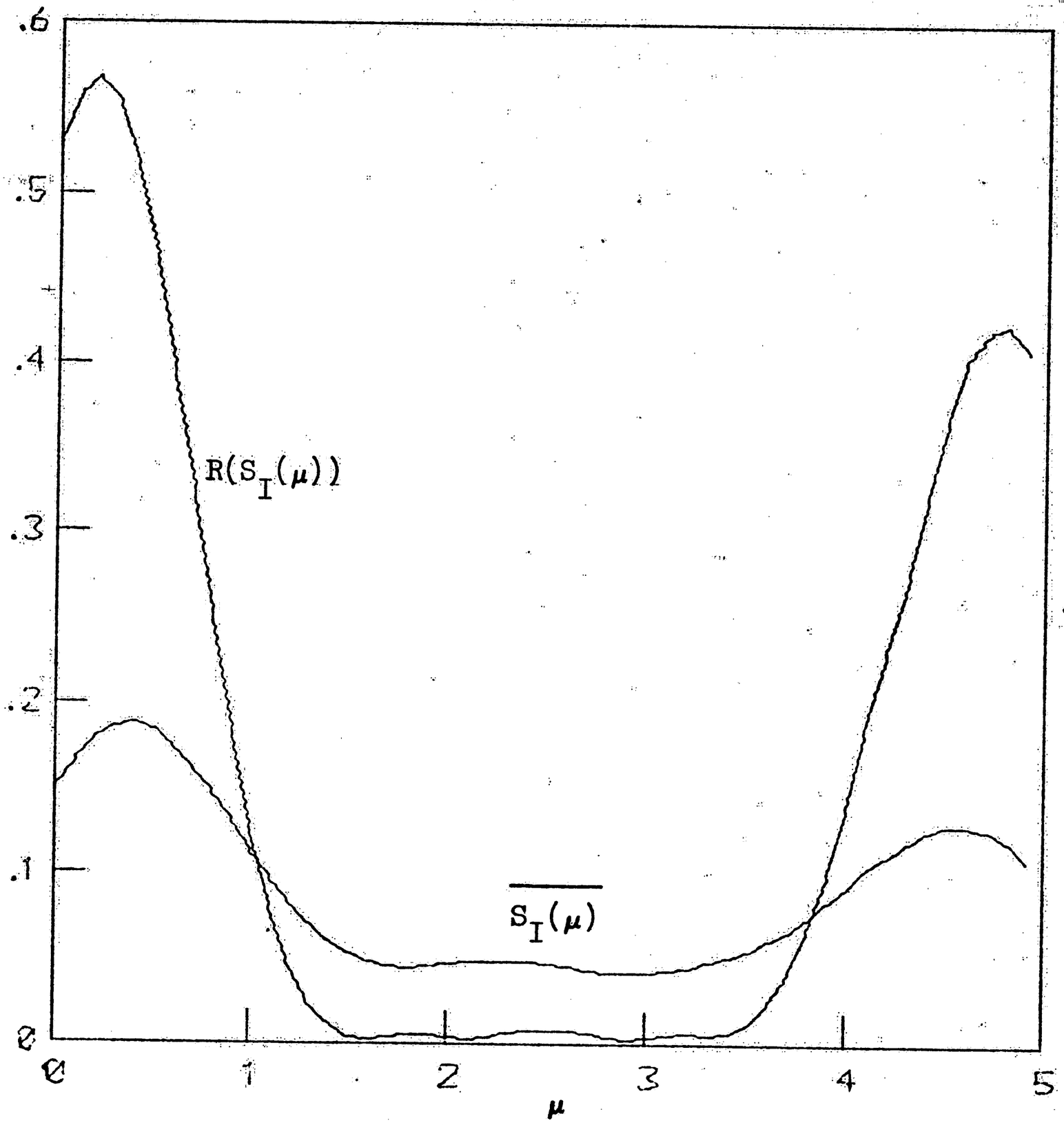


Figure 4.6 - Composite behavior of $S_I(\mu)$ for 8 replications

were arbitrarily selected for determining these points:

1. The μ's at which $R(S_I(\mu)) = \overline{S_I(\mu)}$ 4-2

2. The μ's at which $R(S_I(\mu)) = \frac{1}{2}\overline{S_I(\mu)}$ 4-3

Using one or the other of these criteria two such points are thus determined for each cell of the design, making a total of 12 points. Three of these points will correspond to the left tail of the bimodal $g(x)$, three to the right tail of the bimodal $g(x)$, three to the left tail of the skewed $g(x)$ and three to its right. Thus these 12 points can be viewed as the results of an experiment with four treatments (tails of distributions) and three replications for each treatment. Averaging the results to produce a single value for each treatment leads to the results summarized in Table 4.2.

Next each of the resulting averages was considered individually. Tables for percentage points of the control distribution and $g(x)$ were made so that percentage pairs, (a_1, a_2), could be determined where by setting the mean of the control distribution at one of the averages in the table, $a_1\%$ of the control distribution lay in the outer $a_2\%$ of the appropriate $g(x)$. There will be many such pairs for each of the averages, but interest was confined to a range of 1-50 for a_1 and 1-5 for a_2. Rounding to the nearest 5 for a_1 and 1 for a_2 gave the pairs exhibited in Table 4.3.

A recurring pattern was then sought within either of the criteria 4-2 and 4-3. Looking first within 4-3, the right tail of the bimodal

$R(S_I(\mu)) = \frac{1}{2}\overline{S}_I(\mu)$			$R(S_I(\mu)) = \overline{S}_I(\mu)$	
Left Tail	Right Tail		Left Tail	Right Tail
		Bimodal g(x)		
1.01	4.07	Quadratic Y	.20	4.57
1.24	3.65	Linear, decreasing Y	1.04	3.83
1.04	3.62	Linear, increasing Y	.68	3.90
1.10	3.81	Averages	.81	4.10
Left Tail	Right Tail		Left Tail	Right Tail
		Skewed g(x)		
1.00	2.84	Quadratic Y	.70	2.94
.90	3.30	Linear, decreasing Y	.30	3.60
1.10	2.92	Linear, increasing Y	.80	3.05
1.00	3.02	Averages	.60	3.20

Table 4.2 - Values of μ at which $S_I(\mu)$ "blows up"

$R(S_I) = \frac{1}{2}\overline{S_I}$			$R(S_I) = \overline{S_I}$	
Left Tail	Right Tail		Left Tail	Right Tail
		Bimodal g(x)		
1.10	3.81	Averages	.81	4.10
(10,1)	(20,1)	(a_1, a_2) pairs	(25,1)	(40,1)
(15,2)	(30,2)		(35,2)	
(25,3)	(40,3)		(45,3)	
(30,4)	(50,4)		(50,4)	
(30,5)				
		Skewed g(x)		
1.00	3.02	Averages	.60	3.20
(25,1)	(5,1)	(a_1, a_2) pairs	(50,1)	(10,1)
(40,2)	(5,2)			(10,2)
(50,3)	(5,3)			(15,3)
	(10,4)			(15,4)
	(10,5)			(15,5)

Table 4.3 - Percentage pairs corresponding to average values of μ at which $S_I(\mu)$ "blows up"

g(x) and the left tail of the skewed g(x) limit the value of a_2 to 1. However it is observed that if a_2 is held at 1, values for a_1 range from 10 to 50, or nearly throughout the entire range of consideration. It is thus difficult to pinpoint any "good" single value for a_1. Turning to the results associated with criterion 4-2, three candidates for a_2 occur: 1, 2, and 3. With a_2 = 3 the same problem which hindered identification of a_1 earlier recurs. The same problem is encountered if a_2 = 2, although the range of a_1 is slightly reduced. For a_2 = 1 the selection of a_1 is still not clear, but at least it is isolated to just half the region considered. Averaging the four values of a_1 associated with a_2 = 1 yields an estimate of a_1 equal to 15. It is this pair (15,1), which is used as a first guess at A_1 and A_2. Returning to the tabulated percentage points associated with the g(x)'s and f(x;μ), the "suitable" regions of interest for each cell of the design can be determined. They are:

For cells with skewed g(x): (1.2,3.3) 4-4

For cells with bimodal g(x): (1.0,3.7) 4-5

These intervals define the μ's at which approximately 15% of the appropriate tail of the control distribution, f(x;μ), lies in the appropriate 1% tail of the distribution of the data. While intended by no means to be conclusive for the general case, such a result is at least reasonable since it provides the analyst with estimates of the expected process response in which no more than 15% of the integral in equation 1-4 is based on fewer than 1% of the observations.

One of the questions of interest is how Table 4-1 is affected if the

regions of interest in the cells of the factorial design are limited to those of 4-4 or 4-5. The answer is shown in Table 4.4. The attractiveness of $I(\mu)$ as a predictor of the optimal state has certainly increased, but $W(\mu)$ is still clearly the superior estimator.

The results of this first phase of the experiment can be summarized in two parts:

1. $W(\mu)$ appears to be the most promising predictor of the optimal state when compared to optimizing criteria based on equation 2-13, and
2. Reasonable bounds for ranging the mean of the control distributions were found considering the interaction of the two distributions: control and data.

It remains to substantiate the claim that $W(\mu)$ is the better predictor of μ^* by considering the other conditions thought to affect its and $I(\mu)$'s behavior.

4.2 Analysis of Phase 2

The intent of this phase of the experiment was the determination of what effect the addition of a random component, ε, to the response Y will have on the behavior of the statistics $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$, and particularly their predictive capabilities.

This latter question can be dispensed with immediately. By comparing the resulting estimates of the optimal process states for each of the candidate criteria with the corresponding estimates obtained in phase 1 at which time ε was identically zero, it was found that the maximum shift in $\hat{\mu}^*$ was with one exception 0.1 unit of the process

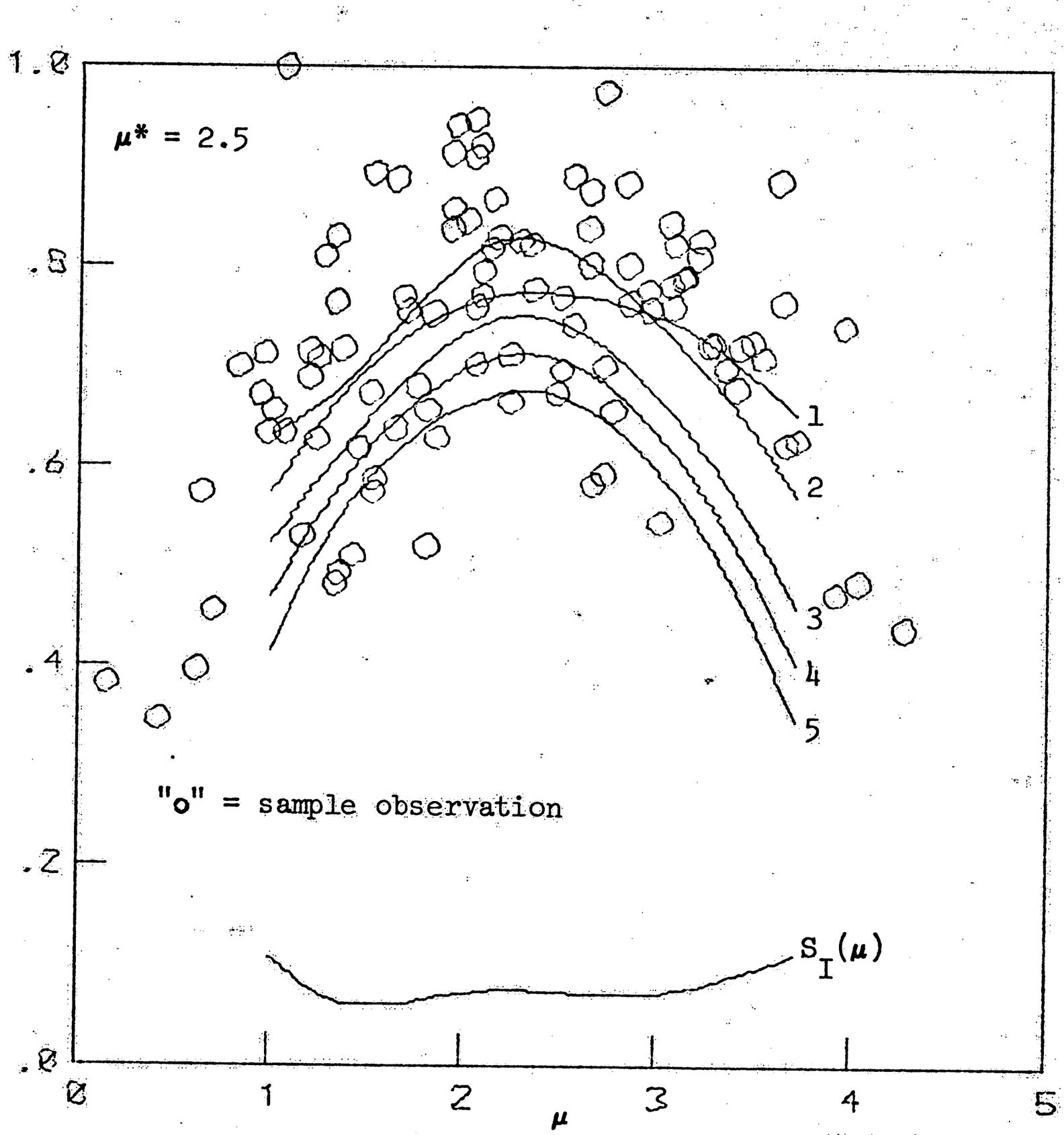
Bimodal						Skewed				
W	I	I-1.0 S_I	I-1.5 S_I	I-2.0 S_I		W	I	I-1.0 S_I	I-1.5 S_I	I-2.0 S_I
					QUADRATIC					
2.5	1.2	1.4	1.4	1.5	min	2.4	1.2	1.9	2.0	2.0
2.6	3.1	3.0	3.0	3.0	max	2.5	3.3	3.0	3.0	3.0
2.51	2.46	2.46	2.45	2.46	ave	2.46	2.87	2.85	2.65	2.72
2.50	2.65	2.65	2.65	2.65	comp	2.45	3.30	2.80	2.80	2.80
	$\mu^* = 2.50$		$\bar{\mu}_s = 1.60$				$\mu^* = 2.50$		$\bar{\mu}_s = 2.85$	
					LINEAR, DECREASING					
1.0	1.0	1.0	1.0	1.3	min	1.2	1.2	1.3	1.5	1.7
1.0	1.9	1.8	1.8	1.8	max	1.2	2.7	2.8	2.9	2.9
1.00	1.45	1.54	1.56	1.64	ave	1.20	1.46	1.82	2.24	2.45
1.00	1.30	1.60	1.60	1.65	comp	1.20	1.20	1.65	2.00	2.40
	$\mu^* = 1.00$		$\bar{\mu}_s = 2.80$				$\mu^* = 1.2$		$\bar{\mu}_s = 3.05$	
					LINEAR, INCREASING					
3.7	2.6	2.6	2.6	2.6	min	3.3	2.9	2.9	2.9	2.9
3.7	3.7	3.7	3.7	3.6	max	3.3	3.3	3.3	3.3	3.2
3.70	3.52	3.20	3.14	3.05	ave	3.30	3.21	3.11	3.07	3.01
3.70	3.70	3.30	3.15	3.05	comp	3.30	3.30	3.20	3.10	3.00
	$\mu^* = 3.70$		$\bar{\mu}_s = 1.60$				$\mu^* = 3.30$		$\bar{\mu}_s = 2.80$	
	Region of interest: $1.0 \leq \mu \leq 3.7$						Region of interest: $1.2 \leq \mu \leq 3.3$			

NOTE: μ^* and $\bar{\mu}_s$ are the same as in Table 4.1.

Table 4.4 - Revised summary of optimal process state estimates - phase 1

parameter even for the maximum considered value of σ_ε^2. This result was true as long as the region of interest for the mean, μ, of the control distribution, was confined to those regions identified in 4-4 and 4-5. Thus under the conditions considered, the random component appears to have little effect on the predictive abilities of the various optimizing criteria, which is clearly a desirable result.

Figure 4.7 shows the results of one of the replications from the cell representing a parabolic response, bimodal distribution of the data, and variance on the error component at level ε_3. It can be compared with Figure 4.4 which shows the results of the same sample under the same conditions except that $\varepsilon \equiv 0$. The preservation of shapes of curves, evident in the comparison of these two figures was characteristic of all plots within the "suitable" regions of interest (either 4-4 or 4-5). Thus not only are the estimates of μ^* stable with respect to random components on the response, but so apparently are the behaviors of the statistics $I(\mu)$, $W(\mu)$, and $S_I^2(\mu)$. As an attempt to quantify this stability the maximum percent change in the value of the statistics $W(\mu)$ and $I(\mu)$ was recorded for each replication for values of μ within the appropriate region of interest. This change is with respect to the benchmark values of the statistics obtained for each of the samples in phase 2 with $\varepsilon \equiv 0$. The results are shown in Table 4.5. It was not necessary to separate the changes in $I(\mu)$ from those of $W(\mu)$, since, as can be inferred from their respective definitions, the percent change in $W(\mu)$ as a result of adding a random component to the response Y is the same as that for $I(\mu)$. Also recorded



Line	Criterion	$\hat{\mu}^*$	Maximum
1	$W(\mu)$	2.4	.79
2	$I(\mu)$	2.3	.84
3	$I(\mu) - 1.0S_I(\mu)$	2.3	.76
4	$I(\mu) - 1.5S_I(\mu)$	2.3	.73
5	$I(\mu) - 2.0S_I(\mu)$	2.3	.69

Figure 4.7 – Typical results from phase 2

Response	Bimodal: $g_1(x)$						RUN NO.	Skewed: $g_2(x)$					
	ϵ_1		ϵ_2		ϵ_3			ϵ_1		ϵ_2		ϵ_3	
	Max % Change	μ_{max}	Max % Change	μ_{max}	Max % Change	μ_{max}		Max % Change	μ_{max}	Max % Change	μ_{max}	Max % Change	μ_{max}
Quadratic	2.0	1.0	3.4	1.0	4.4	3.7	1	1.6	1.2	7.8	1.2	8.7	1.2
	1.0	1.0	2.4	1.8	1.8	3.0	2	2.9	1.2	7.3	1.2	18.2	1.2
	1.0	3.7	3:3	1.0	1.2	1.9	3	4.7	1.2	8.7	1.2	13.6	1.2
Linear, Decreasing	1.6	1.0	4.8	1.0	2.4	2.9	1	7.2	1.2	8.0	1.2	3.1	2.7
	4.4	3.7	2.6	1.0	10.1	2.3	2	1.3	3.3	3.6	3.3	10.1	3.3
	1.9	1.0	2.1	1.0	8.4	3.7	3	2.9	1.2	2.3	1.9	4.7	1.2
Linear, Increasing	2.2	3.7	4.3	3.7	1.8	1.6	1	4.7	1.2	3.2	1.8	5.1	3.3
	1.6	1.0	7.3	1.0	9.4	1.0	2	4.5	1.5	10.9	1.2	7.1	3.3
	.9	1.0	3.7	1.0	9.8	1.0	3	3.3	1.6	1.4	1.7	4.6	3.3
	Region of Interest: $1.0 \leq \mu \leq 3.7$							Region of interest: $1.2 \leq \mu \leq 3.3$					

Table 4.5 - Maximum percent change in $I(\mu)$ or $W(\mu)$ - Phase 2

in the table is the value of μ at which the maximum occurred. Note that the maximum change almost always occurs at or near the extreme of the appropriate region of interest indicating a tendency toward instability as μ nears the limits of the data. This again emphasizes the needs for careful determination of suitable regions of interest.

A three way analysis of variance was performed on these data and the results are summarized in Table 4.6. Although four effects are significant at the .95 level--distribution of the data, level of ε, interaction between response and distribution of the data, and, to a lesser extent, the three way interaction between all factors--it is heartening to observe the dominance of the effect of ε since, after all its effect is what this phase of the experiment sought to evaluate. This analysis of variance can also serve to test the criteria for selecting a "suitable" region of interest as devised in the analysis of phase 1. If the region were, in fact, entirely suitable, it would be hoped that the only significant effects to be observed in this kind of analysis would be due to the levels of ε, and perhaps its interactions. Thus the significance of g(x) and g(x)xY somehow indicate an inadequacy in the regions resulting from values of A_1 and A_2 equal to 15% and 1% respectively. Thus, while not failing, the results of the analysis of variance performed here indicate that improvement in the "15%, 1%" combination for selection of a suitable region of interest could be made. Perhaps an analysis of this kind could be employed for results obtained using different A_1, A_2 pairs in determining suitable regions of interest until the undesired effects

Source of Variation	df	SS	MSE	F-Ratio	Significance
Distribution of data, g(x)	1	59.53	59.53	8.57	Yes
Response, Y	2	8.51	4.26	.62	---
Error level, ϵ	2	177.12	88.56	12.74	Yes
g(x) x Y	2	97.33	48.67	7.00	Yes
g(x) x ϵ	2	5.77	2.89	.42	---
Y x ϵ	4	11.54	2.89	.42	---
g(x) x Y x ϵ	4	78.68	19.67	2.83	Yes
Residual error	36	250.17	6.95		---
TOTALS		688.65			at .95 level

Table 4.6 - Analysis of variance for phase 2

became insignificant. Such refinement in A_1 and A_2 at this point is, however, not deemed necessary and may be considered as an area for further work.

Stability in estimates of μ^* has been shown, and consistency in the shapes of the curves for each of the statistics has been inferred. The significant effects regarding maximal percent change within the region of interest have been identified and now the mean maximal percent change in $I(\mu)$ (or, equivalently, $W(\mu)$) is presented as a function of μ. The results are listed in Table 4.7.

Error Level	σ^2_ε	Mean Maximal % Change in $I(\mu)$
ε_1	.0016	2.5%
ε_2	.0064	4.8%
ε_3	.0144	6.9%

Table 4.7

These indicate, as would be expected, a positive effect with an increasing variance of the random component of the response which is nearly linear. The author feels that ε_3, corresponding to a standard deviation of .12, is sufficiently large, and its mean effect on $I(\mu)$ of some 6.9% to be sufficiently small to justify the conclusion, when corroborated with other results of this section, that the behaviors of $W(\mu)$ and $I(\mu)$, as well as their abilities to predict the optimal state, are not significantly affected by a random com-

ponent present in the process response. The effects due to ε certainly are far less dramatic on the behaviors of the statistics than are the fluctuations observed from sample variations observed in the first phase of the experiment.

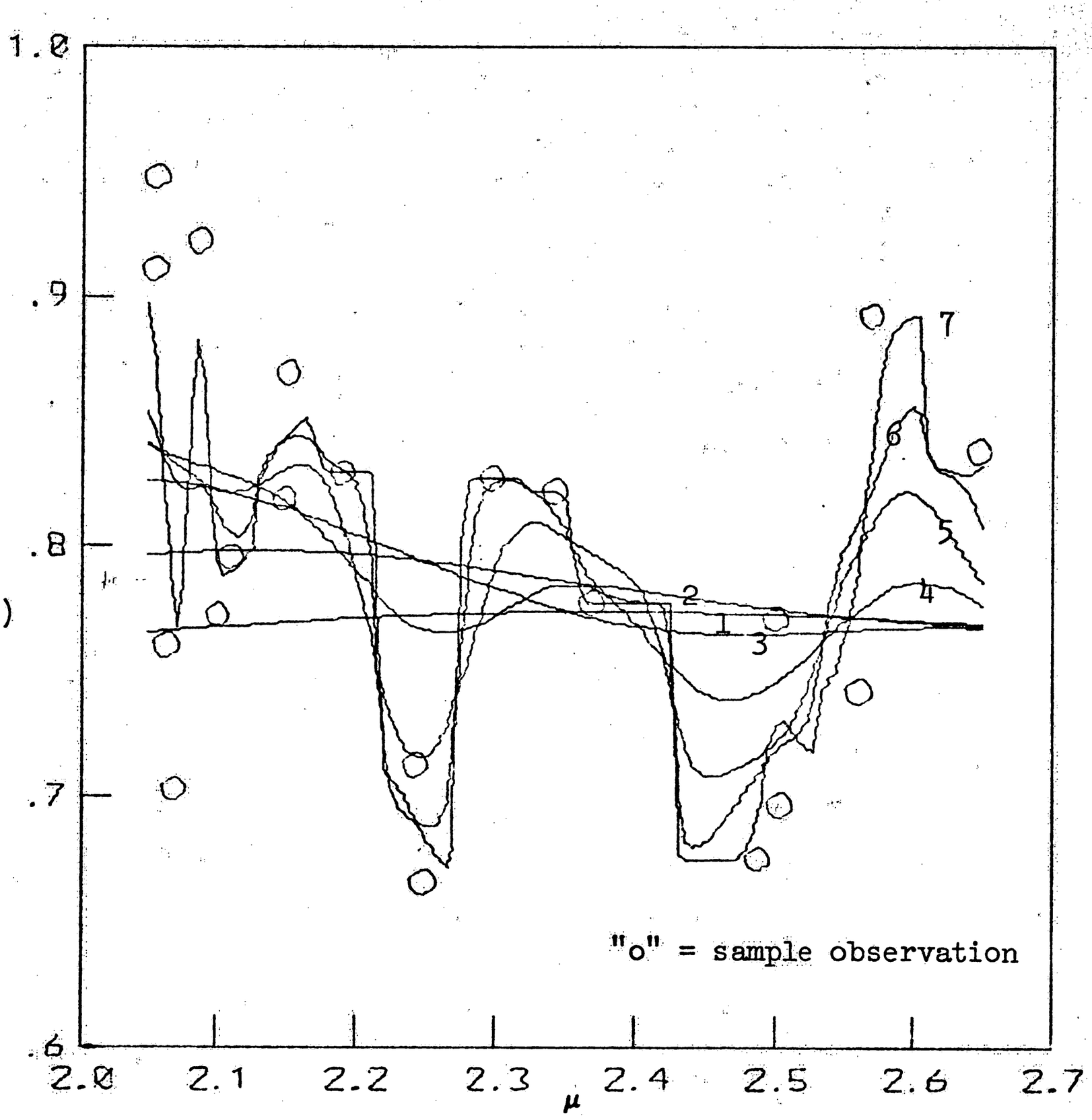
Summarizing the results of this second phase of the experiment:

1. The claim that $W(\mu)$ is the most promising predictor of the optimal process state has been substantiated by the independence of its estimates from the random component of the response, and
2. No statistics or predictors are significantly affected by the random component of the response, and
3. A method for refinements in the estimates of A_1 and A_2 which are used to identify the "suitable" region of interest was suggested.

One other factor remains to be investigated: the effect of the variance of the control distribution.

4.3 Analysis of Phase 3

While the results of this phase of the experiment were, in theory, predictable, it is still instructive to see how $W(\mu)$ and $I(\mu)$ are affected as the variance on the control distribution is reduced. For this reason the data which produced Figure 4.7 was reanalyzed with the intention of observing when the predicted effects appear and how predominant they are for decreasing values of σ_f^2, the variance of the control distribution. Figure 4.8 contains a portion of the plot showing the effects on $W(\mu)$. The scale has been magnified to empha-

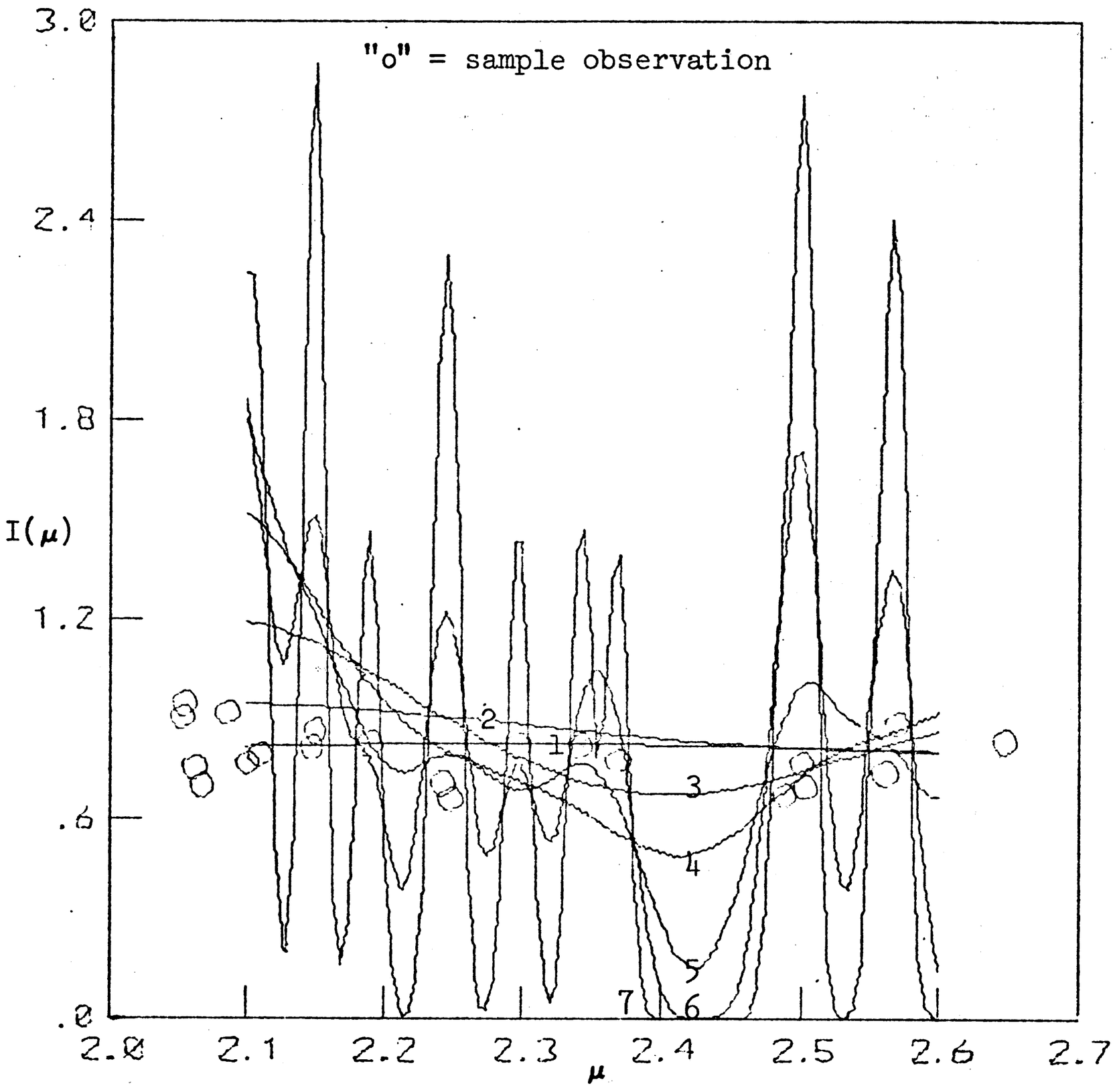


for line m, $\sigma_f^2 = \frac{1}{4^{m-1}}(0.25)$

Figure 4.8 - Effect on $W(\mu)$ for differing values of σ_f^2

size the behavior. The lines of the plot have been numbered from 1 to 7 corresponding to decreasing values of σ_f^2. Line 1 corresponds to the appropriate segment in Figure 4.7 for which $\sigma_f^2 = .25 = \sigma_{f,1}^2$. Line 2 represents the behavior of $W(\mu)$ for $\sigma_f^2 = \frac{1}{4}\sigma_{f,1}^2 = \sigma_{f,2}^2$, and in general, line m corresponds to the behavior of the statistic for $\sigma_f^2 = \frac{1}{4^{m-1}}\sigma_{f,1}^2 = \sigma_{f,m}^2$. Line 7 clearly shows the predicted stepping effect, particularly in the region for values of μ between 2.3 and 2.5. A corresponding plot showing the behavior of $I(\mu)$ as a function of σ_f^2 can be found in Figure 4.9. Again line 7 clearly shows the predicted rabbits, or spiking effect occurring over the data points.

Obviously information as to the location of the optimal process state cannot be derived from Line 7 of either of Figures 4.8 or 4.9. However, the variance on the control distribution giving line 7 was approximately 6×10^{-5} and it is equally as obvious that no self-respecting analyst would assume such a tight control to be possible if his historical data covered a range as large as that of either $g_1(x)$ or $g_2(x)$. Thus this extreme behavior is not likely to be encountered in actual analyses. But on the other hand even line 3, particularly in Figure 4.9, shows unreasonable behavior, for it predicts 1.20 or 120% to be the expected process response if the control distribution is centered with mean μ at 2.1 at which, as pointed out in an earlier discussion, is ridiculous for a response such as percent yield. For Line 3, σ_f^2 was about 1/16 that of Line 1. Even Line 2 may be subject to some discussion as to the validity of its predictions. This suggests then that the optimal control capability, however



for line m, $\sigma_f^2 = \frac{1}{4^{m-1}}(0.25)$

Figure 4.9 - Effect on $I(\mu)$ for differing values of σ_f^2

selected, should be somehow consistent with the observed range of the data. It also leads to a "bonus" result for this phase of the experiment.

As a by-product of the first phase of the experiment, a criterion for selecting a suitable region of interest for the mean of the control distribution was established. Similarly the results of this phase of the experiment indicate a means for which the adequacy of the data can be measured. The question of adequacy, or rather when the analyst has "enough" data, has carefully been avoided to this point. The reader will recall that the sample size of 100 utilized throughout the experiment was somewhat arbitrarily selected. Since the smoothness of the behavior of $I(\mu)$ and $W(\mu)$ as observed in Figures 4.8 and 4.9 is clearly a function of the variance of the control distribution, σ_f^2, and since smoothness is generally considered a desirable attribute, a measure of smoothness as observed in $I(\mu)$ should provide information as to the adequacy of the data. For the observed "unsmooth" behavior corresponding to small σ_f^2, it is clear that the estimates provided by either $I(\mu)$ or $W(\mu)$ have little real meaning: $I(\mu)$ either grossly underestimates or grossly overestimates the expected response while $W(\mu)$ simply takes on values of the observed responses, $Y^{(i)}$. For larger σ_f^2, the resulting "smooth" estimates provided by both $I(\mu)$ and $W(\mu)$, such as seen in Figure 4.7 and other earlier plots can yield very meaningful information. Thus for this latter case, the sample size of 100 is sufficient for analysis, while for small σ_f^2, a sample size of 100 is entirely inadequate.

It would be desirable to be able to say that analysis of a response which is a function of n process parameters requires N_n observations. The results just discussed, however, obviously preclude such a simple tabulation. Rather it seems that each set of data need stand on its own merit, and that only after initial investigation will its adequacy, or lack of it, become apparent. Recalling the analysis procedure developed in Chapter 2, the analyst must somehow establish the optimal control capability, $f(x;\mu)$ and σ_f^2, before subjecting his data for analysis. Since $f(x;\mu)$ will generally be presumed normal, his only real task is in estimating σ_f^2. If after initial investigation his data appears inadequate, in the sense discussed here, rather than discard the data or this analysis approach, the analyst should review his determination of $f(x;\mu)$ and σ_f^2. If $f(x;\mu)$ was a "seat-of-the-pants" guess, it is likely that the assumed capability was overly optimistic and that such a control could not be realistically practiced, anyway. If he relaxes the "optimal" capability by increasing σ_f^2, he may not only be approaching a more realistic level of control but will also be smoothing the behaviors of the statistics $W(\mu)$ and $I(\mu)$. A seemingly insignificant increase in σ_f^2 might be sufficient, but should the results still indicate inadequate data σ_f^2 could be increased still again so that "meaningful" estimates can be obtained from any set of data. However, caution is in order since oversmoothing by increasing σ_f^2 well beyond the initial estimate may yield results without relevancy to the original problem, particularly if the response surface is complicated. In this case oversmoothing may

obscure legitimate local maxima.

While "smoothness" of the statistic $I(\mu)$ can determine adequacy of the data and inadequate data might be salvaged by increasing σ_f^2, determination of when the behavior is or is not "smooth" is no simple matter. When the response is a function of only one or two parameters, the results may be graphed and a subjective measure of smoothness made. Such determination is not possible should there be three or more process parameters and some quantification of smoothness is thus needed. No attempt will be made here to find this quantification other than to state that a simple measure might be the number of relative maxima encountered for $I(\mu)$ while searching the response surface. This is probably what the analyst looks for in his subjective evaluation of the one dimensional case, anyway. The actual determination of smoothness, hence the adequacy of the data, will be left to the discretion of the analyst.

Summarizing the results of this third phase of the experiment, the observed behaviors of $I(\mu)$ and $W(\mu)$ and their predictive capabilities can be affected significantly by the analyst's choice of $f(x;\mu)$ and σ_f^2. A smaller σ_f^2 will also likely increase the effect of ε as studied in phase 2. Also while no standard exists for determining when the analyst does or does not have enough data, $I(\mu)$ can be used to measure the adequacy, and an apparently inadequate set of data may be salvaged by revising the estimate of σ_f^2. Nothing was found to repudiate the claim that the optimizing criteria based on $W(\mu)$ was the best of those considered.

5.0 Reconstruction of the Distribution of the Data: g(x)

The interpretation given in Chapter 2 to the collection of historical data which is available for analysis was that of a random sample. It is this interpretation which makes it necessary to develop or select a technique for identifying some population for which the data comprise a "representative" sample. While the study of Chapters 3 and 4 assumed knowledge of the actual distribution of the data (recall samples were simulated from known distributions), such luxuries are not available in an actual application of the procedure. This chapter therefore presents and evaluates a technique which shows promise for reconstruction of a generalized distribution from a collection of historical data.

5.1 Selecting a Technique

Before selection of a reconstruction technique can be made, the needs that the technique must serve have to be established. The most important question regarding these needs are whether it is necessary to justify the reconstructed population by recognizing and explaining the mechanisms responsible for the data, or whether it is sufficient to obtain only realistic relative frequencies for the sample values. Reviewing the analysis procedure developed in Chapter 2, nowhere is it necessary to characterize the mechanisms behind the sample although such characterization would undoubtedly prove useful in establishing the optimal control capabilities of the process. Rather it is seen that only the relative frequencies of the sample observations are required. These frequencies, along with the probability

density function of the optimal control distribution determine the weights to be assigned to each of the observed responses in calculating $W(\mu)$, $I(\mu)$, and $S_I^2(\mu)$. Consequently the requirements for the reconstruction techniques consist only of obtaining realistic relative frequency estimates for each of the observations. No consideration need be given to what the resulting estimates of the frequencies imply about the population from which the data came.

Several alternatives for reconstructing this population are available. One might be to hypothesize a particular family of distributions, estimate the required parameters, and perform suitable tests of hypothesis. Such a technique does not offer much promise in the kinds of analyses with which this thesis is concerned because, remembering the implied nature of historical data, few families, if any, are general enough to encompass the wide range of possible shapes (for example, multiple modes) which are likely to be encountered. Another technique could be called the standard histogram approach. It involves fitting curves to the peaks of the cells of the histogram. Such an approach, however, is sensitive not only to the size but also the positioning of the cells, particularly for smaller sample sizes. This results in widely varying estimates of the population probability density function from but a single sample. Still another alternative involves curve fitting of the empirical cumulative distribution function and then differentiating the result to obtain the relative frequencies of the observations. Limited success was achieved with this approach by fitting high degree poly-

nominals, but while multiple modes can be detected, large negative estimates of the frequencies were occasionally obtained--particularly in extreme values.

Common to all these approaches to reconstruction of a population is an underlying assumption that the analyst is capable of determining a suitable family of functions for handling the wide variety of populations which might occur, whether for curve fitting or parameter estimation. It would be desirable to have available a technique which is entirely non-parametric in nature for reconstructing a population for which the data can reasonably be said to form a representative sample. Boneva, Kendall, and Stefanov have presented a technique which very nearly meets this requirement.

5.2 Histograms and Deltasplines

"The situation which we have in mind is ... that in which one is presented for the first time with a random sample from a distribution about which nothing is known at all...."[1] This statement accurately sums up both the content of the paper by Boneva et al and the situation being dealt with in this chapter. While Boneva, Kendall, and Stefanov's primary interest is in development of a smoothing technique for answering qualitative questions about the shape of the distribution, this author has enjoyed considerable success employing their techniques quantitatively. At the heart of their technique is the treatment of each sample observation as a separate, equal, and independent building block for reconstructing the distribution of the data. At the risk of oversimplification, an interpretation of the

principles behind their technique is presented here.

Consider the ordinary histogram which is constructed by partitioning the real line into a series of adjacent, equal width "cells" and setting the height of each cell to represent the count of the number of observations which fall within its limits. Let N be the total number of observations comprising the histogram. This same histogram could be constructed from an entirely different approach. Instead of constructing a single histogram containing all N observations, N distinct "unit" histograms are formed; one to be associated with each sample observation. Each of these unit histograms share these properties:

1. Exactly one cell is occupied.

2. The height of the occupied cell is 1, all others 0.

Let H be the histogram containing all N observations and H_i be the unit histogram associated with the i^{th} observation. Then if the positions and widths of cells are common to all histograms H_i as well as H, the unit histograms "add" linearly, as in Figure 5.1 so that

$$H = \sum_{i=1}^{N} H_i \qquad 5\text{-}1$$

Thus each observation contributes a separate, equal, and independent building block to the entire histogram H.

Consider the sequence of events which resulted in the formation of H in the above manner:

1. A cell width, C, had to be chosen common to all unit histograms.

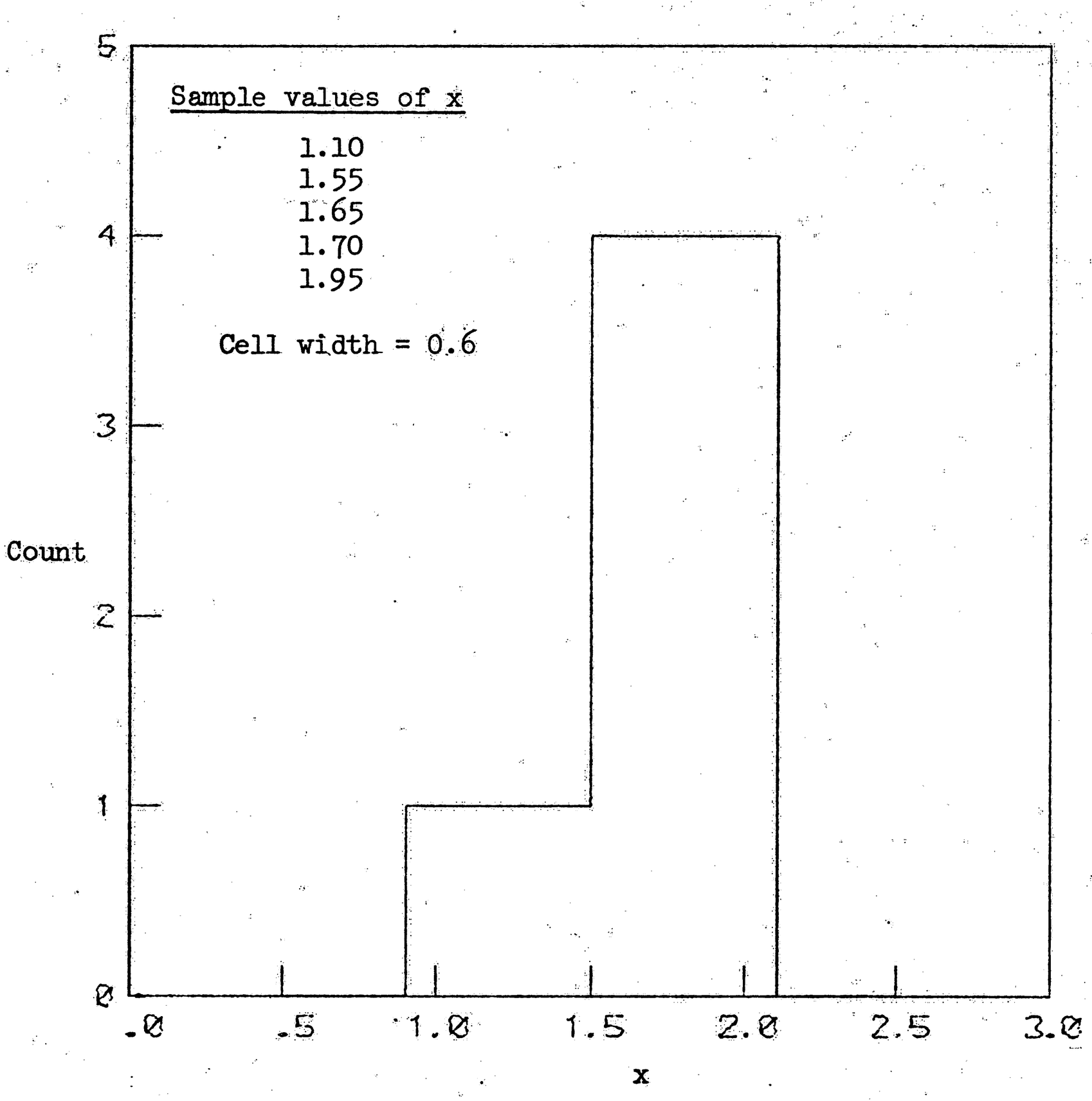


Figure 5.1 - Ordinary histogram

2. The real line in each unit histogram had to be partitioned into cells of width C such that cells were identically positioned in each H_i.

3. The height of the occupied cell in each H_i was set to 1, all others to 0.

4. The unit histograms, H_i, were "added" to produce H.

The resultant H cannot be expected to be "smooth" in the usual sense of the word since each of the unit histograms has two discontinuities. However, a slight modification to the above sequence will have a smoothing effect on H, while not yet producing a "smooth" histogram. Consider this new sequence:

1. Choose a cell width, C, common to all unit histograms.

2. Partition the real line of each unit histogram into cells of width C, but instead of positioning the cells identically for each H_i, position the cells in the i^{th} unit histogram so that the i^{th} observation is centered in the occupied cell.

3. Set the height of the occupied cell in each H_i to 1, all others to 0.

4. Add the unit histograms, H_i, to produce H.

The smoothing effect of this modification is seen in Figure 5.2.

Boneva, Kendall, and Stefanov, recognizing the appeal of a "smooth" histogram, carried this concept one step further. They sought and found a transformation of the unit histogram, H_i, which not only preserves the linearity of the addition of the H_i's, but

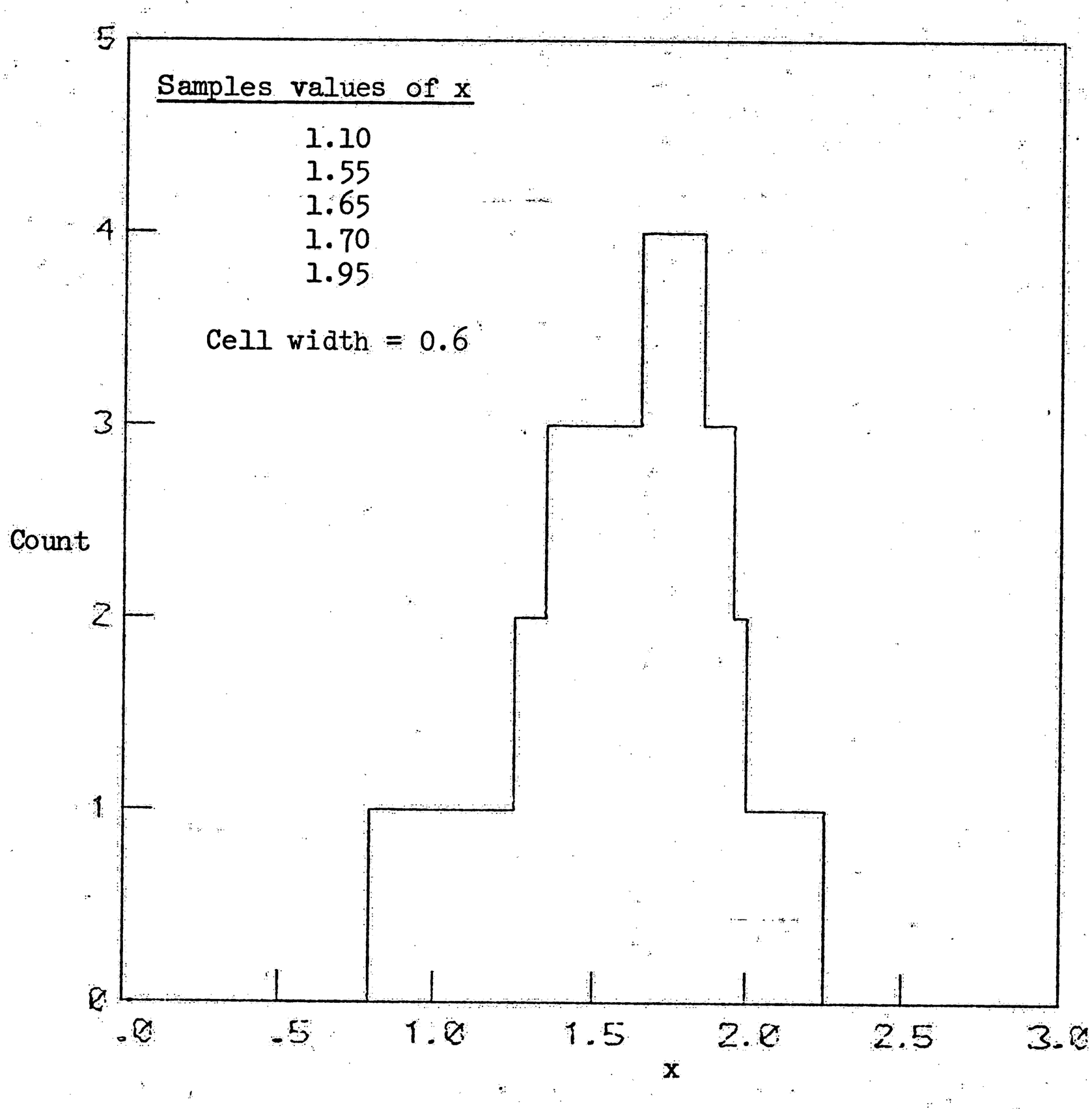


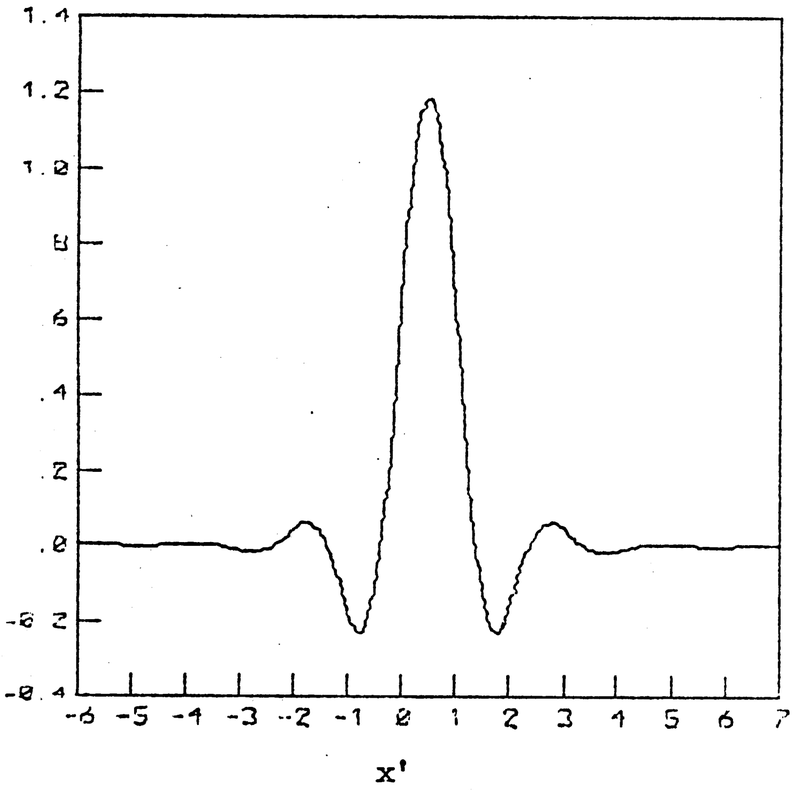
Figure 5.2 - Ordinary histogram "smoothed"

also yields a sum of the transformed H_i's which is smooth in every sense of the word. The transformation is the deltaspline shown in Figure 5.3. The deltaspline is a piece-wise parabolic function with each piece occupying one cell of the unit histogram. If the deltaspline is centered over the occupied cell in the unit histogram and integrated piece-wise (or cell-wise), the results are 0 for every unoccupied cell and 1 for the occupied cell. The analogy between deltasplines and unit histograms is clear. If H_i' is the deltaspline associated with the transformed unit histogram H_i, then H', where

$$H' = \sum_{i=1}^{N} H_i' \qquad 5\text{-}2$$

approximates in a "smooth" way the actual shape of the frequency distribution of the data. If H' is further normalized by dividing by the number, N, of deltasplines comprising the sum, then estimates of the relative frequencies of the sample values can be taken directly from a plot of H'/N.

Thus the deltaspline allows a nearly parametric-free reconstruction of an estimate of the population underlying the sample. The only parameter required to apply the technique is the cell width, C. This parameter will determine the number of modes exhibited by the smooth approximation of the sample distribution and hence also determines the frequency estimates for each observation as taken from the plot of H'/N. The parameter C will not only determine the number of modes in the approximation of the population but may also result in an occasional negative frequency estimate because of the behavior



Cell		Cell-wise Parabola Definition*		
x'_{min}	x'_{max}	Const.	Coeff of x	Coeff of x^2
-6	-5	2.35×10^{-4}	8.13×10^{-4}	-1.92×10^{-3}
-5	-4	-8.76×10^{-4}	-3.03×10^{-3}	7.18×10^{-3}
-4	-3	3.27×10^{-3}	1.13×10^{-2}	-2.68×10^{-2}
-3	-2	-1.22×10^{-2}	-4.22×10^{-2}	1.00×10^{-1}
-2	-1	4.55×10^{-2}	1.58×10^{-1}	-3.73×10^{-1}
-1	0	-1.70×10^{-2}	-5.88×10^{-1}	1.39×10^{0}
0	1	6.34×10^{-1}	2.20×10^{0}	-2.20×10^{0}
1	2	6.34×10^{-1}	-2.20×10^{0}	1.39×10^{0}
2	3	-1.70×10^{-1}	5.88×10^{-1}	-3.73×10^{-1}
3	4	4.55×10^{-2}	-1.58×10^{-1}	1.00×10^{-1}
4	5	-1.22×10^{-2}	4.22×10^{-2}	-2.68×10^{-2}
5	6	3.27×10^{-3}	-1.13×10^{-2}	7.18×10^{-3}
6	7	-8.76×10^{-4}	3.03×10^{-3}	-1.92×10^{-3}

Occupied cell is $x'_{min} = 0$, $x'_{max} = 1$

Parabola definition is with respect to local coordinates within each cell.

* - approximate.

Figure 5.3 - Definition of the deltaspline

of the deltaspline. Such a result does not concern Boneva et al since it has little effect on the answers to their qualitative inquiries. But this fact could prove bothersome in a quantitative application.

As a means of evaluating quantitative application of the techniques of Boneva, Kendall, and Stefanov, a study has been made and will now be presented.

5.3 A Study and Evaluation of the Deltaspline as Applied to Reconstructing a Population for the Data

The purposes of this study are threefold:

1. Determine the ability of the deltaspline to reconstruct a known population;
2. Develop a procedure for obtaining a "best" estimate of an unknown population; and
3. Determine the effects on $I(\mu)$, $W(\mu)$ and estimates of μ^* in the first phase of the experiment described in Chapter 3 if values for $g(x^{(i)})$ are taken from the deltaspline approximations instead of the known population.

The first question to be considered is the ability of the deltaspline approach to reproduce a known distribution. The parameter, C, acts as a smoothing parameter much in the same way that σ_f^2 did as presented in the analysis of phase 3 in Chapter 4. As was discussed there, undersmoothing results in "rabbits" or an unreasonable number of modes while oversmoothing may obscure or blend two or more legitimate modes. This effect, using the deltaspline, is displayed in

Figure 5.4. Thus the analyst must proceed with caution in selecting a cell width. How does he know when H'/N is not smooth enough, properly smoothed, or too smooth? If he knows what distribution he is reconstructing, it is easy to tell when the estimate is under- or oversmoothed. This prior knowledge will be employed here to find those cell widths, C, which for a number of samples from several distributions provide "best" approximations of the sampled distributions. The results are hoped to indicate how C might be selected for a sample from an unknown distribution in addition to answering the question of the ability of the deltaspline to reproduce a known distribution.

One first must define "best" approximation. If the sample is "representative" of the population from which it was drawn in the sense that it can pass a Kolmogorov-Smirnov goodness of fit test, then the "best" smoothed estimate of the underlying population could be that estimate which minimizes the sum square deviation of the frequency estimates about their actual frequencies. Using this definition of "best" approximation, twenty random samples were analyzed: five from each of four different populations. Those populations included the bimodal and skewed distributions of Chapter 3 (equations 3-9 and 3-10), a uniform distribution (equation 5-3), and a normal distribution (equation 5-4).

$$\text{uniform:}\quad g_3(x) = \begin{cases} .25 \text{ for } .5 \times 4.5 \\ 0 \text{ else} \end{cases} \qquad 5\text{-}3$$

$$\text{normal:}\quad g_4(x) = \phi(x;2.5,.7225) \qquad 5\text{-}4$$

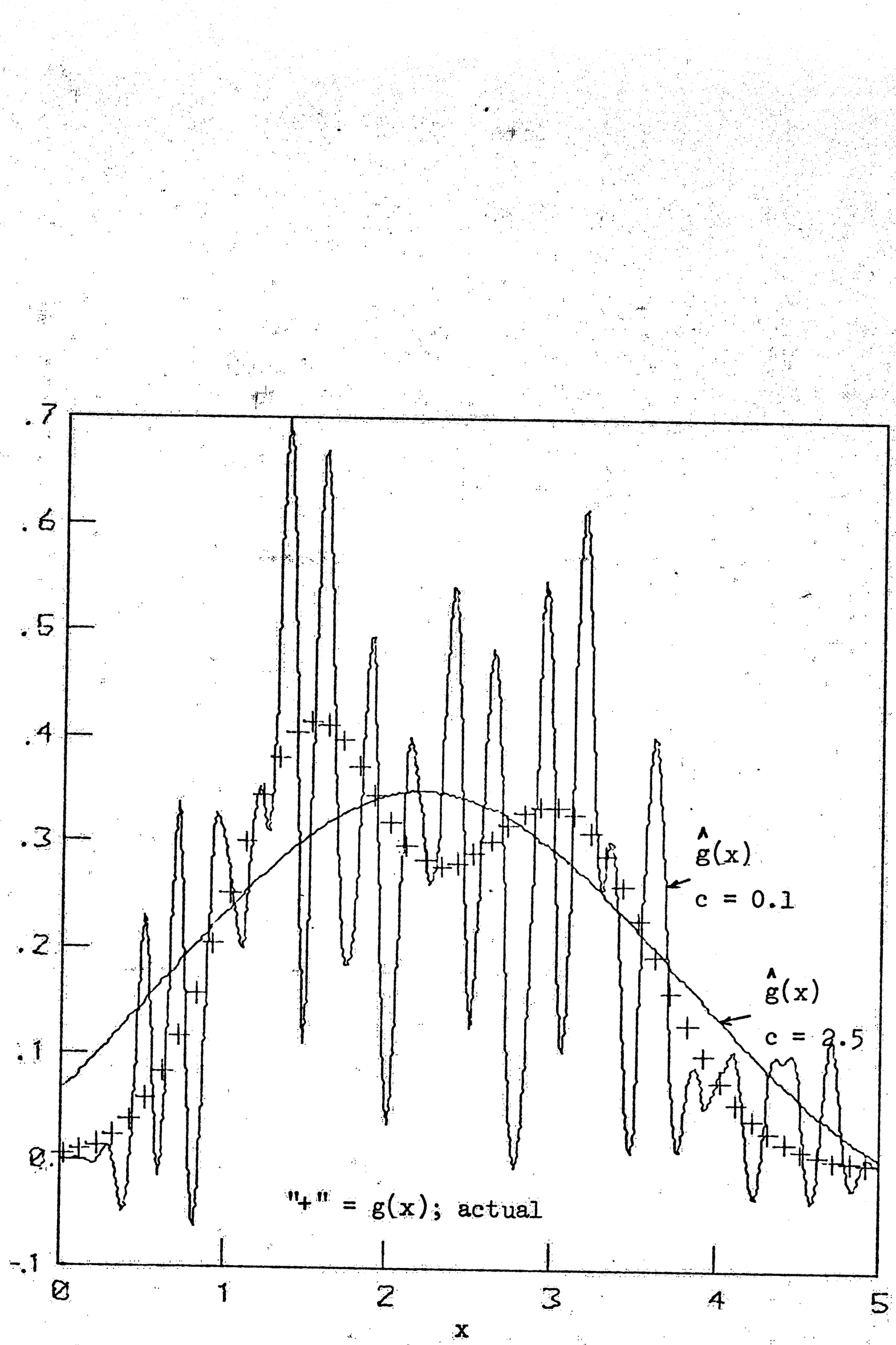


Figure 5.4 - Effect of smoothing parameter, C, on deltaspline constructed estimates

The sum squared deviations for the estimated frequencies of the observations about their theoretical frequencies were calculated for cell widths ranging from .1 to 2.5 for each sample. The minimum sum of squares obtained with each sample is found in Table 5.1 along with its associated cell width. In same cases, the sum of squares was observed to be rather flat near the minimum point, so as an indication of the sensitivity of the reconstruction process to cell width, a range of values for C in which the total sum of squares was increased not more than 20% from the minimal value is also tabulated in Table 5.1. The ability of the deltaspline approach to reconstruct a known distribution can be evaluated by plotting the resulting estimates of the population for the largest and smallest values of SS_{min} in each distribution and comparing the results to the theoretical, or actual, populations from which the samples were drawn. This is done in Figures 5.5 - 5.8. It is observed that the smallest values of SS_{min} yield excellent fits of the underlying population for both the bimodal and normal distributions. The fits obtained with the smallest values of SS_{min} for the skewed and uniform populations are not encouraging. The difficulty in reconstructing these latter distributions is a result of the combination of wide, slowly varying segments as in the center of the uniform distribution and left side of the skewed distribution and narrow, rapidly changing segments as in the right side of the skewed distribution and both ends of the uniform distribution. The wide, slowly varying sections suggest using a wide cell width for a broader deltaspline while the narrow, rapidly changing sections

Bimodal				Skewed		
SS_{min}	C_{min}	Limit on C's*	Sample #	SS_{min}	C_{min}	Limit on C's*
.3062	.80	.71- .97	1	.9587	.70	.54-1.03
.0147	.70	.68- .81	2	.4243	.60	.48- .73
.5352	.90	.66-1.40	3	.3155	.50	.38- .56
.1445	.90	.78-1.09	4	.6024	.50	.38- .71
.3240	1.00	.62-1.43	5	.2845	.50	.45- .63

Normal				Uniform		
SS_{min}	C_{min}	Limit on C's*	Sample #	SS_{min}	C_{min}	Limit on C's*
.0262	1.5	1.40-1.55	1	.0817	.90	.77-1.01
.0257	1.4	1.29-1.54	2	.1008	.90	.77-1.04
.1560	1.7	1.41-1.91	3	.1781	1.60	1.36-1.80
.0423	1.4	1.25-1.59	4	.0965	1.50	1.35-1.61
.0957	1.8	1.55-2.00	5	.1480	1.50	1.28-1.69

*Range of C for which $|SS_{min} - SS_C| \leq .10\ SS_{min}$

Table 5.1 - Best fits of reconstructed population to known population

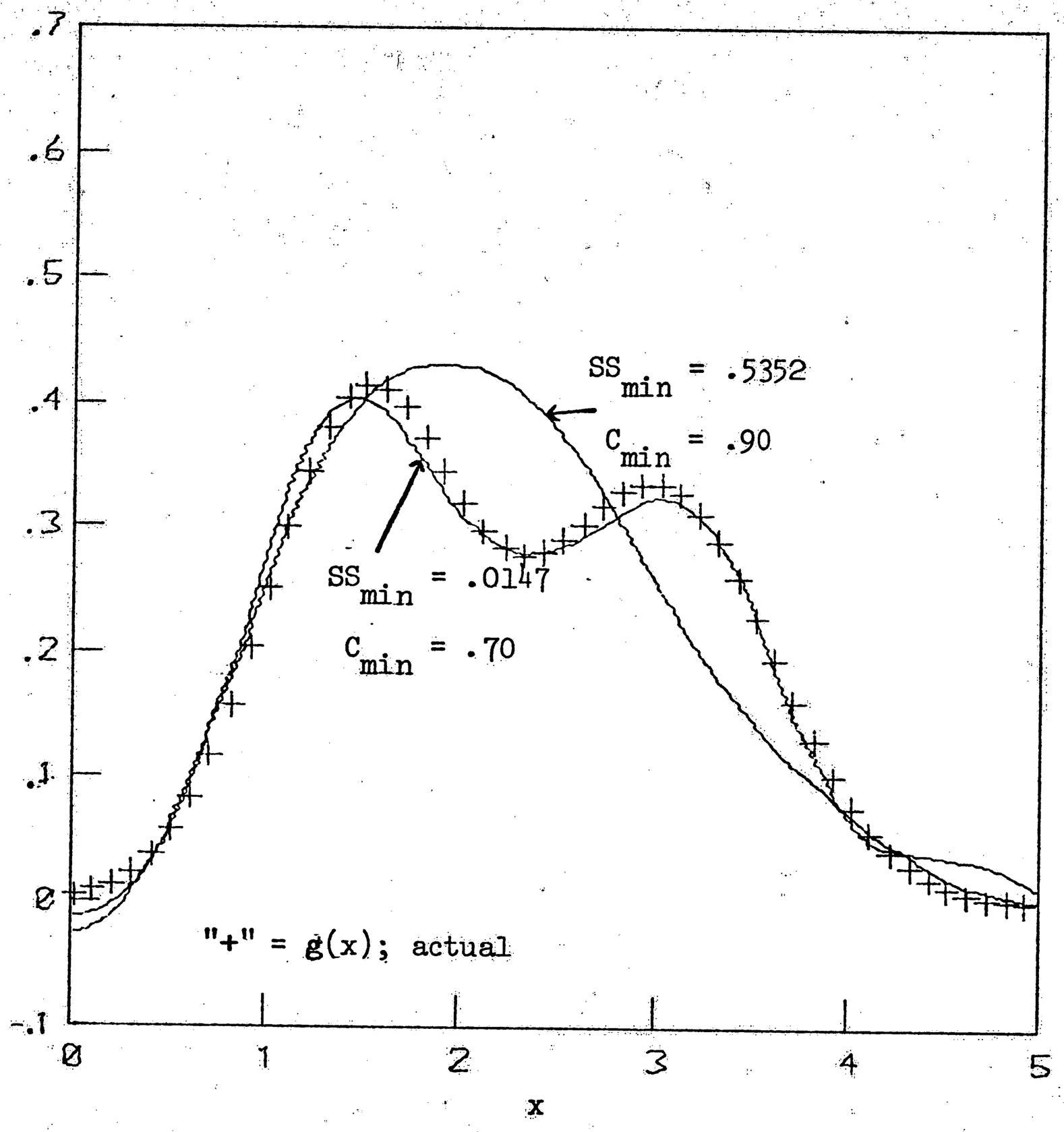


Figure 5.5 - Reconstructed bimodal distribution

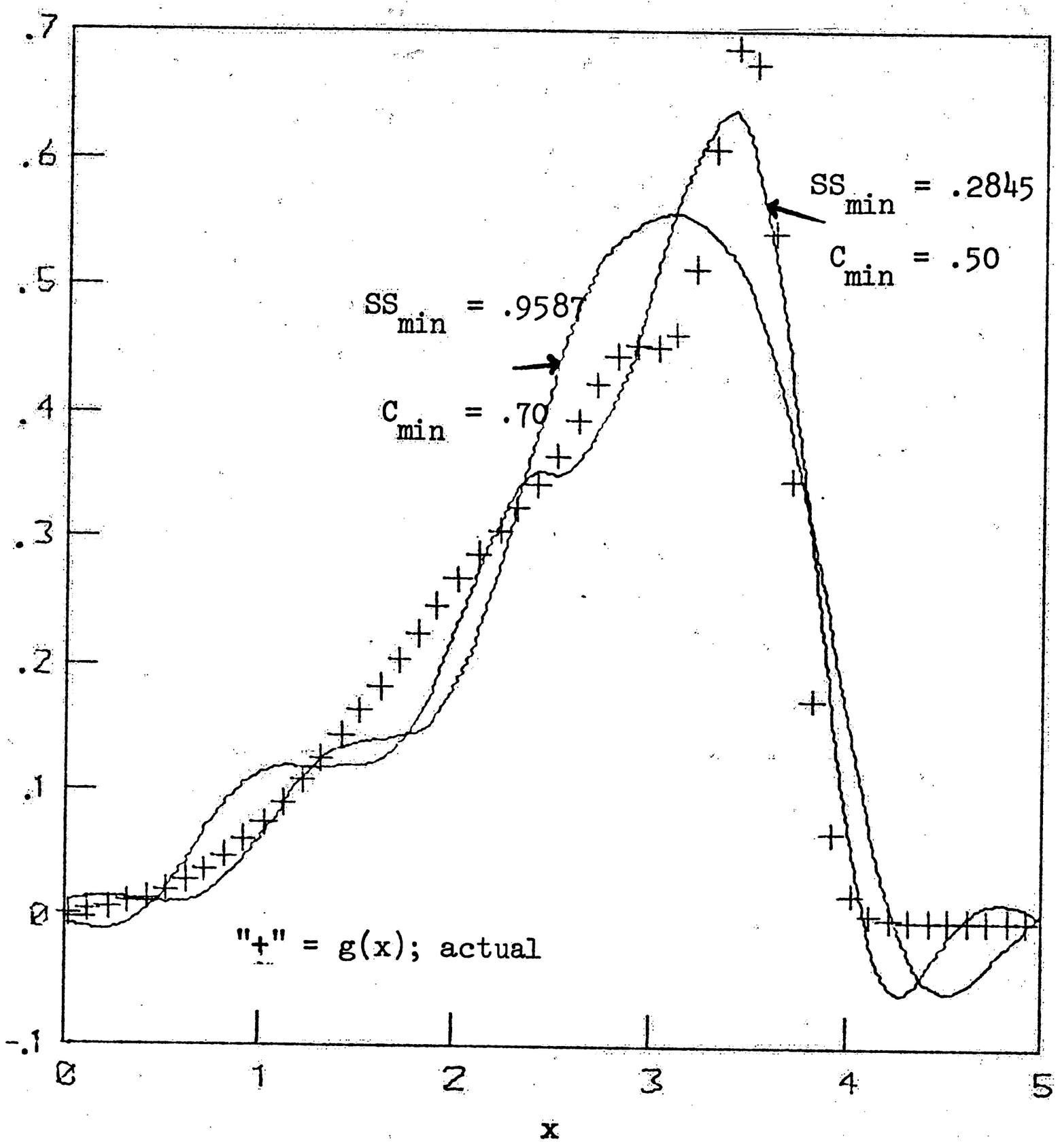


Figure 5.6 - Reconstructed skewed distribution

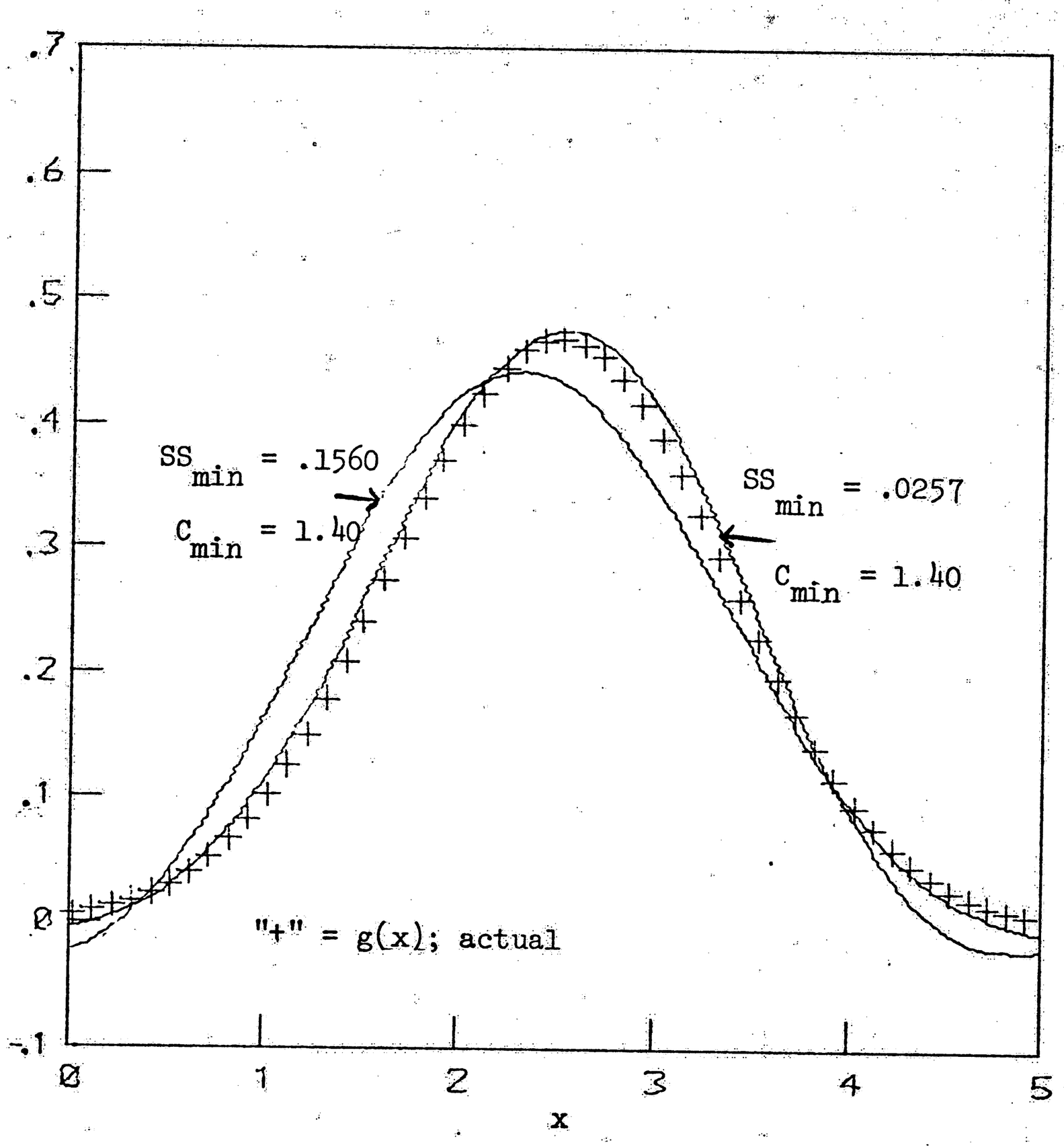


Figure 5.7 - Reconstructed normal distribution

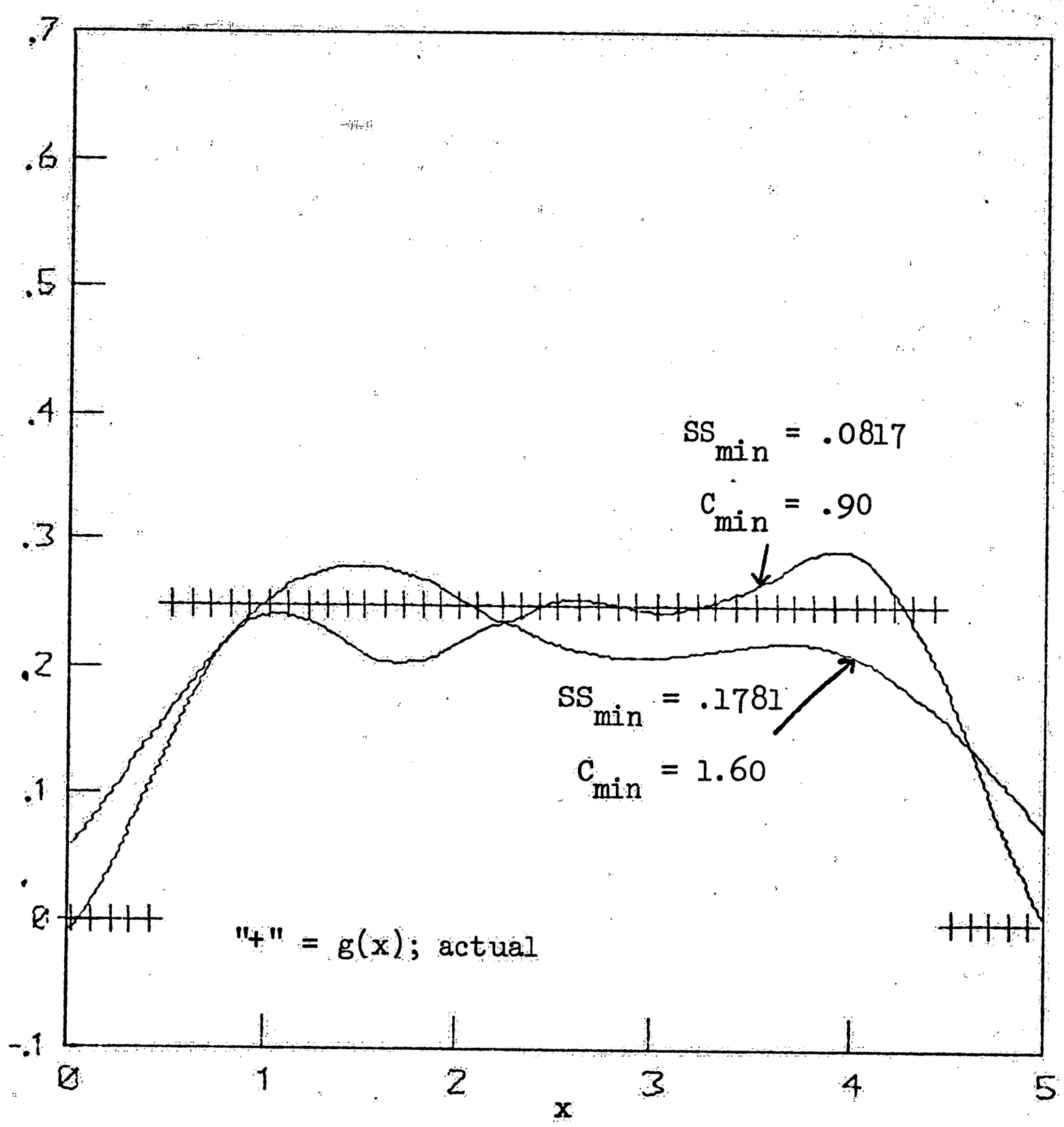


Figure 5.8 - Reconstructed uniform distribution

require narrow cell width for resolution. The results in Figures 5.6 and 5.8 are the best compromises to these conflicting requirements. While not as satisfying as the best fits for the normal and bimodal populations, the fits of the skewed and uniform distributions for the smallest values of SS_{min} are reasonable approximations of the underlying populations. The deviations from the actual shape of the distribution observed for the largest value of SS_{min} in all cases can be attributed more to sample variations than to the inability of the deltaspline approach to reconstruct the underlying distribution of the data.

Attention is now turned to the second purpose of the study which was development of a general procedure for reconstructing an unknown population from a single sample. This is equivalent to selecting a rule for determining what cell width, C, to use. The simplest such rule which may be devised would be to select C such that there is some fixed number, M, of cells between the minimum and maximum observed sample values, or such that

$$C = R(x^{(i)})/M \qquad 5\text{-}5$$

where $R(x^{(i)})$ is the sample range of observed values. The data in Table 5.1 was employed in making Table 5.2 which shows $R(x^{(i)})$, the number of cells, M, within that range for C_{min}, and the bounds of M using the limits on C for each sample from Table 5.1

A one-way analysis of variance was performed on the values $M(C_{min})$ to test their independence of the distributions considered. If the hypothesis could not be rejected, then an average of the

Bimodal			Sample #	Skewed		
$R(x^{(i)})$	$M(C_{min})$	Limits on M		$R(x^{(i)})$	$M(C_{min})$	Limits on M
3.90	4.9	4.0 - 5.5	1	3.60	5.1	3.5 - 6.7
4.19	6.0	5.2 - 6.2	2	3.68	6.1	5.0 - 7.7
4.08	4.5	2.9 - 6.2	3	3.41	6.8	6.1 - 9.0
3.71	4.1	3.4 - 4.8	4	3.49	7.0	4.9 - 9.2
3.78	3.8	2.6 - 6.1	5	3.33	6.7	5.3 - 7.4

Normal			Sample #	Uniform		
$R(x^{(i)})$	$M(C_{min})$	Limits on M		$R(x^{(i)})$	$M(C_{min})$	Limits on M
4.83	3.2	3.1 - 3.5	1	3.97	4.4	3.9 - 5.2
4.00	2.9	2.6 - 3.1	2	3.94	4.4	4.0 - 5.2
4.74	2.8	2.5 - 3.4	3	3.97	2.5	2.2 - 2.9
3.66	2.6	2.3 - 2.9	4	3.78	2.5	2.3 - 2.8
3.43	1.9	1.7 - 2.2	5	3.89	2.6	2.3 - 3.0

Table 5.2 - Number of cells within range of data

tabulated $M(C_{min})$ might be used with equation 5-5 for reconstructing an unknown population. However, the result of the analysis was a highly significant F-ratio of 15.2, thus rejecting the hypothesis. Even if it had been found that 5-5 seemed to hold across the distribution considered, it would be anticipated that the rule require an additional parameter--sample size. For sample sizes greater than 100, it could be expected that better resolution in the shape of the underlying population would be possible thus permitting the use of smaller cell widths or a larger number of cells within the range of data. Likewise for samples of less than 100 observations less resolution is possible and wider cell widths would be required to obtain a "best smooth" approximation of the underlying population for the data. So it appears that a rule as simple as 5-5 is not likely to be of any general use.

Another approach to solving the general problem of selecting a cell width was tried, but it was doomed to failure from the outset. However it did provide a potentially useful result. The approach was to select a number of potentially "optimal" cell widths, determine the deltaspline approximation to the underlying population for each, calculate the cumulative distribution function from the deltaspline approximation and do a kind of Kolmogorov-Smirnov analysis on the data to determine the maximum deviation of the empirical cumulative distribution function from the reconstructed function. The candidate cell width which minimized the maximum deviation would then be selected as "optimal." The flaw in this approach becomes apparent

if an infinitesimal cell width is among the candidates. Such a value for C concentrates all the area of the approximated density function at the sample values so that, when integrated, the cumulative distribution function will increase stepwise and very nearly equal the empirical cumulative distribution function about which the deviations are measured. Thus the infinitesimal cell width will always appear "optimal." This result can be generalized to a set of realistic cell widths where it was found in all cases that the "optimal" width was the minimum cell width in the set. Thus the approach gives no insight into how the cell width should be selected when analyzing a sample from an unknown distribution. The potentially useful result which can be derived from this approach of integrating the reconstructed distribution is possible detection of oversmoothing. While the smallest candidate cell width was always "best," the largest candidate was always "worst." Referring back to Figure 5.4 it is observed that oversmoothing had the effect of increasing the length of the tails of the distribution as well as obscuring legitimate modes. Continuing to smooth forces more and more of the area under the population into the tails so that if the deltaspline approximation is integrated up to the minimum sample value, x_{min}, and the difference between that value and the value of the empirical cumulative distribution function noted, the difference will continue to increase. Similarly for the maximum sample value, x_{max}, and the area under the approximation for $x > x_{max}$. No attempt will be made here to establish a general criterion for determining when the ap-

proximation is oversmoothed other than to suggest that the behavior here described could be interpreted in such a way as to yield workable criteria for detecting an oversmoothed estimate.

It appears then that no simple, straightforward procedure exists to aid in the selection of the smoothing parameter, C, for reconstructing the distribution of the data. Rather, as in Chapter 4, the determination of when the behavior is sufficiently smooth will be left to the discretion of the analyst. As an initial guess, however, when working with samples of approximately 100 observations, the analyst can draw upon the results in Table 5.2 and select a cell width equal to about 1/4 or 1/5 the range of the data and adjust C from there. The behavior discussed in the preceding paragraph can then be used as a check against oversmoothing while too many modes acts as a check against undersmoothing.

The third purpose of this study is now addressed. It is an evaluation of the impact on the behaviors of $I(\mu)$, $W(\mu)$, and the various optimizing criteria resulting from the use of estimated instead of actual values for the $g(x^{(i)})$. The data analyzed is the same group of samples utilized in the first phase of the experiment described in Chapter 3. This time, however, values for $g(x)$ which were estimated by a deltaspline approximation of the underlying population were used. To determine quantitatively the impact of this change on the behavior of the statistics $I(\mu)$ and $W(\mu)$ within the suitable region of interest (4-4 or 4-5), curves were plotted for each of the statistics using both actual and estimated frequen-

cies. Then the mean squared deviation about the theoretical expected response for a number of equispaced points within the appropriate region of interest was calculated for each curve. A comparison of the resulting values of these mean squared deviations (MSD) for the actual and estimated frequencies then provides a quantitative measure of the effect of the estimated frequencies.

Two points regarding the analysis should be made at this time. The first concerns selection of cell widths for each of the samples. Rather than make 48 separate determinations of "best" cell width, one for each replication of the experiment, the consistency of best cell widths within distributions as seen in Table 5.1 was used to determine a single "optimal" cell width for each of the two distributions which were sampled. The "optimal" cell widths used were just the simple averages of the five values of C_{min} for each distribution. Thus .85 was used as the cell width for all bimodal samples and .55 for skewed samples.

The second point concerns the treatment of negative frequency estimates. Recalling the earlier discussion of the deltaspline, it is not impossible that the technique produces an occasional negative estimate for one or more of the observations in a sample. Clearly this occurrence cannot be completely ignored in the calculations of $I(\mu)$ and $W(\mu)$ and possible methods for treating these occurrences need to be considered.

To grasp the magnitude of the problem regarding negative frequency estimates, the number of samples from the first phase of the

experiment in Chapter 3 containing such estimates after reconstruction using the "optimal" cell widths was counted. Of the 48 samples in the experiment--24 of which were bimodal and 24 skewed--only 8 contained any negative frequency estimates: 6 bimodal and 2 skewed. Of these 8 samples only one contained more than a single negative estimate and it contained only 2. Thus a total of only 9 out of 4800 observations had negative frequencies estimates as a result of using the appropriate "optimal" cell width--either .85 or .55--in a deltaspline reconstruction of the distribution of the data. This is hardly an alarming total and their treatment can be handled on an exception level.

The simplest treatment for a negative frequency estimate might be setting that value to 0 since its actual value is likely to be very small anyway. However, $g(x)$ is a divisor in the formula for calculating $I(\mu)$ and $W(\mu)$ so such treatment is not feasible as it results in division by zero. Another possible treatment for the negative estimate also utilizes the likelihood that the actual frequency was very small. Under this treatment, any observation which has a negative frequency estimate would be treated as an outlier and eliminated from the sample. The calculations of $I(\mu)$ and $W(\mu)$ would then be made with just the remaining observations. Such a treatment is feasible and, considering the infrequent occurrences of negative estimates, not unreasonable. Another treatment which is also feasible would be to change the cell width, either increasing it or decreasing it by small increments until all sample frequency estimates are posi-

tive. Whereas the former treatment of considering observations with negative estimates as outliers leaves the remaining frequency estimates for the sample unchanged, this latter treatment will produce new frequency estimates for all observations with each incremental change in cell width. Both of these feasible alternatives were evaluated with respect to each other by considering their effects on the eight samples which were found to contain negative frequencies for the "optimal" cell widths. The mean squared deviation (MSD) as defined earlier provided the measure for comparison. The results with respect to both $I(\mu)$ and $W(\mu)$ are listed in Table 5.3 along with the MSD resulting from use of the actual, or known, $g(x)$. Scanning the table for general information it appears that the first alternative of eliminating all observations from the sample which have negative frequency estimates provides the more consistent effect as the range of MSD's for the other treatment was significantly greater for both $I(\mu)$ and $W(\mu)$. It is also seen that for six of the eight samples for each of $I(\mu)$ and $W(\mu)$ (though not the same six samples) the first alternative is clearly the better treatment. For the other two samples in each case, while changing C is the better treatment, it does not provide a radically smaller MSD. Thus it is concluded that little would be lost if the first treatment were used on all samples containing negative frequency estimates. Also it is computationally more expensive to employ the second alternative since the number of computations required to obtain all positive frequency estimates is highly variable as seen in the

$I(\mu)$

Sample	Design Cell	M.S.D.x10^{-4} Actual g(x)	M.S.D.x10^{-4} - Estimated g(x)	
			Eliminate if g(x)≤0	Change C: By How Much
1	Quad-bimodal	42.8	1172.0	835.0: -.100
2	Quad-bimodal	18.6	136.0	2450.0: +.050
3	Quad-bimodal	57.8	163.0	783.0: -.275
4	Quad-bimodal	48.2	148.0	2760.0: -.050
5	Quad-skewed	65.5	170.0	1130.0: -.125
6	Quad-skewed	181.0	131.0	62.0: +.225
7	Lin, decr-bimodal	27.8	97.0	1079.0: -.425
8	Lin, incr-bimodal	45.0	67.5	2015.0: -.150

$W(\mu)$

Sample	Design Cell	M.S.D.x10^{-4} Actual g(x)	M.S.D.x10^{-4} - Estimated g(x)	
			Eliminate if g(x)≤0	Change C: By How Much
1	Quad-bimodal	1.68	2.29	17.50: -.100
2	Quad-bimocal	1.11	2.46	27.10: +.050
3	Quad-bimodal	1.53	6.07	3.60: -.275
4	Quad-bimodal	4.97	1.18	35.00: -.050
5	Quad-skewed	2.82	6.60	12.30: -.125
6	Quad-skewed	5.05	2.64	1.60: -.225
7	Lin, decr-bimodal	0.75	0.54	4.40: -.425
8	Lin, incr-bimodal	0.43	0.39	5.10: -.150

Table 5.3 - Treatment of negative frequency estimates

"how much" column. Each increment of $\pm.025$ required recalculation of all estimates so that for the changes of -.425 for sample 7, a minimum of 16 frequency recalculations were required--if the analyst knew to decrease rather than increase C! Thus the first feasible treatment was adopted for dealing with negative frequency estimates. Comparing the MSD's using the estimated $g(x)$ to those obtained for the actual $g(x)$ caused some concern, particularly for $I(\mu)$, as it seems that the estimated $g(x)$ significantly increased the MSD in all but one sample. The MSD for the estimated $g(x)$ compares more favorably with that of the actual $g(x)$ for $W(\mu)$, however, and this same comparison was then made on the remaining 40 samples for which no negative frequency estimates occurred to see if these trends continued. The fact that the MSD's in all columns were higher for $I(\mu)$ than for $W(\mu)$ should have been expected recalling the results for the analysis of phase 1 in Chapter 4.

A comparison of the MSD's for the actual and estimated $g(x)$'s for all replications of the first phase of the experiment is summarized in Table 5.4 for each of $I(\mu)$ and $W(\mu)$. The table shows for each of $I(\mu)$ and $W(\mu)$ the minimum, maximum, and average MSD for the actual and estimated $g(x)$'s. The series of three numbers which appear twice in the bottom of each cell represent the number of samples for which the MSD for the estimated $g(x)$ was less than the MSD for the actual $g(x)$, the number of samples for which there was less than a 10% difference in MSD's, and the number of samples within the cell for which the MSD's were increased using the estimated

Bimodal					Skewed			
I(μ)		W(μ)			I(μ)		W(μ)	
g(x) act.	g(x) est.	g(x) act.	g(x) est.		g(x) act.	g(x) est.	g(x) act.	g(x) est.
				QUADRATIC				
14.3	0.4	0.75	0.39	min.	14.5	4.5	0.32	0.09
81.8	1172.0	4.97	6.07	max.	181.0	170.0	5.04	6.60
45.4	205.4	1.61	1.89	ave.	58.6	59.5	1.91	1.21
4-0-4 ****		4-0-4 * ***			6-0-2 * *		5-0-3 * *	
				LINEAR, DECREASING				
3.2	1.8	0.25	0.07	min.	4.5	1.8	0.05	0.09
50.0	96.8	0.93	0.68	max.	138.1	14.1	0.55	0.36
28.6	24.6	0.68	0.29	ave.	39.6	5.9	0.23	0.18
4-0-4 *		7-0-1 *			7-0-1		2-4-2	
				LINEAR, INCREASING				
2.5	1.1	0.11	0.07	min.	7.3	0.4	0.18	0.00
105.0	218.6	0.79	1.75	max.	82.7	85.0	1.41	1.09
31.4	57.5	0.36	0.68	ave.	27.3	16.3	0.50	0.36
4-0-4 *		2-3-3 *			5-1-2		5-0-3	

Table 5.4 - M.S.D. summary - estimated g(x) vs. actual g(x)

g(x). The asterisks indicate into which group the samples which originally contained negative frequency estimates fall. Recall that those observations containing negative frequency estimates were dropped from the sample prior to analysis. The results indicate no particular loss of accuracy when the estimated frequencies replace the actual values in calculating the statistics I(μ) and W(μ). The average MSD over all 48 samples for W(μ) and and the actual g(x) was 0.88 while with the estimated g(x) and W(μ), the overall average MSD was decreased to 0.77. Repeating this comparison for I(μ) after throwing out the sample producing the MSD of 1172.0, the results are 38.5 with the actual g(x) and 38.4 using the estimated g(x) indicating no significant change. Thus on the average the estimated g(x) provided as good or better a fit of the curves for the statistics I(μ) and W(μ) to the theoretical expected response than did the actual values for g(x). While this is true on the average, the triples in the bottom of each cell indicate that for any given sample, a better fit of I(μ) with estimated frequencies occur 50% more often than not (30 to 17). Similar results are found for W(μ) (25 to 16). While distribution-response interaction may be present and might be detected with an analysis of variance it is the average effect which is important, since the analyst supposedly knows little about either the response or the distribution of the data prior to the analysis. If the conditions under which this experiment was run can be argued to be representative of what the analyst may actually encounter, then the results here indicate he can obtain equally accurate results

using the estimated or reconstructed population rather than the actual. This attests to the usefulness of the deltaspline reconstruction technique as a quantitative tool. The deltaspline approximation even appear to be able to smooth sample variations particularly in light of the 12.5% decrease in overall average MSD for $W(\mu)$. It remains to determine the effect of the deltaspline reconstruction technique on the estimates of μ^*, the optimal process states, before concluding its suitability for analysis purposes.

Table 5.5 summarizes the estimates of the optimal process states for each cell of the experiment design. It shows the minimum, maximum, and average estimates obtained when estimated values for $g(x)$ replaced the actual. The table can be compared to table 4.4 which shows the results from the same samples using the actual $g(x)$. The criterion based on $W(\mu)$ again is the superior estimator, substantiating all previous assertions. At this point it seems reasonable to conclude that the best optimizing criterion for an actual application of the analysis procedure developed in Chapter 2 then is to maximize $W(\mu)$ within a suitable region of interest using a deltaspline reconstruction to estimate the population represented by the collection of historical data.

Bimodal						Skewed				
W	I	I-1.0 S_I	I-1.5 S_I	I-2.0 S_I		W	I	I-1.0 S_I	I-1.5 S_I	I-2.0 S_I
					QUADRATIC					
2.5	2.6	2.4	2.4	2.3	min	2.5	1.2	2.6	2.6	2.6
2.5	3.2	3.1	3.1	3.0	max	2.5	2.8	3.0	3.0	3.1
2.50	2.85	2.81	2.79	2.74	ave	2.50	2.13	2.71	2.75	2.83
		$\mu^* = 2.50$						$\mu^* = 2.50$		
					LINEAR, DECREASING					
1.0	1.0	1.1	1.4	1.4	min	1.2	1.2	1.6	2.0	2.2
1.0	1.8	1.8	1.8	1.8	max	1.2	1.6	2.3	2.5	2.7
1.00	1.34	1.54	1.61	1.65	ave	1.20	1.31	1.91	2.15	2.33
		$\mu^* = 1.00$						$\mu^* = 1.20$		
					LINEAR, INCREASING					
3.7	3.3	3.2	3.1	2.9	min	3.3	1.2	3.0	3.0	3.0
3.7	3.7	3.7	3.4	3.1	max	3.3	3.3	3.2	3.1	3.1
3.70	3.61	3.45	3.18	3.05	ave	3.30	2.86	3.05	3.03	3.03
		$\mu^* = 3.70$						$\mu^* = 3.30$		

Region of interest: $1.0 \leq \mu \leq 3.7$ Region of interest: $1.2 \leq \mu \leq 3.3$

NOTE: μ^* is the μ for which $E(Y;\mu)$ is maximized within the appropriate region of interest.

Table 5.5 - Summary of optimal process state estimates - Phase 1 with estimated $g(x)$

6.0 Summary, Conclusions and Areas for Further Study

6.1 Summary

This thesis was undertaken with the intent to present and evaluate a technique for analyzing historical data such as that which might be taken from a manufacturing process. The results of the analysis are required to provide an estimate of the "optimal" process state, or at least a direction in which to shift the nominal process operating levels in order to realize an improvement in the process "level of performance." Because of the possibility that the historical data may not suggest a model which adequately describes its behavior, the analysis technique to be employed must not rely upon such a model which is to say the analysis must be performed "model free."

A review of available optimization analysis techniques revealed that none were exactly suited to the above requirements, necessitating the development of a new procedure. A technique to meet these requirements was thus developed utilizing concepts of the Monte Carlo variance reducing method of importance sampling. Optimization criteria based on either the usual importance sampling statistics or a statistic derived from another Monte Carlo method--weighted uniform sampling--were then proposed.

An evaluation of the new "model free" optimization technique with respect to these optimizing criteria followed. Since the new technique sacrifices the variance reducing power of importance sampling, it was necessary not only to determine which optimization criterion performs best, but also to obtain an understanding of the behaviors of the

statistics employed in those criteria. Simulated data for responses with one independent parameter were the basis for the evaluation and study.

A necessary and critical step in the analysis procedure which was developed is the construction of a population for which the historical data might be considered a "representative random sample." A technique based on the delaspline transform was adopted to perform this task. This construction technique, due to Boneva, Kendall, and Stefanov, was originally conceived to answer qualitative distributional questions so that a quantitative evaluation of the technique was necessary. This evaluation included a comparison of the results obtained in the first phase of the evaluation of the optimization technique with those which would have been obtained using the deltaspline construction techniques instead of an assumed knowledge of the exact underlying distribution of the data.

6.2 Conclusions

6.2.1 Discussion

The model-free empirical optimization technique developed in this thesis resembles, of necessity, the oft-used method of natural variation, one of the previously available empirical optimization techniques. This being the case, it is likely to be subjected to criticism by the statistically pure in heart. The resemblance is in the data source. Such a method relies on the day-to-day process variation to provide a "wide enough" range of values for each of the process parameters so that significant relations to the response variable will, by what-

ever means, be detectable. Thus the parameter which could provide the single greatest contribution to an improved level of performance might not be discovered if the values observed in the data for that parameter differ by only a small amount. Statistically pure or not, this analysis approach is the most widely used in a manufacturing environment because it can be made without disrupting the normal operations of the process. Other techniques would require the introduction of experimental lots into the normal product flow in order to obtain data for analysis.

This resemblance to the method of natural variation does not, however, restrict the application of the model-free concept. It was developed in this thesis to meet the worst case which might be encountered, i.e., no freedom to experiment. Should the production personnel be more receptive to experimental analysis, nothing in the model-free analysis procedure makes its application inappropriate. Thus it also serves as an alternative to the other available empirical optimization techniques of designed experiment and E.V.O.P. once the data is available.

Being an empirical optimization technique, the model-free approach seeks the true process optimal operating state in an iterative fashion. This is to say that no single set of data, no matter how collected, can be expected to reveal the most favorable operating conditions immediately. It can only be hoped that the results of an analysis will move closer to that state. Further collection of data and analysis of the same should eventually converge upon that state.

The experiment described in Chapter 3 and analyzed in Chapter 4 indicates decisively that, at least for the conditions considered, the optimizing criterion based on $W(\mu)$, the statistic derivable from weighted uniform sampling, is superior to those based on the usual importance sampling statistics. It is superior not only in its ability to predict the "optimal" process state, μ^*, but also in the resulting estimate of the expected process response, $E(Y;\mu^*)$. The results show conclusively that the model-free approach is a feasible approach to empirical process optimization, at least for a response which is a function of one process parameter.

The evaluation of the model-free technique in Chapters 3 and 4 was done independent of the method used for reconstructing the distribution of the data, $g(x)$. It was assumed that, somehow, the actual distribution underlying the data was exactly recoverable. The deltaspline transformation was then evaluated as a means for constructing $g(x)$ in an actual application of the analysis procedure. The results indicated that even though Boneva et al intended their technique to answer qualitative questions, their technique, with special consideration for an occasional negative frequency estimate, may also be applied quantitatively with reasonable success. In fact, when using estimates of the frequencies for $g(x)$ based on the deltaspline construction, the net effect on the statistic $W(\mu)$ was to remove some of the sample-to-sample variation so that it provided more accurate estimates of $E(Y;\mu)$ for all μ, and hence better estimates of μ^* than what was obtained when using the theoretical, or known frequencies.

While this technique was presented as a possible means of solving the problem in a critical step of the model-free analysis procedure, it is not intended to be conclusive. Other methods may exist which can equal or improve upon the performance of the deltaspline construction in identifying some population for which an arbitrary set of historical data could be considered a representative random sample. Its inclusion and evaluation in this thesis was necessary to establish the feasibility of the model-free analysis procedure, and this, the author is certain, it has done.

Perhaps the most important conclusions to be drawn from this thesis relate to generalization of the results to the case of n process parameters. First, it should be observed that nowhere in the development of the procedure is the word "independent" used in conjunction with "process parameters." This in itself is worthy of note. Thus the optimal control capability, $f(x;\mu)$, which the analyst must estimate prior to analysis of the data can reflect any known covariance between the process parameters and can also reflect a feed-forward control capability in the process.

Another point regarding $f(x;\mu)$, the optimal control capability should be made. The analysis with one process parameter which was done in this thesis assumed the shape of $f(x;\mu)$ to be fixed as was varied. However, this is not a requirement of the model-free analysis procedure. For example, if the analyst actually observed behavior in his data similar to that of the skewed distribution equation 3-10, he could reasonably conclude that the variance of his optimal control

capability should decrease with increasing μ, thus his specification of the optimal control capability would include a statement of how σ_f^2 varies with μ, or $\sigma_f^2(\mu)$. This freedom, wisely utilized, could affect and improve the resulting estimate of the optimal process state, μ^*.

A need for identifying a suitable region of interest in which to look for the optimal process state was established in Chapter 4, and a means of determining what it should be was suggested. Extension of this idea to n process parameters is not entirely clear. The simplest way to extend the idea would be to identify one dimensional regions as done in Chapter 4 and then use as the n-dimensional region of interest their cartesian product. It should be obvious that this is inadequate. The problem is compounded by the strange shapes which may result from an n-dimensional construction of g(x) using the deltaspline transformation on an arbitrary set of historical data. The criterion of Chapter 4 which was to consider only those μ's for for which less than 15% of $f(x;\mu)$ was to be found in the outer 1% of g(x) may be the best means for determining this region of interest, but its application with n parameters is certainly not straightforward.

Generalization of the deltaspline construction technique to an n-dimensional application is straightforward if independence of the process parameters can be a reasonable assumption since the one dimentional results can simply be multiplied together. But the historical data cannot be expected to exhibit such independence. The n-dimensional deltaspline construction is completely analagous to that of one, but instead of the single parameter, C, the resulting g(x)

will be determined by n parameters, C_1, C_2, ..., C_n, and proper smoothing of g(x) may be a serious problem. The deltaspline construction however may be the only feasible technique for obtaining g(x) since, as discussed in Chapter 5, other usual methods of constructing an underlying population are too restrictive in the resulting shapes.

The behaviors of the statistics $I(\mu)$, $W(\mu)$, and $S_I^2(\mu)$ when extended to n-dimensions are not expected to change. The optimizing criterion based on $W(\mu)$ will still most likely be superior. The greatest concern should be expressed with regard to the number of observations required for meaningful estimates of μ^*. While this required number actually depends, as shown in Chapter 4, upon both $f(x;\mu)$ and g(x), it can be expected to increase drastically with n. It then becomes essential to have an "adequacy test" such as that suggested in Chapter 4 which might be to count the relative maxima observed in $I(\mu)$ for the suitable region of interest.

6.2.2 Recommended Procedure

The discussion in the previous section indicates a concern with regard to generalization of the results presented in this thesis to the case of n process parameters. This concern can only be justified or put to rest by a natural extension of the work in this thesis.

However the results of this thesis do indicate that a model-free approach to process optimization is workable and that a reasonable procedure for this approach could be summarized as below:

1. Collect data, whether historical or experimental.

2. Construct the distribution of the data, $g(x)$, using the deltaspline approach.
3. Determine the optimal control capability $f(x;\mu)$.
4. Determine a suitable region of interest in which to look for the process optimum using the "15%, 1%" criterion or some other workable rule.
5. Determine the "adequacy" of the data by counting the number of relative maxima observed for $I(\mu)$ within the suitable region of interest. If necessary "relax" the optimal control and redetermine the suitable region of interest and adequacy of the data. Continue until data is considered "adequate" with respect to the smoothness of $I(\mu)$.
6. Estimate μ^*, the optimal process operating state by $\hat{\mu}^*$ where $\hat{\mu}^*$ maximizes $W(\mu)$ within the suitable region of interest.
7. Recommend to production personnel a shift in the nominal process operating levels to $\hat{\mu}^*$ and a respecification of tolerances derived from $f(x;\hat{\mu}^*)$.
8. Begin collecting data again so that this sequence may be repeated.

6.3 Areas for Further Study

Because this thesis was devoted to a new analysis technique, perhaps it has raised more questions than it has answered. Thus a great deal of work remains to be done in proving, or disproving its worth. The most urgent need for further work clearly relates to the generalization of the results to problems with more than one process parameter of interest. Because of the need for restricting the estimate of μ^* to a region where the statistics employed are meaningful, a workable rule for establishing a suitable region of interest for more than one parameter is needed. Also there are problems related to smoothing. How does the analyst know if his estimate of $g(x)$ is properly smoothed when working with four process parameters? Or how does he determine when $I(\mu)$ is "smooth enough" to conclude that the sample size is adequate, or that $f(x;\mu)$ has been "relaxed" enough?

If the multidimensional concerns for the model-free optimization technique can be allayed, then how does the approach compare in accuracy and efficiency as an alternative to the other available empirical optimization techniques, even if experimentation is permitted and models can be found?

The results of this thesis suggest a potentially powerful analysis technique whose worth needs to be further evaluated.

BIBLIOGRAPHY

1. Boneva, L. I., Kendall, D. G., and Stefanov, I., "Spline Transformations: Three New Diagnostic Aids for the Statistical Data-Analyst," Journal of the Royal Statistical Society, Series B, Vol. 33, No. 1, 1971, pp. 1-36.

2. Box, G. E. P., and Draper, N. R., Evolutionary Optimization, Wiley, 1969.

3. Box, G. E. P., and Hunter, J. S., "Condensed Calculations for E.V.O.P. Programs," Technometrics, Vol. 1, No. 1, February, 1959.

4. Clark, C. E., "Importance Sampling in Monte Carlo Analysis," Technical Memorandum No. TM-505, System Development Corporation, Santa Monica, California, June 24, 1960.

5. Daniel, C., and Wood, F. S., Fitting Equations to Data, Wiley-Interscience, 1971.

6. Davies, O. L., The Design and Analysis of Industrial Experiments, Hafner, Second Edition, 1965, pp. 495-578,

7. Evans, D. H., "Applied Multiplex Sampling," Technometrics, Vol. 5, No. 3, August, 1963, pp. 341-359.

8. Evans, D. H., "Multiplex Sampling," Annals of Mathematical Statistics, Vol. 34, No. 4, 1963.

9. Evans, D. H., "Tolerancing with Given Data," (unpublished).

10. Goll, E. O., "An Investigation of Empirical Optimization Techniques for Manufacturing Processes," unpublished M.S. Thesis, Lehigh University, 1971.

11. Halton, J. H., "A Retrospective and Prospective Survey of the Monte Carlo Method," SIAM Review, Vol. 12, No. 1, January, 1970, pp. 1-63.

12. Hanmersley, J. M., and Handscomb, D. C., Monte Carlo Methods, Methuen, London, 1964.

13. Handscomb, D. C., "Remarks on a Monte Carlo Integration Method," Numerische Mathematik, Vol. 6, 1964, pp. 261-268.

14. Hicks, C. R., Fundamental Concepts in the Design of Experiments, Holt, Rinehart and Winston, 1964.

15. Kahn, H., "Use of Different Monte Carlo Sampling Techniques," Symposium on Monte Carlo Methods, H. A. Meyer, Ed., Wiley, 1956, pp. 146-190.

16. Pennel, A. R., "Practical Applications of Statistical Design of Experiments," Journal of Industrial Engineering, October, 1968.

17. Powell, M. J. D., and Swan, J., "Weighted Unifrom Sampling - A Monte Carlo Technique for Reducing Variance," Journal of the Institute of Mathematical Applications, Vol. 2, 1966, pp. 228-236.

18. Shreider, Y. A., The Monte Carlo Method, Pergamon Press, 1967.

19. Wilde, D. J. Optimal Seeking Methods, Prentice Hall, 1964.

VITA

Personal History

Name: Thomas C. Pingel

Date of Birth: May 20, 1947

Place of Birth: Mansfield, Ohio

Parents: Marvin E. and Ruth E. Pingel

Educational Background

Upper Arlington Senior High School — Graduated 1965
Columbus, Ohio

Purdue University — 1965 - 1969
West Lafayette, Indiana

Bachelor of Science, "With Highest Distinction"
major: Mathematics
minor: Physics

Lehigh University — 1971 - 1973
Bethlehem, Pennsylvania

Candidate for Master of Science in
Industrial Engineering

Honors

Delta Rho Kappa, Honor Society
Sigma Pi Sigma, Physics Honor Society
Alpha Pi Mu, Industrial Engineering Honor Society

Professional Experience

Battelle Memorial Institute — 1967
Columbus, Ohio
Technical Assistant

Western Electric Company, Inc. — 1969 - 1971
Columbus, Ohio
Information Systems Staff Member

Western Electric Company, Inc. — 1971 - 1973
Princeton, New Jersey
Information Systems Staff Member