ANALYTIC AND SIMULATIVE APPROACHES TO AN ELECTRONIC REFERENDA--SYSTEM OCCUPANCY PROBLEM

Peter Parkinson

Bell Northern Research

error in sim

anal. sol. is not correct

1. INTRODUCTION

Consider a referenda system in which questions are broadcast to domestic television and radio receivers and a measure of audience reaction returned to the broadcast distribution centre electronically. Such a system might be used to gauge the popularity of programs, commercial products, sociopolitical opinions and new ideas. The return paths might form an independent and dedicated physical network or be carried over modified existing facilities such as CATV distribution and telephony networks.

In such a system it might be unnecessary to identify each individual voter or to attempt to return highly accurate estimates of the size of the responding audiences. In the United States House of Representatives, one of the five traditional voting methods has been the 'voice vote' in which the presiding officer makes an auditory comparison of the relative volumes of YEA and NAY responses. "Only if the volumes appear similar is another vote method invoked. (1) Clearly in this case, where a decision is sufficiently popular, the accuracy of the vote is not important.

In a domestic referenda system which could be used repeatedly at low cost, a polling 'uncertainty' of ±10% might be quite acceptable for most purposes.

By way of comparison, if the percentage of spoiled ballot slips is used as a crude measure of the uncertainty in a typical federal election, (2) these may be looked upon as having an accuracy of ±1% and costing in Canada (federal expenditure only) roughly $2 per voter in 1972. (3)

2. NETWORK CONFIGURATION

Figure 2.1 shows a simplified schematic of a vote response gathering network in which a large number of reply keys are shown generating signals in a narrowly defined time interval, in answer to a broadcast question.

Primary response circuits are shown terminated at concentrators C1, C2, C3, etc., from which secondary response signals are routed to a hierarchy of totalizers. The purpose of the concentrators is to reduce the amount of dedicated circuitry in the network. Each concentrator is an OR gate, and generates a secondary response signal if any of the incoming primary circuits is found to be carrying a poll response signal.

The schematic of Figure 2.1 shows a possible network having only one choice of reply key per broadcast receiver. This would permit polls to be conducted on a 'show-of-hands' basis in which the YEAS cast their vote during a specific time interval and the NAYS immediately after. A multi-choice key system can be represented by separate networks, one for each type of key (see Figure 2.2). In practice such a system might be implemented by multiplexing key signals within a single physical network.

Common to both schematics is the technique of dedicated circuit reduction by the use of OR-gates. An analysis of the statistics of such networks leads to a statement of the occupancy problem that is the subject of this paper.

3. THE OCCUPANCY PROBLEM

Let the following be defined:

Number of OR-gates............................N
Primary circuits per gate.....................n
Number of reply keys..........................K

Average *a priori* probability of an affirmative

reply on any of the N.n circuits..............α
Output from totalizer..........................y
Number of affirmative responses...............x

Figure 3.1 is a schematic diagram for a specific case in which N=4, n=3, K=9, x=6 and y=3.

The general problem can be stated as given N, n, K and y, what is E(x), the expected value of x, and with what measure of confidence can it be assumed?

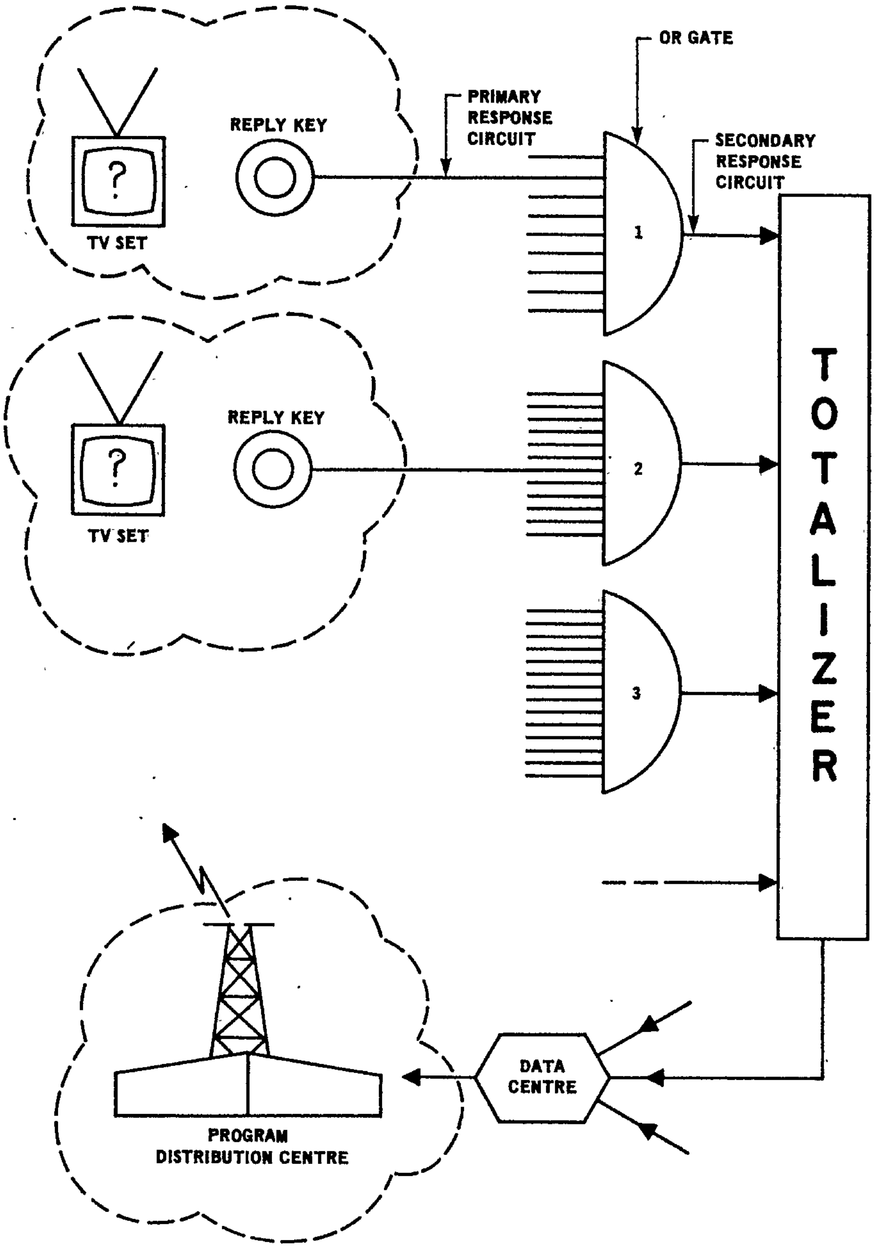

Figure 2.1

VOTE RESPONSE GATHERING NETWORK
SINGLE KEY CHOICE

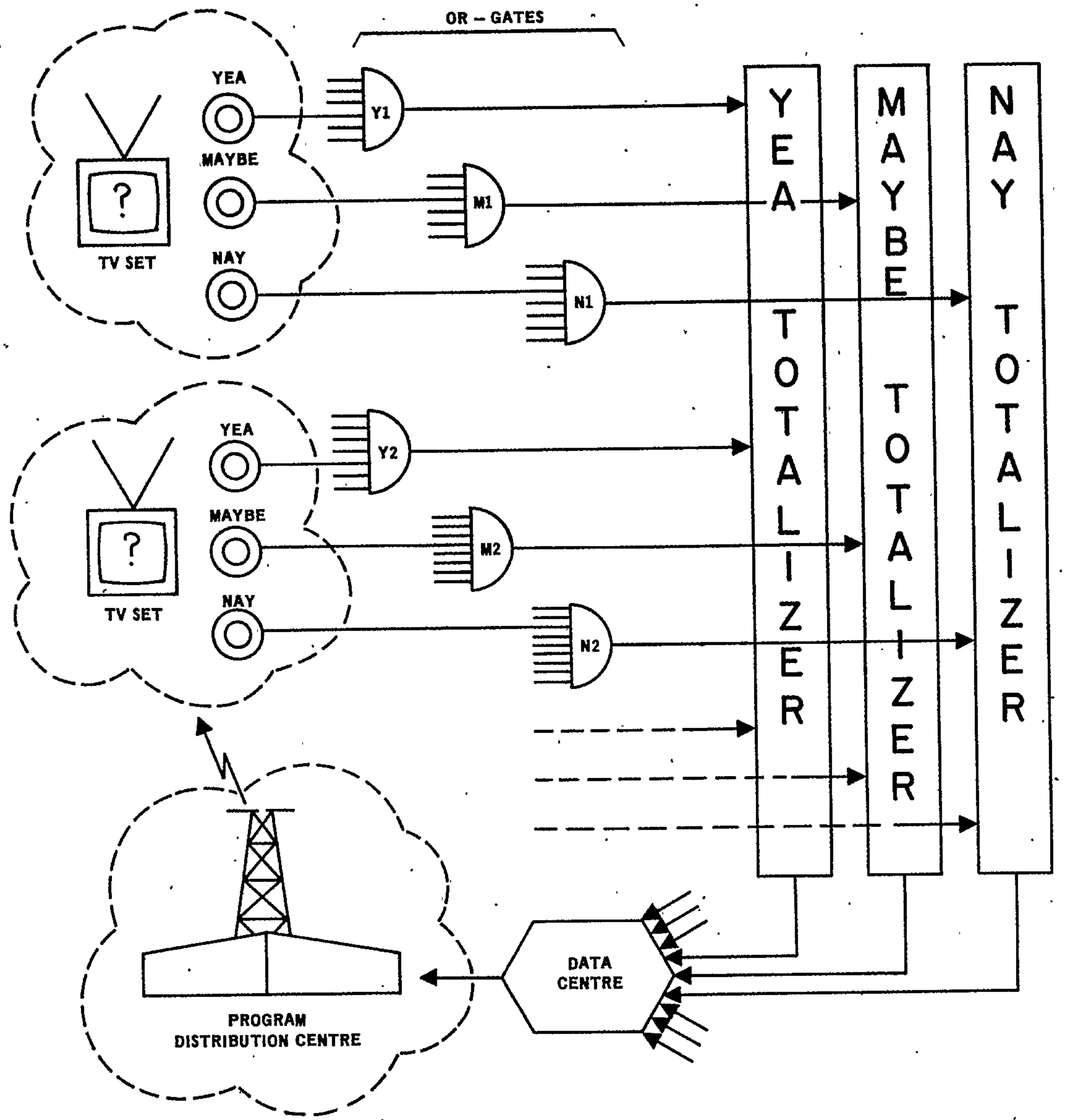

Figure 2.2
VOTE RESPONSE GATHERING NETWORK
MULTIPLE KEY CHOICE

ANALYTIC APPROACH

y = number of concentrators (OR-gates) receiving at least one affirmative reply.

= N. (probability of an OR-gate receiving a reply)

= N. (1 - probability of zero replies in gate)

i.e., $y = N.\ (1 - (1 - \alpha)^n)$.

The above equation can be solved to give in terms of y, N and n.

$$\alpha = 1 - \sqrt[n]{1 - y/N}\ .$$

The expected value of x is

$$E(x) = Nn\ \alpha.$$

Substituting for α in equation (A.1) gives

$$E(x) = Nn\ (1 - \sqrt[n]{1 - y/N}) \qquad (1)$$

A measure of the confidence to be placed on any particular value of E(x) may be derived by examining the variance $\sigma^2(x)$. This is the mean square deviation of x from E(x), and may be calculated, for a given value of y, as follows:

$$\sigma^2(x) = \frac{1}{M} \sum_{x=0}^{x=Nn} f(x).\ \ (x-E(x))^2.$$

Where f(x) = number of configurations having exactly x inputs and y outputs

$$\text{with}\quad M = \sum_{x=0}^{x=Nn} f(x)$$

$$\text{and}\quad f(x) = \binom{Nn}{n}\ \alpha^x\ .\ (1 - \alpha)^{Nn-x}.$$

The above three equations may be solved to give

$$\sigma(x) = \sqrt{Nn\ \alpha(1 - \alpha)} \qquad (2)$$

It will be noted that the above equations for E(x) and σ(x) are independent of both K, the total number of keys distributed and of detailed references to particular distributions for any given K.

The lack of particular distribution reference can be traced to the definition of α. The solutions assume that the keys are randomly assigned before each poll is taken. The effect of this untrue assumption is to ignore any biases that may be present in real estimations of x in particular keyed networks. Such biases may, however, be kept small by ensuring as near as possible a uniform distribution of keyed-circuits amongst the OR-gates.

The fact that the above equations for E(x) and σ(x) are independent of K requires an explanation. It would appear reasonable to expect that no estimate for x should exceed K and yet equation (1) permits the following:

Let y = N

Then by equation (1):

$$E(x) = Nn(1 - \sqrt[n]{1 - 1})$$

i.e. E(x) = Nn

In most situations this estimate for E(x) can exceed K, whose range is from N to Nn inclusive, given y = N.

From the above argument the analysis solutions appear degenerate, at least when y = N. This raises the question of the extent and effect of the degeneracy. To gain insight into this area it was decided to supplement analysis with simulation.

SIMULATIVE APPROACH

This approach uses a scatter diagram constructed from the results of repeated poll simulations. A computer program invoking a random number sub-routine is used to generate a typical poll result, y, when given a number, x, to represent the affirmative replies (keys depressed). The resulting coordinates (x, y) are used to plot a point on a two dimensional graph. The process is repeated many times until the graph becomes a scatter diagram indicating a probabilistic relationship between x and y.

Such a scatter diagram is shown in Figure 3.2. This is a record of 10,000 separate simulations. The parameters chosen to define the simulated network were as follows:

OR-gates..............................N = 500
Number of circuits per OR-gate............n = 20
Total circuits.....................N x n = 10,000
Number of reply keys..................K = 2,500

Examination of the program used to generate the scatter diagram reveals that the effect of changing K from 2,500 to say 3,500 would be simply to generate more pairs of points. All of the points generated for the original scatter diagram could be used again in the new one. This is illustrated in Figure 3.3 where a scatter diagram is shown extended to cover the effect of increasing K. From the general shape of this diagram with its sharply truncated limit at x = K it can be seen that it will be inappropriate to apply Gaussian statistics to variations in x for a given y beyond say y = y' (for the small scatter diagram).

The exact definition of y' is obviously debatable since the scatter diagram has no absolutely clearly defined 'edge'.

4. THE ANALYSIS RE-EXAMINED

In Figure 4.1 the scatter diagram of Figure 3.2 is repeated overlaid with the analytic curve for E(x)

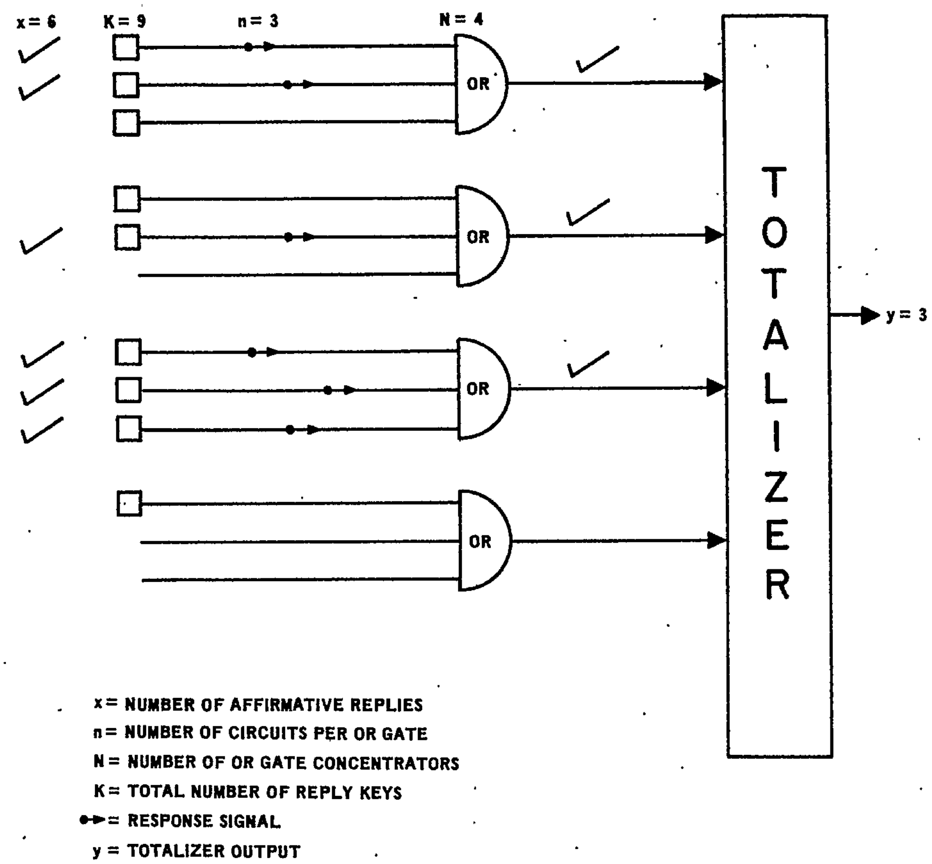

Figure 3.1

OCCUPANCY PROBLEM PARAMETERS
(numerical example)

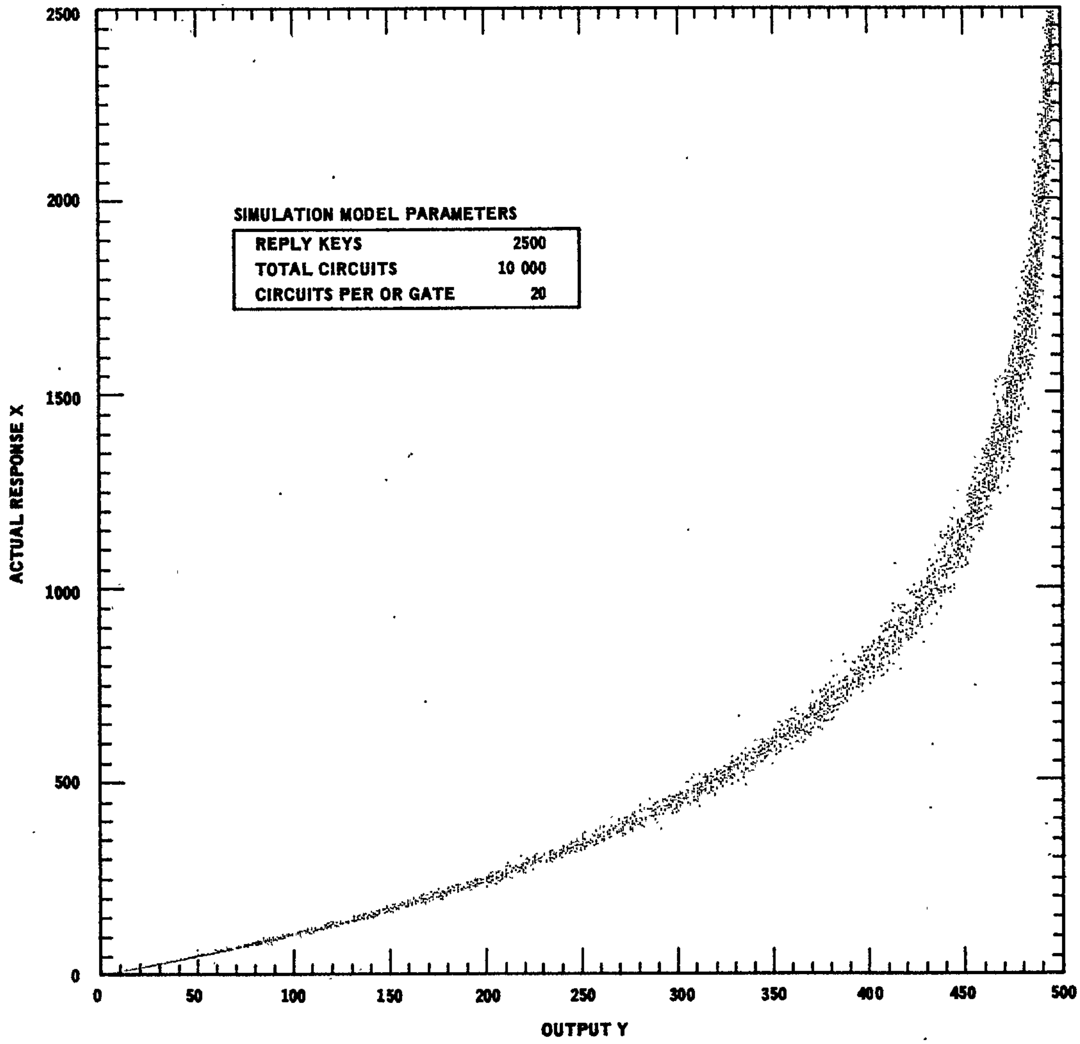

Figure 3.2

SCATTER DIAGRAM (10 000 points)

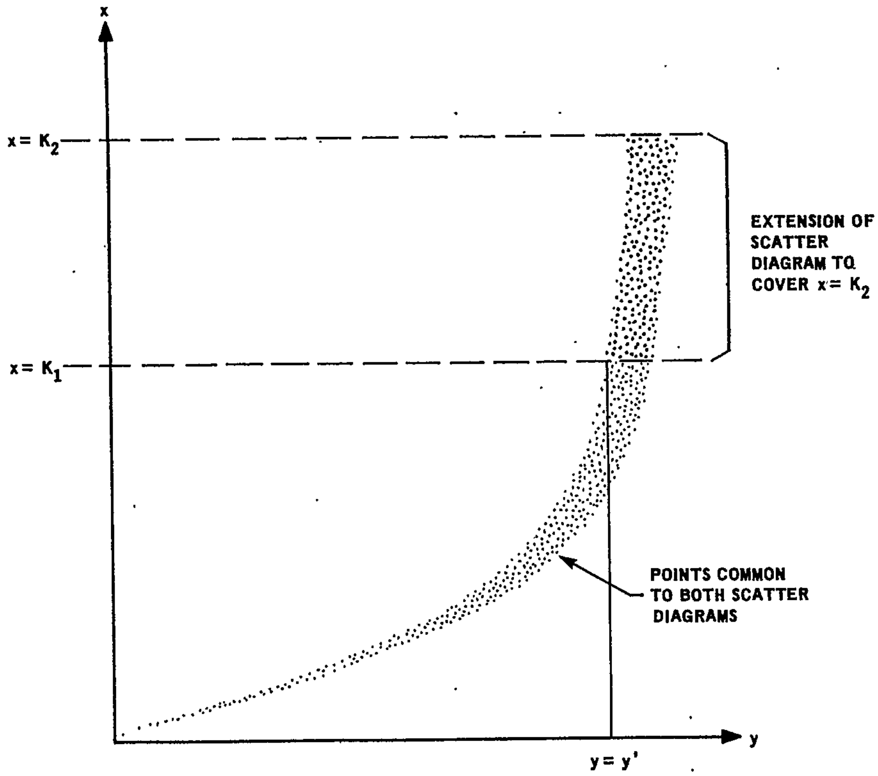

Figure 3.3

SCATTER DIAGRAMS FOR TWO DIFFERENT VALUES OF K derived from equation (1). Note the good agreement between analysis and simulation for low values of y. (i.e. for $0 < y < 70\%N$).

Examination of Figure 3.2 reveals that the accuracy of the estimation appears to vary with the size of the poll.

In Figure 4-2, three curves have been fitted to the original scatter diagram. The upper and lower curves were drawn manually using curve templates so as to exclude 26 of the more extreme points. Approximately half of the excluded points lie below the lower bound and half lie above the upper bound. The curves can therefore be said to represent the 99.74 per cent confidence limits of deviation from a mean curve. The mean curve is established half way between these limiting curves.

If the probability distribution of x given y were known to be Gaussian, it would be permissible to equate its 99.74 per cent extremes with the 3σ deviation limits. Although $p(x/y)$ is not necessarily Gaussian for all or any given values of y, it is apparent from the scatter diagram that the distributions are for the most part two-tailed and that they do possess a central maximum likelihood area.

With the above as justification, the standard deviation σ was determined using the scatter diagram. This yielded the data shown in Table 4-1 in which a comparison is made between the simulation results and the ranges for x anticipated by the original analysis.

5. CONCLUSIONS

Table 4.1 shows good agreement between analysis and simulations for the range $0 < y < N/2$. In the example used, this corresponds to a mean response of approximately $E(x) = 345$ out of a total possible response of 2500; i.e. a 14% response.

Since the uncertainty in estimating x increases with x, the worst case in determining the outcome of a close vote could occur given a 50% AYE response and a 50% NAYE response. Either one of these poll returns would yield $x = 1250$, $y = 450$ (approx.). A comparison of the 'expected value of x' columns for $y = 450$ in Table 4.1 shows analysis and simulation still in agreement to within 6% and the calculated one-sigma deviation in error by a margin equal to 12% of the simulated $E(x)$.

Although the drawing of absolute conclusions as to the strength of the analysis would be premature, the simulation program has offered qualified support and given a better understanding of the nature of the original problem.

REFERENCES

1. Ryan, Frank B. The Electronic Voting System for the United States House of Representatives. Computer: November/December, 1972.

2. Report of Chief Electoral Officer, 1968. (Information Canada Bookstore reference SE1-1/1968).

3. Spokesman for Chief Electoral Officer - Government of Canada. April 19, 1973.

4. Brendel, P.J. Incasting and the Telephone Network. International Conference on Communications, Seattle, June 1973.

5. Jurgen, R.K. Two-way Applications for Cable Television Systems in the '70's. IEEE Spectrum, November 1971.

6. Unger, S.H. Technology to Facilitate Citizen Participation in Government. Working Paper II, Center for Policy Research, Washington, D.C.

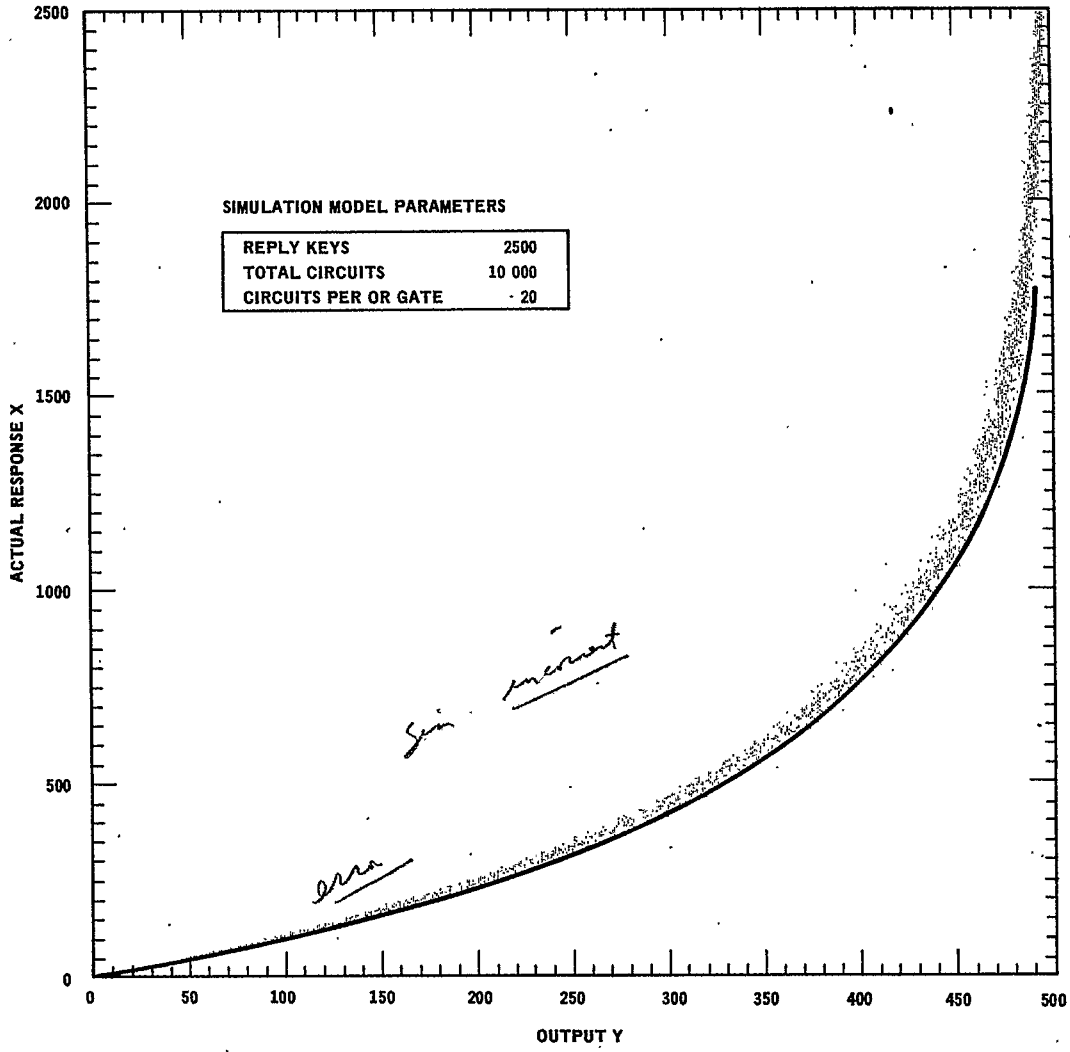

Figure 4.1
SCATTER DIAGRAM (10 000 points) OVERLAYED WITH
ANALYTIC CURVE FOR E(x)

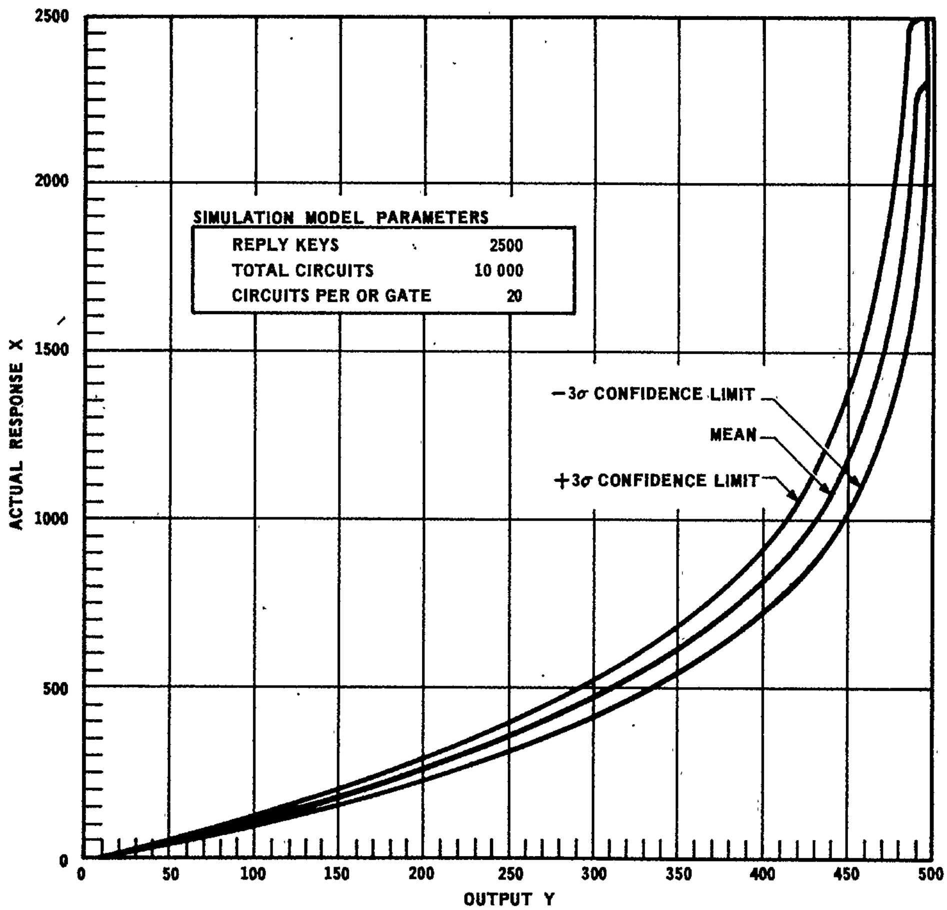

Figure 4.2

99.74% (3σ) Confidence Limits
& MEAN (most likely) RESPONSE

ACKNOWLEDGEMENTS

The author wishes to thank Marius Lewin, now with the Manitoba Provincial Government, and Professor Miklos Csorgo of the Department of Mathematics, Carleton University for their contribution to the theoretical analysis of the data interpretation problem. Especial thanks also to Bruce Sanderson of Bell Northern Research for his patient work in production stream-lining the original simulation computer programs.

TOTALIZER OUTPUT y	EXPECTED VALUE OF X		ONE SIGMA DEVIATION		ONE SIGMA LIMITS		THREE SIGMA LIMITS	
	FROM SCATTER DIAGRAM	CALCULATED FROM EQN #1	FROM SCATTER DIAGRAM	CALCULATED FROM EQN #2	FROM SCATTER DIAGRAM	CALCULATED	FROM SCATTER DIAGRAM	CALCULATED
50	50	53	3.3	7.2	53.3 / 46.7	60.2 / 45.8	59.9 / 40.1	74.6 / 31.4
100	110	111	5.5	10.4	115.5 / 104.5	121.4 / 100.6	126.5 / 93.5	142.2 / 79.8
150	180	177	8.0	13.1	118 / 172	190.1 / 153.9	204 / 156	216.3 / 127.7
200	256	252	10.5	15.6	266.5 / 245.5	267.6 / 236.4	287.5 / 224.5	298.8 / 205.2
250	350	341	14.0	18.1	364 / 336	359.1 / 322.9	392 / 308	395.3 / 286.7
300	465	448	17.5	20.6	482.5 / 427.5	468.6 / 427.4	517.5 / 392.5	509.8 / 386.2
350	610	584	23.3	23.4	633.3 / 586.7	607.4 / 560.6	679.9 / 540.1	654.2 / 513.8
400	805	773	31.7	26.7	836.7 / 773.3	799.7 / 746.3	900.1 / 709.9	853.1 / 692.9
430	980	936	41.7	29.1	1021.7 / 938.3	965.1 / 906.9	1105.1 / 854.9	1023.3 / 848.7
450	1150	1087	55.0	31.1	1205 / 1095	1118.1 / 1055.9	1315 / 1205	1180.3 / 993.7
480	1665	1487	106.0	35.5	1771 / 1559	1522.5 / 1451.5	1983 / 1347	1593.5 / 1380.5

Table 4.1

COMPARISON OF SIMULATIVE & ANALYTIC ESTIMATING