APPROXIMATING THE POINT BINOMIAL WITH THE

GRAM-CHARLIER TYPE B SERIES

by

David Aaker

David Butterfield

S.B., School of Industrial Management

1960

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

1960

Signature of authors . . . . Signature redacted
David Aaker

. . . . Signature redacted
David Butterfield

Certified by . . . . Signature redacted
Faculty Advisor of the Thesis

May 20, 1960

Professor Philip Franklin,
Secretary of the Faculty
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

Dear Professor Franklin:

In accordance with the requirements for graduation, we herewith submit a thesis entitled "Approximating the Point Binomial With the Gram-Charlier Type B Series".

In addition, at this time we would also like to express our gratitude to Professor David Durand of the School of Industrial Management for his assistance and patience as our Thesis Advisor; also we would like to take this opportunity to thank the School of Industrial Management for making this joint thesis possible.

Sincerely,

Signature redacted

David Aaker

Signature redacted

David Butterfield

ABSTRACT

APPROXIMATING THE POINT BINOMIAL WITH THE
GRAM-CHARLIER TYPE B SERIES

David Aaker
David Butterfield

Submitted to the School of Industrial Management on May 20, 1960 in partial fulfillment of the requirements for the degree of Bachelor of Science.

The Poisson distribution is one of the well-known approximations to the Binomial distribution. An improvement to the Poisson distribution, used only occasionally, is the Type B Gram-Charlier series, which consists of the Poisson and its differences. A two term approximation (through the second difference of the Poisson) has been previously tried and found useful. We have worked with the Type B Gram-Charlier series to obtain additional information about this two term approximation, and also to find what possible improvement in accuracy could be obtained by the use of two additional terms.

We found that the use of the second-difference term gives an improvement of a factor of ten over the simple Poisson. The third-difference term does not improve the approximation, but the fourth-difference term adds another factor of ten to its accuracy. These factor improvements are given as an indication of the type of results obtained; the actual improvements varied considerably with the value of p, the Binomial probability. The results indicated that this approximation is very good for p less than 0.3. The fact that the accuracy of the approximation varies with p is generally known. However, contrary to popular belief, the accuracy does not vary substantially with n, the sample size.

The improvement is significant enough to suggest that the second difference should be printed along with the cumulative and first difference in tables of the Poisson. Other characteristics of the Poisson distribution made the use of the differences also valuable in interpolation. We have investigated interpolation for its own sake and also for its use with the Type B Gram-Charlier series.

Thesis Advisor: David Durand
Title: Professor of Industrial Management

TABLE OF CONTENTS

TABLE OF TABLES AND EXHIBITS

Page

CHAPTER I

APPROXIMATION

Essentially, the Type B Gram-Charlier series is a device for approximating discrete distributions resembling the Poisson at least slightly. The series appears as follows:

$$f(x) = C_0P(x) + C_1P'(x) + C_2P''(x) + C_3P'''(x) + C_4P''''(x) + \ldots$$

where $P(x) = \frac{e^{-m}m^x}{x!}$, $P'(x) = P(x) - P(x-1)$ ---- the first difference, and x is an integer. If the series is summed for $x = 0$ to $x = r$, the result is:

$$F(r) = \sum_{i=0}^{r} f(x) = C_0 \sum_{0}^{x} P(x) + C_1P(x) + C_2P'(x) + C_3P''(x) + \ldots .$$

Thus one set of coefficients - C_0 , C_1 , etc. - provide the means for approximating either the individual or the cumulative probabilities.

Theory tells us that this series will converge if enough terms are included. However, in practice a finite number of terms must be used so that an error can be expected to occur as the series is stopped at a certain term. We have investigated four terms of the series in this thesis. The first term is the well-known simple Poisson and had been used as one of the spproximations to the Binomial for years. The term in C_2 (C_1 is zero) has been tried by Raff [1] and was found to be an useful improvement to the first term. We have obtained additional information about the term in C_2 and have also investigated the previously untried terms in C_3 and C_4 .

Our first problem was to find the relevent coefficients of the Type B series. We were greatly aided in this task by previous work done by Aroian [2] and Kendall [3]. Aroian obtained the coefficients in a generalized form in terms of moments of the desired distribution. When the Type B Gram-Charlier is to be used to approximate the Binomial, the moments around the mean of the Binomial are the correct moments to use in Aroian's coefficients. Kendall reproduces these moments in terms of the parameters of the Binomial distribution. We used Kendall's moments in Aroian's generalized coefficients and reduced the subsequent results until they were in their most usable form. The following are the coefficients which we obtained:

$$C_0 = 1$$

$$C_1 = 0$$

$$C_2 = -\frac{1}{2}\,np^2$$

$$C_3 = -\frac{1}{3}\,np^3$$

$$C_4 = +\frac{1}{8}\,np^4(n-2)$$

Thus, rewriting the Type B Gram-Charlier series using these coefficients we have:

$$B(x) = P(x) - \frac{1}{2}np^2P''(x) - \frac{1}{3}np^3P'''(x) + \frac{1}{8}np^4(n-2)P''''(x) + \ldots$$

The first two coefficients, C_0 and C_1, suffice to equate the mean of the Binomial to the mean of the Type B series. This makes the first term of the series the simple Poisson approximation. For example, in a sample of 100 with 5% defective, the mean of the

Binomial, np=5 , is used as the mean of the Poisson, m , for the approximation. However, the difference of the variances of the Poisson and the Binomial, np vs. npq , introduces a substantial error in the simple Poisson approximation. The term in C_2 removes this error by equating the two variances. Similarly, the function of the terms in C_3 and C_4 is to equate the third and fourth moments of the Binomial with the third and fourth moments of the Type B series.

CHAPTER II

ANALYSIS OF DATA

Unlike some mathematical series, no means of determining the error of the Gram-Charlier series analytically is now known. It was therefore necessary to calculate values of the series in order to obtain the error. This was an arduous task which was simplified greatly by the availability of a table of the cumulative Poisson with differences, a sample page of which is included as Exhibit I of the Appendix. This table is the result of a program which was run on the the MIT Computation Center's IBM 704 computer. The print out, which is accurate to seven decimal places , was obtained for the following values of $m = np$:

0.00(0.02)0.20(0.05)1.40(0.10)6.0(0.20)10.0(0.50)20.0

The error referred to hereafter as the maximum error consists of the maximum value of the difference between the Gram-Charlier series for a given number of terms and a given n and p , and the cumulative Binomial. This definition of maximum error obtained is different from that of Raff, who defines his error as "...the largest possible error which can arise in estimating any consecutive number of binomial terms with the specified parameters." [4] Raff's error, therefore, is approximately twice as large as ours.

TABLE OF THE ABSOLUTE VALUE OF THE

MAXIMUM ERRORS

$\times 10^{-5}$

P	N	M	1st*	2nd**	3rd	4th	4th Adj
.02	25	0.50	307	4	3	1	1
	50	1.00	371	3	8	4	4
	100	2.00	274	4	3	1	2
	200	4.00	295	3	4	2	1
	500	10.00	273	3	4	2	2
.03	25	0.75	539	78	66	4	4
.05	20	1.00	939	22	23	2	1
.08	5	0.40	1124	75	17	4	5
	25	2.00	1129	57	46	6	4
	50	4.00	1213	49	56	4	3
	100	8.00	1136	49	51	3	2
	200	16.00	1097	49	52	13	9
.10	50	5.00	945	68	71	8	7
.15	20	3.00	2359	198	205	35	25
.16	5	0.80	3112	236	171	49	32
	25	4.00	2514	212	237	36	29
	50	8.00	2351	215	203	29	24
	100	16.00	2280	208	216	30	25
.20	5	1.00	4020	399	338	117	62
	20	4.00	3202	345	379	71	58
	25	5.00	3104	336	367	62	43
.24	5	1.20	4764	644	560	223	134
	25	6.00	3665	515	521	109	97
	50	12.00	3584	498	527	103	93
.25	20	5.00	3984	545	591	125	113
.30	20	6.00	4755	845	843	221	196
.32	25	8.00	5055	997	967	276	235
.40	5	2.00	6905	1975	1491	841	652
	25	10.00	6667	1675	1665	574	513
	45	18.00	6504	1330	1469	336	283
	50	20.00	6498	1459	1621	512	481

* P(x) - B(x)

** P(x) with first correction factor - B(x)

It can be seen from the preceding table that for a given n and p the term in C_2 gives a ten fold improvement over the simple Poisson. Raff mentioned this in his article and presented data to support his observation. Wehave developed additional data in support of his observation, and, in addition, have shown that the addition of the term in C_3 effects no improvement in the approximation, whereas when the terms in C_3 and C_4 are applied a ten fold increase in accuracy over the term in C_2 is realized. The affect that the terms in C_2 and C_3 and C_4 have upon the maximum error is presented in Exhibits II and III of the Appendix.

There appears to be a rather general belief that the error in the Poisson decreases as n grows larger. We are providing information that supports Raff in his contention that the error is independent of n. We found that n had no significant effect upon the size of the resulting error. The maximum error obtained for a given p is presented in table form on the preceding page. It can be observed from this table that the error independency of n applies to the second, third, and fourth terms beyond the simple Poisson as well as to the simple Poisson itself.

It is quite well known that the error of approximation is dependent upon p - more specifically, the error decreases as p grows small. It is not generally known, however, how the error behaves as a function of p. Our data supports the error dependency upon p, and, in addition, gives an indication of how the error varies with increasing p. The curve of maximum error as a function of p is included as Exhibit 4 of the Appendix.

CHAPTER III

INTERPOLATION

The Poisson distribution has a characteristic which makes its differences in the x direction useful when interpolating in the m direction. The even derivatives with respect to m (second, fourth, etc.) of the Poisson expression, $\frac{e^{-m} m^{x}}{x!}$, are identical to the even differences in the x direction. The corresponding odd derivatives are identical to the negatives of the odd differences. This suggests a method of interpolation which makes use of the derivatives (differences). Taylor's series provides such a method. Specifically the Poisson derivatives in m in Taylor's series can be expressed as the Poisson differences in x----

$$P(x:m') = P(x:m) - hP'(x:m) + \tfrac{1}{2}h^{2}P''(x:m) - \frac{h^{3}P'''(x:m)}{6}$$

$$\frac{h^{4}P''''(x:m)}{24} + \ldots.$$

where $P'(x:m)$ is the first difference in x of the Poisson distribution. ($P'(x:m) = P'(x+1:m) - P'(x:m)$)

This provides an extremely useful method for interpolating in the Poisson distribution. None of the terms are complicated and the first two are exceptionally easily obtained. Also, the first difference is always available when interpolating the cumulative Poisson. There is also the possibility of the future availability of the second difference as this thesis has noted its usefulness when the Poisson is used as an approximation.

For any stated number of successive terms in the Taylor's series, the accuracy of the interpolation depends upon the grid, or the value of h, and the size of m. The error will be greater as h is smaller and m is larger. Some specific examples can be given to show the kind of accuracy to expect. If h is .02 and m is .10 the third-difference term essentially will not affect the fifth place. If h is .20 and m is 6.0 the third-difference term is insignificant in the fourth place. If h is .5 and m is 10.0 the fourth-difference term is insignificant in the fourth place.

We now see that Poisson difference tables are useful for interpolation as well as approximation. This suggests that the two operations could be consolidated into one calculation. This is indeed the case. Taylor's can easily be incorporated into the approximation formulas developed earlier. The question arises, however, how far should we carry Taylor's series in the interpolation of the terms of the approximation? It must be remembered that the successive differences of the Poisson must be interpolated just as the original term is. At the outset we take a pessimistic view of the interpolation and carry it to include the term with the difference that is present with the particular approximation we are using. This will insure that the maximum errors of the approximation will not be affected by interpolation if any reasonable grid is used. Applying this criterian to the approximation formulas developed earlier we have:

"GOOD ACCURACY"

$$B(r:n,p) = P(r:m) - hP'(r:m) + (\tfrac{1}{2}h^2 - \tfrac{1}{2}np^2)P''(r:m)$$

"BEST ACCURACY"

$$B(r:n,p) = P(r:m) - hP'(r:m) + (\tfrac{1}{2}h^2 - \tfrac{1}{2}np^2)P''(r:m) +$$

$$(-\frac{h^3}{6} + \frac{np^2h}{2} - \frac{np^3}{3})P'''(r:m) +$$

$$(\frac{h^4}{24} - \frac{np^2h^2}{4} + \frac{np^3h}{3} + \frac{n^2p^4}{8})P''''(r:m)$$

The reader will have to use judgment when applying these formulas. For a large number of applications, when h is small, the terms in h^3 and h^4 can be neglected. There may even be occasion to eliminate the terms in h^2. The individual situation must govern the use of these formulas.

CHAPTER IV

CONCLUSION

The Type B Gram-Charlier series can be an invaluable aid to one who uses the Binomial distribution. The use of one term beyond the single Poisson approximation is not difficult and adds roughly one decimal place to the accuracy of the approximation. The inclusion of two additional terms increases the accuracy roughly another decimal place. This improvement is significant enough to suggest that the second difference should be printed along with the cumulative and first difference in tables of the Poisson. The extra space this would require would not be great enough to offset the advantages to the user of the Poisson as an approximation to the Binomial. The second difference also proves to be valuable to anyone interpolating in the Poisson distribution. Its inclusion in tables of the Poisson would be worthwhile even if its use were limited to interpolation. If the situation demands, the approximation and interpolation can easily be consolidated into one operation. Thus the differences of the Poisson are very useful for both the Type B series for approximation and the Taylor's series for interpolation.

The reader should remember the affect of p and n upon the approximation. The accuracy becomes markedly worse as p gets larger but is relatively independent of n. Like the simple Poisson approximation, the Type B series is designed to be used with small p. However, the use of additional terms expands the usefulness of the Type B to a p of about 0.3.

FOOTNOTES

[1] Raff, Morton, "Approximating the Point Binomial," Journal of the American Statistical Association, June 1956, Vol. 51, No. 274, pp. 293-303.

[2] Aroian, Leo A., "The Type B Gram-Charlier Series," Annals of Mathematical Statistics, Vol. VIII, 1937, pp. 183-192.

[3] Kendall, Maurice K, The Advanced Theory of Statistics, Vol. I, 4th edition, Charles Griffin and Co., Ltd., London, 1948, p. 147.

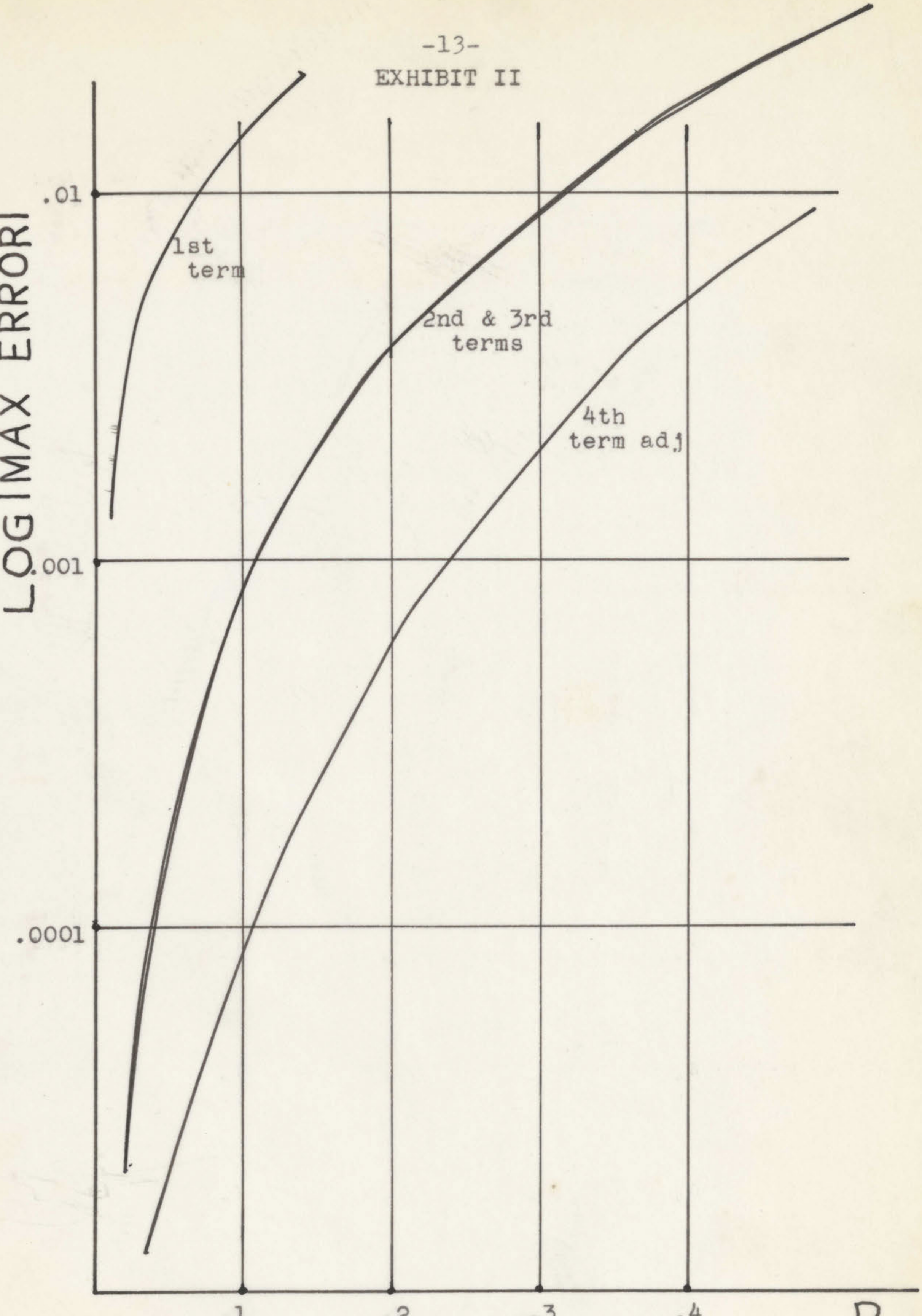
[4] *Ibid.*, p. 298.

EXHIBIT I

M= 3.60

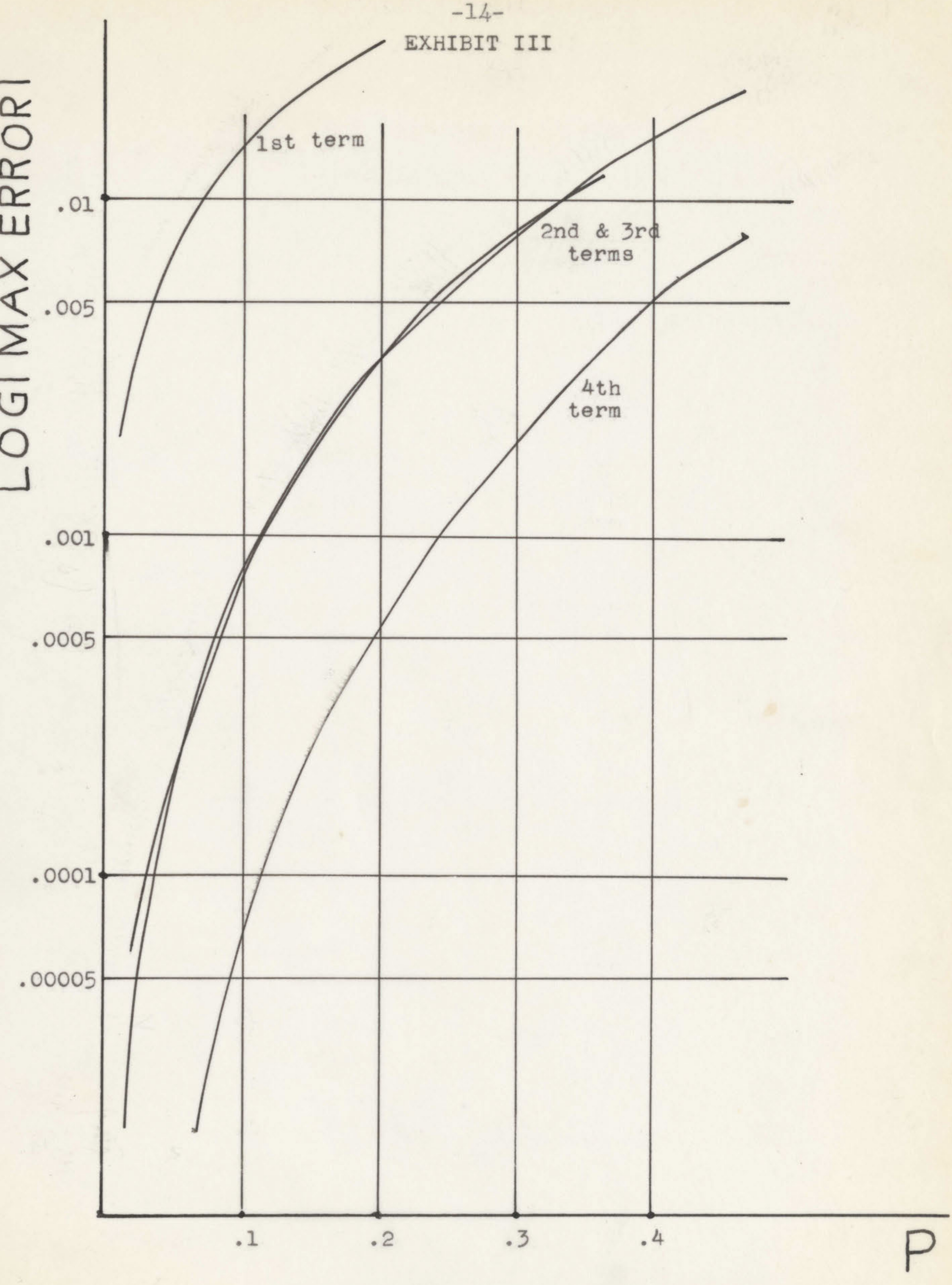
X	CUM P(N)	P(N)	1ST DIFF	2ND DIFF	3RD DIFF	4TH DIFF
0.	0.02732372	0.02732372	0.02732372	0.02732372	0.02732372	0.02732372
1.	0.12568914	0.09836541	0.07104168	0.04371796	0.01639423	-0.01092950
2.	0.30274688	0.17705774	0.07869232	0.00765063	-0.03606732	-0.05246155
3.	0.51521615	0.21246927	0.03541153	-0.04328078	-0.05093142	-0.01486409
4.	0.70643848	0.19122233	-0.02124694	-0.05665847	-0.01337768	0.03755373
5.	0.84411855	0.13768007	-0.05354226	-0.03229532	0.02436315	0.03774083
6.	0.92672658	0.08260803	-0.05507203	-0.00152977	0.03076555	0.00640240
7.	0.96921071	0.04248413	-0.04012390	0.01494812	0.01647789	-0.01428766
8.	0.98832856	0.01911785	-0.02336627	0.01675763	0.00180951	-0.01466838
9.	0.99597570	0.00764714	-0.01147071	0.01189555	-0.00486207	-0.00667158
10.	0.99872867	0.00275297	-0.00489417	0.00657654	-0.00531901	-0.00045694
11.	0.99962964	0.00090097	-0.00185200	0.00304217	-0.00353437	0.00178464
12.	0.99989992	0.00027028	-0.00063068	0.00122131	-0.00182085	0.00171351
13.	0.99997477	0.00007485	-0.00019544	0.00043523	-0.00078607	0.00103477
14.	0.99999402	0.00001924	-0.00005560	0.00013983	-0.00029539	0.00049067
15.	0.99999863	0.00000461	-0.00001463	0.00004097	-0.00009886	0.00019653
16.	0.99999966	0.00000104	-0.00000358	0.00001104	-0.00002992	0.00006893
17.	0.99999988	0.00000022	-0.00000081	0.00000276	-0.00000829	0.00002164
18.	0.99999992	0.00000004	-0.00000017	0.00000064	-0.00000212	0.00000617
19.	0.99999993	0.00000001	-0.00000003	0.00000013	-0.00000050	0.00000161

EXHIBIT II



LOG MAX ERROR vs P for n = 25

EXHIBIT III



LOG MAX ERROR vs P for n = 50

EXHIBIT IV

MAX ERROR vs P