AN APPROXIMATION OF THE CUMULATIVE BINOMIAL PROBABILITY DISTRIBUTION

by

JACK M. HAMILTON, B.S. in I.E.

A THESIS

IN

INDUSTRIAL ENGINEERING

Submitted to the Graduate Faculty
of Texas Technological College
in Partial Fulfillment of
the Requirements for
the Degree of

MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

Approved

Accepted

AC
805
T3
1965
No. 11
Cop. 2

ACKNOWLEDGMENTS

I am sincerely grateful for the personal interest and able guidance given to me by my committee chairman, Professor Prabhakar M. Ghare, and for the assistance of the other committee members, Professors Richard A. Dudek and Horace E. Woodward, Jr.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

CHAPTER I

INTRODUCTION

Statement of the Problem

The cumulative binomial distribution $B(x;n,p)$ is the probability of x or less successes in n trials with p the probability of success in a single trial. There is no practical and concise procedure for calculating the $B(x;n,p)$ for large values of n. There is no brief tabular presentation of the $B(x;n,p)$ as there is for the cumulative Poisson and normal distributions. Since the binomial distribution represents innumerable systems in reality, individuals using probability theory or mathematical statistics in many fields need to have the values of $B(x;n,p)$ readily available.

The $B(x;n,p)$ is tabulated, but the tabulations are published in volumes which are necessarily lengthy since the tables are double entry tables. For each value of x there is a value of $B(x;n,p)$ for each combination of n and p. Few individuals will have these specialized volumes readily available unless they have frequent need for them.

If one needs a value for $B(x;n,p)$ for which the value of x makes direct calculation impractical, then one

must consult a published volume of $B(x;n,p)$ or use an approximation of $B(x;n,p)$. The most commonly used approximations are the cumulative Poisson and the cumulative normal distributions. These approximations are not suitable in most instances if an accuracy of three decimal places is required. To obtain three-decimal accuracy one may use a published volume of $B(x;n,p)$ or a compact book of 72 pages, Binomial, Normal and Poisson Probabilities, authored and privately published by Ed Sinclair Smith (9). Smith's book provides procedures with necessary tables and charts for obtaining $B(x;n,p)$ to three-decimal accuracy. The book uses six procedures for six different areas of an n versus p plot.

There is clearly a need for a practical and concise procedure for obtaining $B(x;n,p)$ to an accuracy of three decimal places. Such a procedure, if included in standard statistical reference books, would provide a readily available means for obtaining $B(x;n,p)$ to three-decimal accuracy when the value of x makes direct calculation impractical. Thus, values of $B(x;n,p)$ would be more generally available to those individuals in the many fields who only occasionally need the values and therefore do not have the extensive tabulations of $B(x;n,p)$.

In the remainder of this chapter the cumulative binomial distribution is operationally defined, and the currently used approximations to it are discussed.

The Cumulative Binomial Probability Distribution

Let the probability that an event will occur be denoted by p and the probability that the event will fail to occur be denoted by q = 1-p. If the event occurs in a given trial, let the trial be termed a success. If the event fails to occur, let the trial be termed a failure.

Then if n independent trials are attempted, the probability of obtaining precisely x successes may be denoted by

$$b(x;n,p) = C_x^n \, p^x \, q^{n-x} \, . \qquad (1\text{-}1)$$

This is called the binomial probability distribution or, more simply, the binomial distribution. It is also known as the Bernoulli distribution in honor of Jacob Bernoulli who was one of the first mathematicians to develop probability theory for discrete variables (7,p.85).

To derive the formula for b(x;n,p), first determine the probability of x consecutive successes followed by n-x consecutive failures. Since the n events are independent, the probability is

$$p_1 \cdot p_2 \cdots p_x \cdot q_1 \cdot q_2 \cdots q_{n-x} = p^x q^{n-x} \, . \qquad (1\text{-}2)$$

The probability of obtaining precisely x successes and n-x failures is the same for any other order of occurrence, because the same number of p's and q's would occur in the

product merely arranged to correspond to the other order. Thus, the number of possible orders times the probability of a specific order produces $b(x;n,p)$. Now the number of possible orders is the number of permutations of n items taken all at a time when x items (p's) are alike and n-x items (q's) are alike. The number of such permutations is the same as the number of combinations of n things taken x at a time,

$$C_x^n = \frac{n!}{(n-x)!\ x!} \qquad (7,p.85). \qquad (1-3)$$

Thus

$$b(x;n,p) = C_x^n\ p^x\ q^{n-x}\ .$$

The name <u>binomial distribution</u> comes from the relationship of $b(x;n,p)$ and $B(x;n,p)$ to the following binomial expansion:

$$(q+p)^n = q^n + nq^{n-1}\ p + \frac{n(n-1)}{2!}\ q^{n-2}\ p^2 + \cdots + p^n$$

$$= B(n;n,p)$$

$$= \sum_{x=0}^{n} b(x;n,p)\ . \qquad (1-4)$$

The first term of the binomial expansion as shown is the probability of 0 successes out of n independent trials $b(0;n,p)$; the second term is equal to $b(1;n,p)$; and in

general the r+1st term is equal to $b(r;n,p)$. The sum of the first r+1 terms is equal to the probability of r or less successes in n independent trials, $B(r;n,p)$ (7,p.86).

Now the probability that x or less successes will occur in n independent trials is called the cumulative binomial probability distribution, or, more simply, the cumulative binomial. The cumulative binomial may be denoted by

$$B(x;n,p) = \sum_{r=0}^{x} c_r^n \, p^r \, q^{n-r} \quad . \qquad (1\text{-}5)$$

Direct Computation of the Cumulative Binomial

To compute $B(x;n,p)$ directly is not practical when the values of n and x make the computation of many terms necessary. For instance, to compute $B(1000;2000,p)$ would require the computation of 1001 individual point binomial terms and then the summing of the 1001 terms as shown by the equation

$$B(1000;2000,p) = \sum_{r=0}^{1000} b(r;2000,p)$$

$$= \sum_{r=0}^{1000} \frac{2000!}{r!(2000-r)!} \, p^r \, q^{2000-r} \quad .$$

Such a computation would be indeed tedious. A computation of, say, twenty-five terms or more would be tiresome.

There is a relation due to the symmetry of the binomial expansion which is quite useful when applicable. It is

$$B(x;n,p) = 1 - B(n-x+1;n,1-p) \ . \qquad (1-6)$$

The computation of $B(x;n,p)$ directly requires less computation using this relation if x is greater than n/2. An extreme example is the case $B(n-1;n,p)$. Then

$$B(n-1;n,p) = 1-B(0;n,1-p) \ .$$

In this case only one term is computed and subtracted from 1; whereas, if the relation is not used, n terms are computed and summed.

If $B(x;n,p)$ is computed directly, a recurrence relation makes computation of the individual point binomial terms more efficient. Use as a starting point,

$$b(0;n,p) = q^n \ .$$

Then,

$$b(r+1;n,p) = \frac{(n-r)p}{(r+1)q} \ b(r;n,p) \ . \qquad (1-7)$$

The recurrence relation is applied repeatedly until $b(x;n,p)$ is computed. The sum of the point binomial terms thus computed is the desired $B(x;n,p)$.

If equation (1-6) is to be employed, compute $B(n-x-1;n,1-p)$ first. Use as a starting point,

$$b(0;n,1-p) = p^n \text{ .}$$

Then,

$$b(r+1;n,q) = \frac{(n-r)q}{(r+1)p} \; b(r;n,q) \text{ .} \qquad (1-8)$$

The recurrence relation is applied repeatedly until $b(n-x-1;n,1-p)$ is computed. Then the sum of the point binomial terms thus computed, which is $B(n-x-1;n,1-p)$, is subtracted from 1 to obtain $B(x;n,p)$.

The tediousness and, in most cases of large n, the impracticality of direct computation of the cumulative binomial have led those who need its values to the use of published volumes of the function $B(x;n,p)$ and to the use of approximations.

The five most widely used volumes containing tabulations of $B(x;n,p)$ are mentioned here with the extent of their tabulations:

Pearson, Karl, Tables of the Incomplete Beta-Function, Cambridge: The University Press (1934). This volume provides tabular values of the cumulative of the beta distribution $I_x(p,q) = 1 - B(p;q+p-1,x)$ with p and q = 0.5(0.5) 10.5, 11(1)50 and with x = 0.01(0.01)1 to seven decimal places.

National Bureau of Standards, Tables of the Binomial Probability Distribution, Applied Mathematics Series 6 (1950). This volume provides B(x;n,p) with p = 0.01 (0.01)0.5 and with x and n = 1(1)50.

Romig, Harry C., 50-100 Binomial Tables, New York: John Wiley and Sons, Inc. (1953). This volume provides B(x;n,p) with x = 1(1)n, n = 50(5)100, and p = 0.01(0.01)0.05.

"Tables of the Cumulative Binomial Probabilities," Ordnance Corps Pamphlet ORDP 20-1, U.S. Government Printing Office (September 1952). This volume provides B(x;n,p) with n = 1(1)150 and p = 0.01(0.01)0.5.

Harvard Computation Laboratory, Tables of the Cumulative Binomial Probability Distribution, Massachusetts: Harvard University Press (1955). This volume provides B(x;n,p) with x = 0(1)n, n = 1(1)50(2)100(10)200(20)500(50)1000, and p = 0.01(0.01)0.5 and p also equal to ten values which are multiples of 1/12 and 1/16 (4,p.xx).

Approximation of the Cumulative Binomial

Most statistical texts discuss the use of the normal and Poisson distributions to approximate the

binomial distributions. However, the empirical rules regarding the accuracy of the approximations generally are not specific, and are almost as numerous as the textbooks containing them.

Some examples of these empirical rules regarding the Poisson approximation follow.

> It gives a good approximation for large n and very small p (7,p.90).
>
> The approximation is good when p is small and n is large. It is generally considered justifiable to use the approximation when $p<0.1$ (2,p.92).
>
> If $n>50$ while $np<5$, the approximation is very close (10,p.124).

Now, some examples of the empirical rules regarding the normal approximation follow.

> If n is large and p is not small or large, the normal approximation may be used (7,p.110).
>
> The approximation is poor for $p<1/(n+1)$ or $p>n/(n+1)$ and outside the interval $np - 3\sqrt{npq} < x < np + 3\sqrt{npq}$. It is good for p close to 1/2, and Hald indicates it is good if $npq>9$ (2,p.90).
>
> It is very good if both np and nq are greater than 5 (10,p.124).

The empirical rules provided in statistical texts are generally inadequate if one desires to approximate the binomial distribution to some specific accuracy using either the normal or the Poisson distribution. Uspensky defines with an inequality the absolute value of the error in approximating the cumulative binomial with the cumulative normal plus a correction term (12,p.129). Smith defines, with greater precision than Uspensky's inequality,

the useful limits of the same two-term approximation to obtain an accuracy of three decimal places (9,p.31).

Uspensky also defines with an inequality the absolute error in approximating the cumulative normal with the cumulative Poisson as follows:

$$\text{error} < (e^{K}-1) \max [P(x,np), 1-P(x,np)]$$

where

$P(x,np)$ = cumulative Poisson distribution

and

$$K = \frac{np + 1/4 + (np)^3/n}{2(n-np)}$$ (12,pp.135-139).

Smith uses Uspensky's approximations, together with others, to give a full discussion of the approximation of the binomial to an accuracy of three decimal places. He also provides an excellent practical discussion of the accuracy of the normal and Poisson approximations to the binomial.

Smith's book is an excellent self-contained and compact paper of only 72 pages. It contains all of the procedures, graphs, and tables necessary for obtaining the cumulative binomial probability to an accuracy of three decimal places. The following summary of procedures is taken directly from Smith's book (9,pp.4-6).

SUMMARY OF RECOMMENDED PROCEDURES FOR OBTAINING VALUES OF THE CUMULATIVE BINOMIAL PROBABILITY WITHIN 3-DECIMAL ACCURACY UNIVERSALLY

In evaluating the cumulative Binomial probability

$$B \text{ or } B(c,n,p) = \sum_{x=c}^{n} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$ for any point

(c,n,p) in the domain: $0 \leq p \leq 1$, $1 \leq n < \infty$, $0 \leq c \leq n$, the whole domain is divided (see Fig. 1) into six regions* in which respective recommended procedures give values of B within .001.

In region 1, values of B can be found directly from a table (C5) of cumulative Binomial probabilities for $1 \leq n \leq 20$. If a table of B is available for other values of n and p, it will of course be used; otherwise the following approximations to B are available for use in the other regions as stated below. Before computing any values of these approximations, one can refer to graphs of percentage points for .001 and .999, see Figs. 14 and 13 of the report, to see whether it is necessary to compute such values.

In region 2, one can use the Poisson approximation

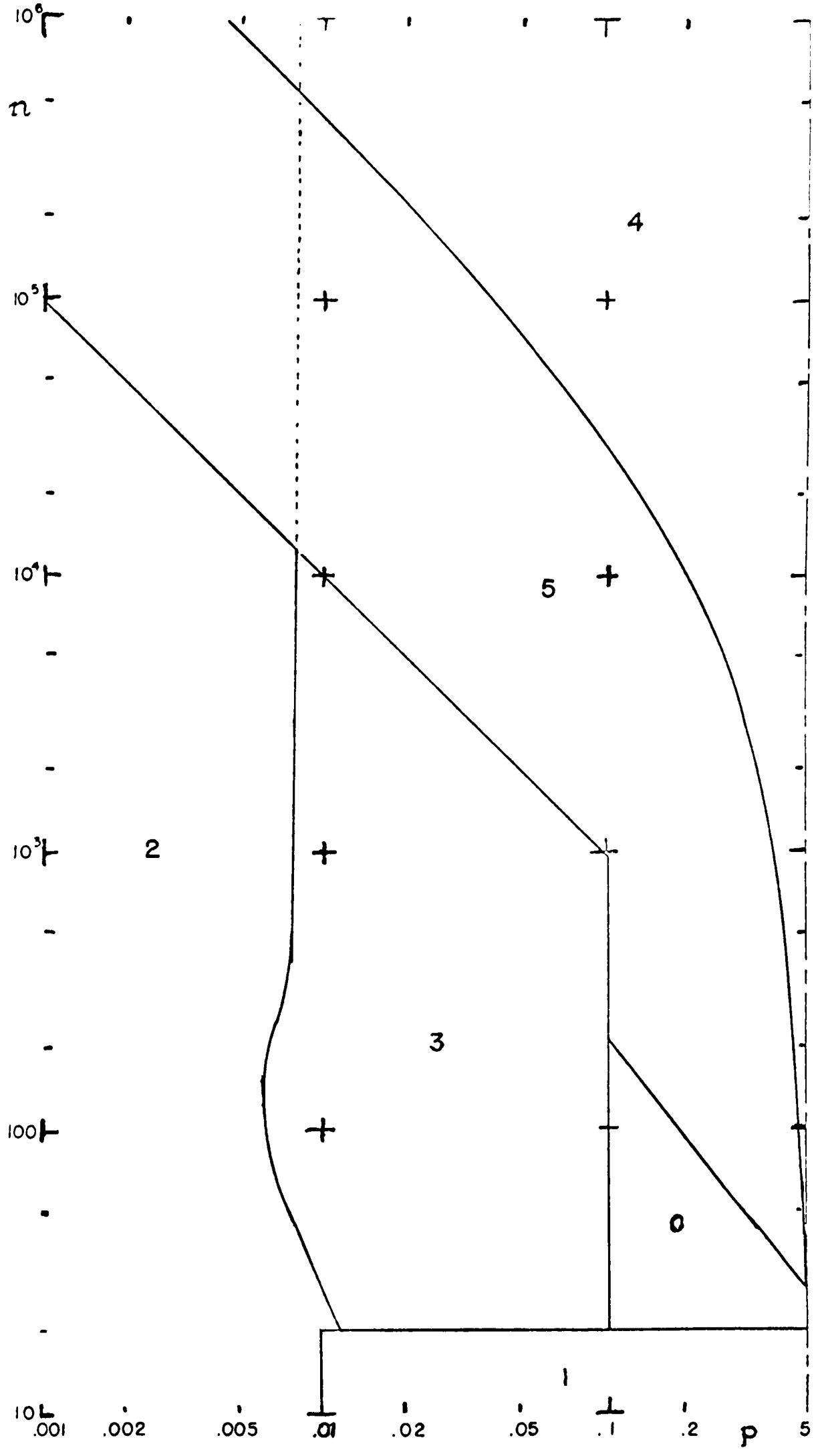
$$P(c,a) = \sum_{x=c}^{\infty} \frac{a^x e^{-a}}{x!}$$ by entering a cumulative Poisson term table (C7) with values of the pair (c,a). Molina has published convenient tables of Poisson terms for $a=np \leq 100$ which is accordingly taken as the upper limit of region 2. For a given n, the maximum error decreases as p approaches zero, from .001 at the right-hand boundary of this region at $p \approx .008$ for $n>20$.

In region 3, one can use the approximation

$$P_B(c,a) = P(c,a) - \frac{np^2}{2}\left[P(c,a) - 2P(c-1,a) + P(c-2,a)\right]$$

where $P(0,a) = P(-1,a) = P(-2,a) = 1$, by entering the cumulative Poisson table with (c,a), $(c-1,a)$ and $(c-2,a)$. This approximation is a 2-term modification of the Gram-Charlier series, type B. The maximum error of this approximation decreases from about .001 at $p=.1$, for $n>20$, to a much lower value at the stated righthand boundary of region 2. While $P_B(c,a)$ can be used to the left of the last named boundary with less than .001 error, this is not necessary since the first term, $P(c,a)$, alone provides this accuracy there.

*For $p>.5$, one uses the relation $B(c,n,p) = 1 - B(n-c+1,n,q)$ and enters the tables or approximations with q instead of p.



SMITH'S PROCEDURES MAP

FIG. 1

In region 4, one can use the Normal approximation

$N(t_c) = \int_{t_c}^{\infty} \phi(t)dt = .5 - \int_0^{t_c} \phi(t)dt$ where $t_c=(c-a-.5)/\sigma$,

$a=np$, $\sigma=\sqrt{npq}$, $q=1-p$ and $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, by entering

a Normal integral table (C6) of values of $\int_0^t \phi(t)dt$

with values of t_c. The maximum error of this approximation decreases as n increases and as p approaches .5, being about .001 at p=.5 and n=28 at the lower end of the lefthand boundary of region 4.

In region 5, one can use the following approximation which comprises the Normal Approximation, $N(t_c)$, and the second term of the Gram-Charlier series, type A:

$N_A(t_c) = N(t_c) - A_1 \phi^{(2)}(t_c)$ where $-A_1 = \frac{q-p}{6\sigma} = \frac{.5-p}{3\sigma}$

and the second derivative $\phi^{(2)}(t_c) = (t_c^2-1)\,\phi(t_c)$.

One uses t_c in entering tables (C6) of the Normal integral, density and/or second derivative of the density. The error of the approximation $N_A(t_c)$ does not exceed substantially .001 at the lefthand boundary of region 5. This error decreases as n increases, for a given p, and as p approaches .5, for a given n. While this approximation can be used in region 4 with much less than .001 error, the second term is of course not needed there to have the error less than .001.

In region 6, one can use the following "remainder" modification of the $N_A(t_c)$ approximation with less than .001 error for plural values of c*:

$N_{Ar} = N(t_c) + \alpha\phi^{(2)}(t_c) + r(t_c)/np$ where $\alpha \tilde{=} -A_1 S = -A_1(1+s) = -A_1(1+0.11/\sigma^2)$

and $r(t_c)$ can be obtained from Fig. 9 of the report.

*For c=0, use $B(0,n,p) = 1$ and, for c=1 and $2<a<2.5$, use $B(1,n,p)=1=q^n$.

Alternatively, α can be obtained from Fig. 8. As long as $.1 \leq p \leq .5$ and $a=np \geq 2$, this approximation (N_{Ar}) can also be used with less than .001 error for values of n outside region 6, but this is not recommended since it is simpler to use tables of B for lower n and the respective approximation N_A or N for higher n. The approximation N_{Ar} is the only one, recommended for cumulative Binomial probabilities in the report, which involves empirical coefficients or curve-fitting.

Example: To find $B(3,25,.10)$, for which $t_c=0$, $np = 2.5$, $npq = 2.25 = \sigma^2$, $\sigma = 1.5$, $-A_1 = \frac{.5-p}{3\sigma} = \frac{.4}{3x1.5} = .08889$, $s = \frac{.11}{npq} = \frac{.11}{2.25} = .04889$ or s can be read from the strip scale for either npq or σ, $\alpha = -A_1(1+s) = .08889x1.04889 = .093235$, and $B(3,25,.10) = \phi^{(-1)}(0) + \alpha\phi^{(2)}(0) + r(0)/2.5$, since r=0 from Fig. 9, $= .5 + (-.39894)(.093235) + 0 = .462805$ which is within .0001 of .462906, the correct value. (6,pp.4-6)

Smith's excellent book filled a very definite need when published and is still essential for binomial approximations for values of n greater than 1000. The book is an excellent reference book on binomial approximations. Since the publication of tables of the cumulative binomial in 1955 which extend the range of n to 1000, a book of approximation procedures for n in the range from 150 to 1000 is no longer essential.

Need for Concise and Practical Approximation

The current need is for a practical procedure for approximating the cumulative binomial to a reasonable accuracy which is concise enough to be suitable for

inclusion in standard statistical reference books and in books of standard mathematical tables.

The binomial distribution is a frequency distribution which represents innumerable systems in reality. Many fields of endeavor, certainly including those using probability theory or mathematical statistics, need to have reasonably accurate values of the cumulative binomial probability readily available. The binomial distribution is the correct distribution to use in applying many statistical quality control techniques.

This paper provides a simple, unified procedure for approximating the cumulative binomial within an accuracy of three decimal places which is concise enough for inclusion in standard statistical reference books and in books of standard mathematical tables.

In Chapter II the general approach to the problem is presented, and several specific approaches which did not lead to acceptable results are discussed. Then in Chapter III the investigation of Wilks' coverage of asymptotic sampling theory is discussed, leading to the concise and practical procedure for approximating the cumulative binomial. This procedure is presented as an algorithm in Chapter IV along with recommendations for further research.

CHAPTER II

APPROACHES TO THE PROBLEM

General Approach

The factorials in the general form of the binomial coefficient make the general form of the binomial distribution not suitable for integration. Integration of the binomial distribution would produce the cumulative binomial. The general approach to the problem was to try to find a suitably accurate approximation (accurate to three decimal places) to the factorial or to the binomial coefficient which when substituted into the binomial distribution would permit integration of the distribution. A logical starting point was Stirling's formula.

Specific Approaches

Stirling's formula

The last few steps of one method for deriving Stirling's formula are as follows. An expression for n! is obtained:

$$n! = A\, n^{n+1/2}\, e^{-n+\theta/(12n)} \qquad (2\text{-}1)$$

where A and θ are unknown constants.

This value for factorials neglecting the quantity $\theta/(12n)$ is substituted into the following formula of Wallis:

$$\lim_{n\to\infty} \frac{2^{2n}(n!)^2}{\sqrt{2n}\,(2n)!} = \sqrt{\pi/2} \quad . \tag{2-2}$$

From this is derived

$$A = \sqrt{2\pi} \quad ,$$

which when substituted into (2-1) gives Stirling's formula in the following form:

$$n! \simeq \sqrt{2\pi}\; n^{n+1/2}\; e^{-n} \quad . \tag{2-3}$$

This remarkable approximation formula gives surprisingly accurate results even for comparatively small values of n. For instance:

$$10! = 3{,}628{,}800$$

and

$$10^{10}\, e^{-10}\sqrt{20\pi} = 3{,}598{,}699 \quad . \ (4, \text{pp. } 94\text{-}95)$$

These results, however, were not accurate enough for our purposes. An attempt was made to improve the accuracy by determining a value for θ. Equation (2-1) was solved for A, giving

$$A = \frac{n!}{n^{n+1/2}\, e^{-n}\, e^{\theta/(12n)}} \quad . \tag{2-4}$$

Then using the relation

$$n! = n(n-1)!$$

and expanding $(n-1)!$ with equation (2-1), A was again solved for and found to be

$$A = \frac{n!}{n(n-1)^{n-1/2}e^{-n+1}e^{\theta/(12(n-1))}} \quad . \qquad (2-5)$$

The right sides of equations (2-4) and (2-5) were equated. The n! factor neatly canceled permitting a solution for θ in terms of n.

$$\theta = 12n(n-1)\ln[n^{n-1/2}e^{-1}(n-1)^{-n+1/2}] \quad . \qquad (2-6)$$

Substituting this value for θ into equation (2-1) gave an expression for n! containing one unknown constant A. This expression was substituted into Wallis' formula (2-2) leading to the following expression for A:

$$A = \lim_{n\to\infty} \frac{\sqrt{\pi/2 \cdot 2n}\ (2n)^{4n^2-n+1}\ (2n-1)^{-4n^2+3n-1/2}}{2^{2n}\ e\ n^{2n^2-n+2}\ (n-1)^{-2n^2+3n-1}} \quad . \qquad (2-7)$$

An indication of the limiting value for A was obtained by using a computer. The value for A with increasing value for n appeared to be converging on 2 as a limit. The following few values for n and A indicated the trend:

n	A
5	2.069
10	2.036
100	2.0049
1000	2.00037
10000	2.000037
50000	2.0000075
90000	2.0000041

Substituting the value of 2 for A and (2-6) for θ into (2-1) gave the following expression for n!

$$n! \approx 2\ n^{n^2-1/2n+1} e^{1-2n} (n-1)^{-n^2+3/2n-1/2} . \quad (2-8)$$

This formula disappointingly gave results which were less accurate than the formula obtained by omitting the $\theta/(12n)$ term (2-3). For instance, the value for 10! obtained from formula (2-8) was 2,896,800. This value, while of the same order of magnitude as the correct answer, was much less accurate than the answer obtained from (2-3).

The value for A could be found in a different manner. Equation (2-8) was used with the 2 replaced by the constant A. Using known values of factorials, values for A were found as follows:

n	A
10	2.5054
50	2.5620
100	2.7951

So A was not constant; neither its values nor the logarithms of its values were linear in relation to n.

This attempt to improve the accuracy of Stirling's formula proved to be unsuccessful.

Stirling's formula with series

Stirling's formula with series was next considered in looking for a suitably accurate approximation of the factorial. One formula with series is obtained by using the Euler-Maclaurin formula to obtain the sum of logarithms. This formula with series is

$$x! \approx x^{x+1/2} e^{-x} \sqrt{2\pi}\, e^{S_n}, \qquad (2\text{-}9)$$

where

$$S_n = \sum_{r=1}^{\infty} \frac{B_{2r}}{2r(2r-1)x^{2r-1}},$$

B_i are Bernoulli numbers,

and

$$S_n = + \frac{1}{12x} - \frac{1}{360x^3} + \frac{1}{1260x^5} - \cdots \quad (11, \text{pp. } 128\text{-}136).$$

Another Stirling formula with series was taken from Beckenbach (1, p. 136). It was

$$x! \approx \left(\frac{x+1}{e}\right)^{x+1} \left(\frac{2\pi}{x+1}\right)^{1/2} S_n \qquad (2\text{-}10)$$

$$S_n = 1 + \frac{1}{12(x+1)} + \frac{1}{288(x+1)^2} - \frac{139}{51840(x+1)^3}$$

$$- \frac{571}{2\ 488\ 320(x+1)^4} + \cdots$$

This Stirling series appeared to converge more rapidly for small x than did the series derived from the Euler-Maclaurin formula. The first three terms of S_n gave suitable accuracy even for relatively small values of x. The use of the first three terms gave a value for the factorial of nine which was accurate to four significant figures.

An expression approximating the value of b(x;n,p) could be obtained by substituting the first three terms of the factorial approximation (2-10) into the formula for b(x;n,p),(1-1). The expression thus obtained could then be integrated over limits adjusted for discreteness to give an approximation for B(x;n,p) as follows:

$$B(x;n,p) \approx \int_{r=-0.5}^{x+0.5} b(r;n,p)dr$$

$$b(r;n,p) \approx \frac{f_s(n)}{f_s(n-r)\cdot f_s(r)}\ p^r\ q^{n-r} \qquad (2\text{-}11)$$

where $f_s(x)$ represents the approximation for x! shown in equation (2-10).

Unfortunately, both $f_s(n-r)$ and $f_s(r)$ contained the variable of integration r raised to the power r. This factor was not integrable. Manipulation of the function

proved fruitless in attempting to get rid of the factor. The Stirling series (2-9) obtained from the Euler-Maclaurin formula also contained this troublesome factor.

Log $C(^{n}_{x})$

The apparent linearity of the logarithm of the binomial coefficient was investigated as a possible approach to approximating the binomial coefficient. A cursory inspection of a table of logarithms of $C(^{n}_{x})$ shows that some of the values closely follow linear patterns. For instance, a sequence of values for log $C(^{n}_{x})$ where x=3 is shown below:

2.21748
2.34242
2.45637
2.56110
2.65801
2.74819

Values of log $C(^{n}_{x})$ were plotted against x for the ratio n to x equal to 1/2, 1/4, and 1/8; a line was plotted for each ratio. The lines, though appearing to be, were not quite linear for values of x greater than 10.

For instance, the line for x/n=1/2 had slopes as follows:

From	To	Slope
x=10	x=50	0.59343
x=25	x=50	0.59608
x=10	x=25	0.58901

The slopes were quite close.

The same values of log $C(^{n}_{x})$ used for the three lines above were plotted on semi-log and log-log graph paper. The plots were not linear. In a further effort to find an integrable function for the binomial coefficient, the values were differenced through the eighth difference without the differences becoming constant. The first difference as expected was almost constant. Succeeding differences diverged.

Binomial coefficient as a ratio of two polynomials

Another investigation was made of the binomial coefficient in an effort to put an approximation of it into integrable form. The binomial coefficient $C(^{n}_{x})$ was put into the form

$$C(^{n}_{x}) = \frac{n!}{(n-x)!x!} = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)(x-2)\cdots 1} \text{ .} \qquad (2\text{-}12)$$

It was noted with interest that the numerator and denominator both had x number of factors and also that the roots of the factors in the numerator were the same as those in the denominator. Thus, the numerator and denominator were identical except the numerator was in variable n and the denominator was in variable x.

When the factors of the numerator and denominator were multiplied together, the result was

$$C\binom{n}{x} = \frac{n^x - A_1 n^{x-1} + A_2 n^{x-2} \mp \cdots + A_x n}{x^x - A_1 x^{x-1} + A_2 x^{x-2} \mp \cdots + A_x x} = \frac{f(n)}{f(x)} \,. \qquad (2\text{-}13)$$

The coefficients A_i were functions of x and could be determined using the rules governing the expressing of coefficients in terms of roots (6,p.227). These rules gave us

A_1 = sum of roots;

A_2 = sum of products of roots taken two at a time;

A_3 = sum of products of roots taken three at a time; etc.;

$\vdots$

A_x = product of all the roots.

Determining the A_i in general form was somewhat simplified, since there were always x roots whose values were $0,1,2,\cdots,x-1$.

$$A_1 = 1+2+3+\cdots+(x-1) = \frac{x(x-1)}{2} \,, \qquad (2\text{-}14)$$

using Gauss' simple scheme for getting the sum of such a series.

$$A_2 = \sum_{K=1}^{x-2} \frac{K}{2}[x(x-1)-K(K+1)] \qquad (2\text{-}15)$$

$$A_3 = \sum_{K=2}^{x-2} \sum_{J=1}^{K-1} \frac{JK}{2}[x(x-1)-K(K+1)] \qquad (2\text{-}16)$$

$$A_4 = \sum_{K=3}^{x-2} \sum_{J=2}^{K-1} \sum_{I=1}^{J-1} \frac{IJK}{2} [x(x-1)-K(K+1)] \quad (2-17)$$

The equations for A_2, A_3, and A_x were arrived at by grouping the products in orderly arrangements and writing down the function. The general form for A_m appears logically to be

$$A_m = \sum_{K=m-1}^{x-2} \sum_{J=m-2}^{K-1} \sum_{I=m-3}^{J-1} \cdots \sum_{B=2}^{C-1} \sum_{A=1}^{B-1} g(x)$$

where

$$g(x) = \frac{A \cdot B \cdots I \cdot J \cdot K}{2} [x(x-1)-K(K-1)]$$

and

A, B, $\cdots$ K occur only if > 0. (2-18)

The ratio of polynomials in (2-13) was such a neat and orderly function that it seemed there should be some simple method to evaluate it. It is regrettable that this investigation did not reveal such a simple method.

Other specific approaches

The approaches discussed so far are the most interesting of the nonproductive approaches studied in this investigation. Two other approaches are mentioned below without detailed discussion, because it quickly became

apparent when they were investigated that they would not yield an approximation of the desired accuracy.

A function of the form

$$\frac{A}{\sqrt{2\ npq}}\ e^{-\frac{B}{2}\frac{(x-np)^2}{npq}}$$

was fitted to B(x;n,p) in the least squares sense. The values of A and B obtained did not give B(x;n,p) to the desired accuracy. The form of the function was suggested by the form of the Gram-Charlier series type A used by Smith (9,p.23).

The Poisson distribution was considered. Plots of curves of the difference between B(x;n,p) and the cumulative Poisson were looked at. The forms of the two functions were manipulated and compared.

Many small tangential investigations were explored along with those discussed. Investigation finally turned to an evaluation of the asymptotic expansion of the distribution of a sample sum as an approximation of B(x;n,p). This approach yielded results within the desired accuracy of three decimal places. The details of this approach are given in Chapter III.

CHAPTER III

APPROXIMATION OF THE CUMULATIVE BINOMIAL PROBABILITY DISTRIBUTION

Introduction

Turning to asymptotic expressions leads to the conclusion of the investigation covered in this paper. Professor Ghare, the chairman of the advisory committee, suggested the investigation of Wilks' coverage of asymptotic sampling theory for large samples (13,pp.254-276). The use of an asymptotic expansion of the distribution of a sample sum produces an approximation of the cumulative binomial probability distribution with the desired accuracy of three decimal places.

Evaluation of the Asymptotic Expansion of the Distribution of a Sample Sum

Explanation of the development of the asymptotic expansion will not be attempted. Wilks' development is quoted below.

ASYMPTOTIC EXPANSION OF DISTRIBUTION OF SAMPLE SUM

> Theorem 9.2.1 contains a statement of the limiting form of the distribution of $(z - n\mu)/\sqrt{n}\sigma$ as $n \to \infty$. One problem which arises here is the determination, for large values of n, of a higher degree of approximation to the distribution of $(z - n\mu)/\sqrt{n}\sigma$ than that provided by the distribution N(0,1). We shall examine this problem.

Suppose the central moments $\mu_1, \ldots, \mu_r, r>2$, of the population c.d.f. $F(x)$ exist and are finite. Then if $\phi(t)$ is the characteristic function of $(x - \mu)/\sqrt{n}\sigma$, we have

(9.4.1)

$$\phi(t) = 1 - \frac{t^2}{2n} + \sum_{j=3}^{r} \frac{(it)^j \alpha_j}{j!(\sqrt{n})^j} + 0\left(\left(\frac{t}{\sqrt{n}}\right)^r\right)$$

where $x_j = \mu_j/\sigma^j$ and $n^{1/2r} \cdot 0((t/\sqrt{n})^r)$ tends to zero as $n \to \infty$ for any $t \neq 0$. But the characteristic function of $(z - n\mu)/\sqrt{n}\sigma$, namely, $\phi_n(t)$, is given by

(9.4.2)

$$\phi_n(t) = [\phi(t)]^n .$$

Taking logarithms, we find

(9.4.3)

$$\log \phi_n(t) = - \frac{t^2}{2} + n \sum_{j=3}^{r} \frac{(it)^j (K_j^*)}{j!(\sqrt{n})^j} + n \cdot 0\left(\left(\frac{t}{\sqrt{n}}\right)^r\right)$$

where the K_j^* are semi-invariants of the distribution of $(x - \mu)/\sigma$ in the population. Therefore, we have

(9.4.4)

$$\phi_n(t) = e^{-1/2t^2} \exp\left[n \sum_{j=3}^{r} \frac{(it)^j (K_j^*)}{j!(\sqrt{n})^j} + n \cdot 0\left(\left(\frac{t}{\sqrt{n}}\right)^r\right)\right] ,$$

which can be written as

(9.4.5)

$$\phi_n(t) = e^{-1/2t^2}\left[1 + \sum_{j=1}^{r-2} \frac{\mu_j(it)}{(\sqrt{n})^j} + n \cdot 0\left(\left(\frac{t}{\sqrt{n}}\right)^r\right)\right]$$

where $\mu_j(it)$ is a polynomial of degree $3j$ in (it) whose coefficients are functions of the K_j^*'s but do

not depend on n. The lowest power of (it) in $\mu_j(it)$ is j + 2.

If we let

(9.4.6)

$$F_n(x) = P\left(\frac{(z - n\mu)}{\sqrt{n}\sigma} < x\right)$$

and put $x' = (x + y)/2$, $\delta = (x - y)/2$ in (5.1.14), then if x and y are continuity points of $F_n(x)$,

(9.4.7)

$$F_n(x) - F_n(y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\sin\left(\frac{x - y}{2}\right)t}{t} e^{-it[1/2(x+y)]}\phi_n(t)dt \quad .$$

Substituting the expression for $\phi_n(t)$ from (9.4.5) and simplifying, we obtain

(9.4.8)

$$F_n(x) - F_n(y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{(e^{-itx} - e^{-ity})}{2(-it)}$$

$$= \frac{1}{\pi} e^{-1/2t^2}\left[1 + \sum_{j=1}^{r-2} \frac{\mu_j(it)}{(\sqrt{n})^j} + n \circ 0\left(\left(\frac{t}{\sqrt{n}}\right)^r\right)\right]dt.$$

But it can be verified without particular difficulty that

(9.4.9)

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{(e^{-itx} - e^{-ity})}{2(-it)} e^{-1/2t^2} dt = \Phi(x) - \Phi(y)$$

where $\Phi(x)$ is the c.d.f. of the distribution N(0,1). All other terms in (9.4.7) are of the following form, except for a constant multiplier,

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} (-it)^j (e^{-itx} - e^{-ity}) e^{-1/2t^2}\, dt$$

which has the value

(9.4.10)

$$\frac{1}{2\pi}\frac{d^j}{dx^j}\int_{-\infty}^{\infty} e^{-itx-1/2t^2}\, dt - \frac{1}{2\pi}\frac{d^j}{dy^j}\int_{-\infty}^{\infty} e^{-ity-1/2t^2}\, dt =$$

$$\Phi^{(j+1)}(x) - \Phi^{(j+1)}(y)$$

where

(9.4.11)

$$\Phi^{(j+1)}(x) = \frac{d^{j+1}}{dx^{j+1}}\Phi(x) = \frac{d^j}{dx^j}\left(\frac{1}{\sqrt{2\pi}} e^{-1/2x^2}\right).$$

Hence, we obtain for (9.4.7)

(9.4.12)

$$F_n(x) - F_n(y) = \Phi(x) - \Phi(y) + \sum_{j=1}^{r-2} \frac{[\mu_j^*(x) - \mu_j^*(y)]}{\sqrt{n}^j} + O\left(\frac{1}{\sqrt{n}^{r-2}}\right)$$

where $\mu_j^*(x)$ and $\mu_j^*(y)$ are the functions one obtains by replacing $(it)^p$ in $\mu_j(it)$ by $\Phi^{(p+1)}(x)$ and $\Phi^{(p+1)}(y)$ respectively.

If we let $y \to -\infty$ in (9.4.12), we obtain as the asymptotic expansion of $F_n(x)$,

(9.4.13)

$$F_n(x) = \Phi(x) + \sum_{j=1}^{r-2} \frac{\mu_j^*(x)}{\sqrt{n}^j} + O\left(\frac{1}{\sqrt{n}^{r-2}}\right).$$

If $F_n(x)$ has a p.d.f. $f_n(x)$, we find by taking the derivative of (9.4.13) with respect to x that

(9.4.14)

$$f_n(x) = \Phi^{(1)}(x) + \sum_{j=1}^{r-2} \frac{\mu_j^{*\prime}(x)}{\sqrt{n}^j} + O\left(\frac{1}{\sqrt{n}^{r-2}}\right),$$

where $\mu_j^{*\prime}(x)$ is the first derivative of $\mu_j^*(x)$.

As a matter of fact, even if $F_n(x)$ has no p.d.f. (that is, $F_n(x)$ may be a discrete c.d.f.), (9.4.14) is still a useful approximation. In this case the right-hand side of (9.4.14) is a representation of a p.d.f. such that $F_n(x)$ is approximated by integrating this p.d.f. from $-\infty$ to x.

We shall not write out the general expression for the function $\mu_j^*(r)$. But it is of some interest to write out the expressions on the right-hand sides of (9.4.13) and (9.4.14) to terms of order $n^{-\frac{3}{2}}$. These are

(9.4.15)

$$F_n(x) = \Phi(x) - \frac{1}{\sqrt{n}} \frac{\alpha_3}{3!} \Phi^{(3)}(x)$$

$$+ \frac{1}{n}\left[\frac{1}{4!}(\alpha_4 - 3)\Phi^{(4)}(x) + \frac{10}{6!}\alpha_3^2\Phi^{(6)}(x)\right]$$

$$- \frac{1}{n^{\frac{3}{2}}}\left[\frac{1}{5!}(\alpha_5 - 10\alpha_3)\Phi^{(5)}(x) + \frac{35}{7!}\alpha_3(\alpha_4 - 3)\Phi^{(7)}(x)\right.$$

$$\left. + \frac{280}{9!}\alpha_3^3\Phi^{(9)}(x)\right] + O\left(\frac{1}{n^{\frac{3}{2}}}\right)$$

and

(9.4.16)

$$f_n(x) = \Phi^{(1)}(x) - \frac{1}{\sqrt{n}}\frac{\alpha_3}{3!}\Phi^{(4)}(x)$$

$$+ \frac{1}{n}\left[\frac{1}{4!}(\alpha_4 - 3)\Phi^{(5)}(x) + \frac{10}{6!}\alpha_3^2\Phi^{(7)}(x)\right]$$

$$- \frac{1}{n^{\frac{3}{2}}}\left[\frac{1}{5!}(\alpha_5 - 10\alpha_3)\Phi^{(6)}(x) + \frac{35}{7!}\alpha_3(\alpha_4 - 3)\Phi^{(8)}(x)\right.$$

$$\left. + \frac{280}{9!}\alpha_3^3\Phi^{(10)}(x)\right] + O\left(\frac{1}{n^{\frac{3}{2}}}\right)$$

We may summarize these results, which were originally obtained by Edgeworth (1905), as follows:

9.4.1 If $(x_1,\ldots,x_n)$ is a sample from a distribution with finite moments $\mu_1,\ldots,\mu_r$ $(r>2)$, then the c.d.f. $F_n(x)$ of $(z = n\mu)/\sqrt{n}\sigma$ can be expanded in the form (9.4.13), the explicit expansion to terms of order $n^{-\frac{3}{2}}$ being given by (9.4.15).

The quantities α_3 $(= \mu_3/\sigma^3)$ and $\alpha_4 - 3$ $(= \mu_4/\sigma^4 - 3)$, usually denoted by γ_1 and γ_2 respectively, are sometimes called the skewness and kurtosis respectively, of a distribution function having mean μ, variance σ^2, and third and fourth central moments μ_3 and μ_4. These two constants play an important role in the degree to which the c.d.f. $F_n(x)$ can be approximated by the c.d.f. $\Phi(x)$ of the distribution $N(0,1)$. It will be noted from an inspection of (9.4.15) that, in general, $F_n(x)$ is approximated by $\Phi(x)$ except for terms of order $1/\sqrt{n}$. But the following corollary of 9.4.1 gives conditions under which higher orders of approximation hold:

9.4.1a If the skewness of the distribution from which $(x_1,\ldots,x_n)$ is drawn is zero, $F_n(x)$ is

> approximated by $\Phi(x)$ except for terms of order $1/n$; if both the skewness and kurtosis are zero, $F_n(x)$ is approximated by $\Phi(x)$ except for terms of order $1/\sqrt{n}$.

It should be pointed out that Lyapunov (1901) was the pioneer on the problem of determining higher degrees of approximation to the distribution of $\bar{x}$ in large samples than that provided by the normal distribution. Cramer (1937) has shown that the remainder term in (9.4.13) and (9.4.14) is of the same order as the first term neglected. Esseen (1944) has made more recent investigations of the accuracy of such asymptotic expansions. Asymptotic expansions in powers of $1/\sqrt{n}$ have been established for other statistics than sample means by Cramer (1937), Hsu (1945a, 1945b), Chung (1946), and others. An expository article on asymptotic approximations to distributions with an extensive bibliography has been published by Wallace (1958). (13,pp.262-266)

The use of z and x in Wilks' formulation is the reverse of the use of z and x in the remainder of this paper. For instance, Wilks' formula $F_n(x)$ (9.4.15) is $F_n(z)$ in the remainder of this paper. Thus, z is the normalized version of x in this paper.

Wilks' formula for $F_n(z)$ (9.4.15) is used to approximate $B(x;n,p)$. When the central moments of the point binomial are used in the formula, then $F_n(z)$ represents asymptotically the probability that the sum of successes in n trials of one each taken from a point binomial distribution will be $\leq x$.

The central moments or moments about the mean of the binomial distribution and the required values for the accumulative normal distribution are first computed. To compute the central moments let

$$P[x=0] = 1-p$$

and

$$P[x=1] = p.$$

Then

$$E(x) = 0(1-p)+1(p) = p,$$

$$E(x^2) = 0^2(1-p)+1^2(p) = p,$$

and

$$E(x^r) = p. \qquad (3-1)$$

Let

$$y = x-E(x) = x-p.$$

Then

$$P[y=-p] = P[x=0] = 1-p,$$

and

$$P[y^r=(o-p)^r] = 1-p;$$

$$P[y=1-p] = P[x=1] = p,$$

and

$$P[y^r=(1-p)^r] = p.$$

Thus

$$E(y) = -p(1-p)+(1-p)p=0, \qquad (3-2)$$

and

$$E(Y^r)=(-p)^r(1-p)+(1-p)^r p. \qquad (3-3)$$

The required central moments are computed as follows from the general formula (3-3).

$$E(y^2)=(-p)^2(1-p)+(1-p)^2p$$

$$= p^2(1-p)+(1-p)^2p$$

$$= p(1-p)[p+(1-p)]$$

$$= p(1-p). \qquad (3-4)$$

$$\sigma = \sqrt{p(1-p)} \qquad (3-5)$$

Similarly,

$$E(y^3) = p(1-p)(1-2p), \qquad (3-6)$$

$$E(y^4) = p(1-p)(1-3p+3p^2), \qquad (3-7)$$

and

$$E(y^5) = p(1-p)(1-4p+6p^2-4p^3). \qquad (3-8)$$

In Wilks' formulation with x changed to z

$$\phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-1/2z^2} dz \qquad (3-9)$$

The derivatives of $\phi(z)$ are set down using the Tchebycheff-Hermite polynomials (8,p.196)

$$\phi^{(1)}(z) = \frac{1}{\sqrt{2\pi}} e^{-1/2z^2}, \qquad (3-10)$$

$$\phi^{(2)}(z) = -z\phi^{(1)}(z), \qquad (3-11)$$

$$\phi^{(3)}(z) = (z^2-1)\,\phi^{(1)}(z), \qquad (3-12)$$

$$\phi^{(4)}(z) = -(z^3-3z)\,\phi^{(1)}(z), \qquad (3-13)$$

$$\phi^{(5)}(z) = (z^4-6z^2+3)\phi^{(1)}(z), \qquad (3-14)$$

$$\phi^{(6)}(z) = -(z^5-10z^3+15z)\ \phi^{(1)}(z)\ , \qquad (3\text{-}15)$$

$$\phi^{(7)}(z) = (z^6-15z^4+45z^2-15)\ \phi^{(1)}(z)\ , \qquad (3\text{-}16)$$

$$\phi^{(8)}(z) = -(z^7-21z^5+105z^2-105z)\ \phi^{(1)}(z)\ , \qquad (3\text{-}17)$$

and

$$\phi^{(9)}(z) = (z^8-28z^6+210z^4-420z^2+105)\ \phi^{(1)}(z)\ . \qquad (3\text{-}18)$$

Substituting values into $F_n(z)$ gives as the first term $T_1(z)$ of the asymptotic expansion the following:

$$T_1(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-1/2z^2}\, dz\ . \qquad (3\text{-}19)$$

$T_1(z)$ is the cumulative normal distribution.

The second term is

$$T_2(z) = -K(z^2-1)\phi^{(1)}(z) \qquad (3\text{-}20)$$

where

$$K = \frac{1}{n^{1/2}} \frac{1-2p}{6(p-p^2)^{1/2}}\ ,$$

and the third term is

$$T_3(z) = -(A+B)\phi^{(1)}z \qquad (3\text{-}21)$$

where

$$A = \frac{1}{24n}\left(\frac{1}{p-p^2} - 6\right)(z^3-3z)$$

and

$$B = \frac{K^2}{2}(z^5-10z^3+15z) \text{ .}$$

The fourth term, not shown, contains approximately three times the number of individual factors as does the third term. Fortunately, it is not necessary to use the fourth term. The first two terms of the expansion are equivalent to the first two terms of the Gram-Charlier Series, Type A which Smith uses as one of his approximating forms for $B(x;n,p)$ (9,p.23).

Computations for evaluation of the series are accomplished on the IBM 1620 Mod II Electronic Computer with Disc Packs. The initial limited evaluation of the series is made using the entering argument

$$z = \frac{x-np}{\sqrt{np(1-p)}} \text{ .} \qquad (3\text{-}22)$$

The error of the approximation obtained is much larger than desired. To improve the approximation the entering argument is adjusted for discreteness to

$$z = \frac{(x+1/2)-np}{\sqrt{np(1-p)}} \text{ .} \qquad (3\text{-}23)$$

The approximation now proves to have the desired accuracy over certain combinations of the variable and parameters of $B(x;n,p)$.

The initial evaluation work on the series is accomplished using a combination of computer computations and manual operations. The terms $T_2(x)$, $T_3(x)$, and $T_4(x)$ are computed on the computer, and $T_1(x)$ and $B(x;n,p)$ are extracted from tables. The desired terms are manually combined and compared with $B(x;n,p)$.

The series is to be evaluated over a map of n versus p with n ranging from 20 to 1000 and p ranging from 0.001 to 0.5. To evaluate the series over one (n,p) point requires the consideration of three variations of the series: only the first term of the series (a one-term series), the first two terms of the series (a two-term series), and the first three terms of the series (a three-term series).

To evaluate any one of the one, two, or three-term series over a specific (n,p) point, the series is first computed and compared to $B(x;n,p)$ for the maximum value of x which obtains a value of $B(x;n,p)$ less than 0.999. If the series value differs from $B(x;n,p)$ by less than 0.001, the value of x is reduced by one and the series value is again compared to $B(x;n,p)$. This procedure is repeated until one of two things happens. One, the minimum value of x is determined which in the series produces a value differing from $B(x;n,p)$ by more than 0.001. Two, the series produces values differing from $B(x;n,p)$ by less than 0.001 for all values of x for which

$0.001 \leq B(x;n,p) \leq 0.999$. For some (n,p) points this requires over one hundred repetitions of the procedure. Thus, a range of x is determined; with every value of x in this range the series produces a value within 0.001 of $B(x;n,p)$ for a specific (n,p) point. Over certain areas of the n versus p map the one, two, and three-term series all must be evaluated.

The large number of computations involved in the over-all evaluation of the series precludes the use of the manual computation procedures used in the initial evaluation. To obtain other than a limited evaluation requires that the evaluation procedure be computerized as much as possible.

Use of IBM-1620 Computer in Evaluation

The major difficulty in computerizing the evaluation procedure is the lack of computer programs for the cumulative normal and the cumulative binomial distributions. Values for both are used many times for the evaluation of each (n,p) point.

A suitable series approximation for the following form of the cumulative normal $\emptyset(z)$ with the mean equal to zero and variance equal to one is derived as follows:

$$\emptyset(z) = \int_{o}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \ . \qquad (3\text{-}24)$$

The exponential factor is expanded in a Maclaurin series. The integrand is then integrated term by term to give

$$\emptyset(z) = Z - \frac{Z^3}{3\cdot 2} + \frac{Z^5}{5\cdot 2^2\cdot 2!} - \frac{Z^7}{7\cdot 2^3\cdot 3!} + \cdots \quad (3\text{-}25)$$

$$= \sum_{j=1}^{\infty} (-1)^{j+1} \frac{Z^{2j-1}}{(2j-1)\cdot 2^{j-1}\cdot (j-1)!} . \quad (3\text{-}26)$$

This series is a convergent alternating series which is used to obtain $\emptyset(z)$ with an error < 0.00001. The series is not efficient for large values of Z, requiring thirty terms when $Z = 4.2$. In the computer program, however, only the number of terms necessary to obtain desired accuracy is used, and for $Z \leq 1$ only five or less terms are required.

In the evaluation procedure a cumulative binomial is used in the form

$$B_t(x+1;n,p) = 1 - B(x;n,p) .$$

This expediency allows direct checking of the values of $B_t(x;n,p)$ obtained with the computer program with the values of $B_t(x;n,p)$ contained in tables of the cumulative binomial being used (5).

A recurrence procedure is used to obtain the required values for $B_t(x;n,p)$. For the evaluation over

an (n,p) point of any one of the series, an initial value of $B_o(x;n,p)$ is taken from a table and an initial value of $b(x;n,p)$ is computed directly. Subsequent values of $B_t(x;n,p)$ are obtained using the following relations:

$B_o(x;n,p)$ = table value with x equal to one more than the maximum value of x which obtains a value of $B_t(x;n,p)$ less than 0.999.

$b(x;n,p)$ = value computed directly using a five-term series approximation of the factorials (2-10).

$$b(x-1;n,p) = \frac{x}{n-x+1} \circ \frac{q}{p} \circ b(x;n,p) \qquad (3\text{-}27)$$

$$B_t(x-1;n,p) = B_t(x;n,p) + b(x-1;n,p) \qquad (3\text{-}28)$$

Then operations (3-27) and (3-28) are repeated to obtain subsequent values of $B_r(x;n,p)$.

The above procedures for obtaining the required values of the cumulative binomial distribution are shown included in the final evaluating computer program which is contained in the Appendix.

Feller's norming

An attempt is made to obtain better two-term series results by using a norming procedure suggested in an article by Feller (3,p.319). Feller uses limits

different from traditional limits in obtaining a better approximation of B(x;n,p) using the cumulative normal distribution. Feller uses limits as follows:

Traditional	Feller's	
$\frac{(x+1/2)-np}{\sigma}$	$\frac{x+1-(n+1)p}{\sigma}$	(3-29)
$\frac{(x-1/2)-np}{\sigma}$	$\frac{x-(n+1)p}{\sigma}$	(3-30)

Since Feller is approximating the discrete binomial with the continuous normal, it seems that his norming might improve the two-term series which is also continuous. Feller's limits are applied to only the first term and to both terms of the two-term series of the Wilks' asymptotic expansion without improving the series. The results are less accurate in both cases.

Graphic Presentation of Evaluation Results

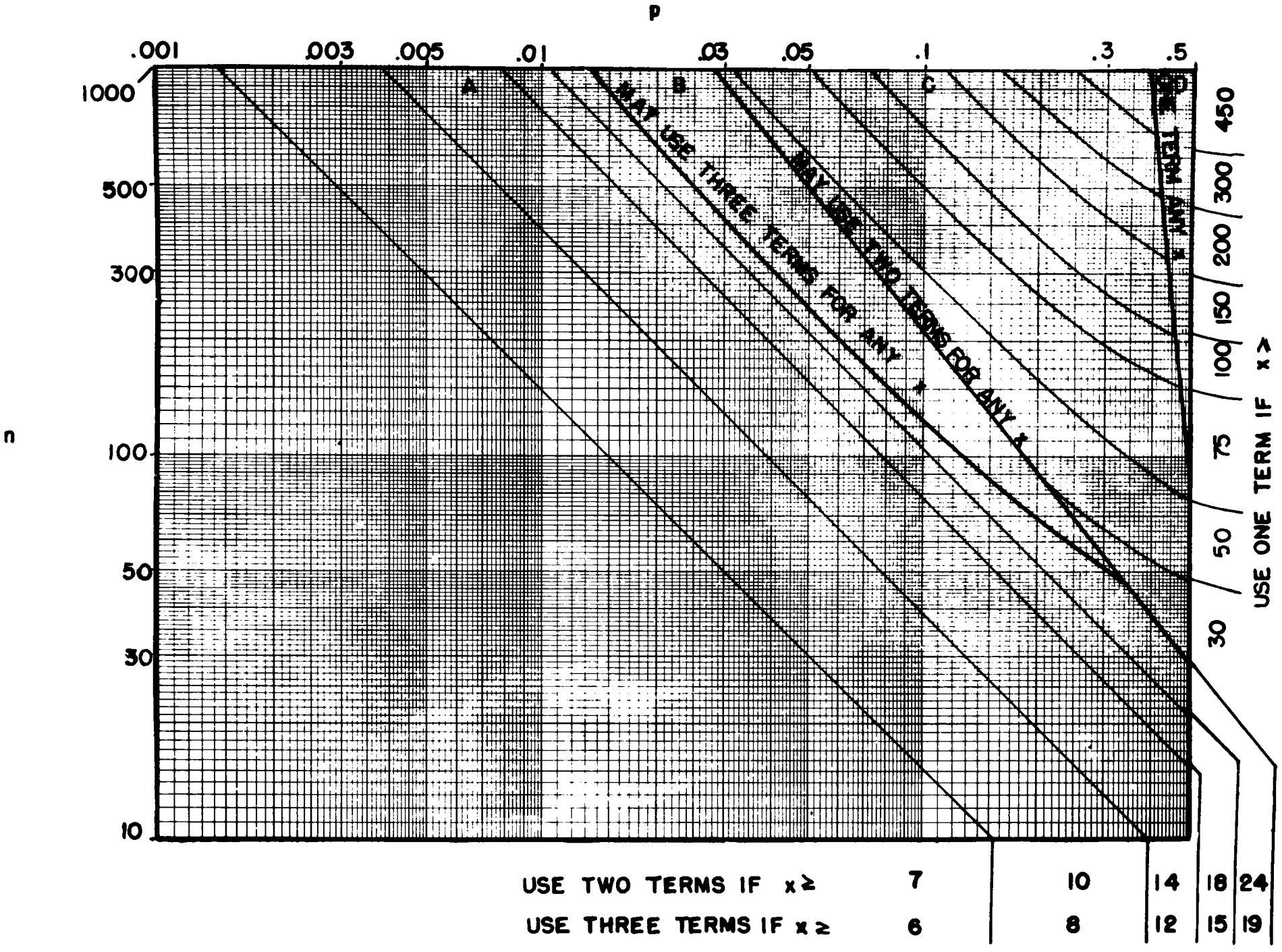
The most useful presentation of the results of the evaluation of the one, two, and three-term series of the Wilks' asymptotic expansion is graphical. Contour lines representing the minimum values of x giving a series approximation of B(x;n,p) with desired accuracy are plotted on a map of n versus p. A separate graph is made initially for the one, two, and three-term series. The contour lines of x can be represented as straight lines

on full logarithmic graph paper without great loss of accuracy. When accuracy is lost, it is through a shift of the line in the direction which will insure that all values of x greater than the value represented by the contour line will still give a series approximation of desired accuracy.

The three separate graphs are consolidated on one graph for simplicity of use and conciseness. The consolidated graph is shown in Figure 2.

The graph is divided into four areas by three dark lines. The leftmost area is designated area A for discussion purposes. The next area toward the right is designated area B, the next C, and the triangular shaped area in the upper right corner is designated area D.

In area A four contour lines are shown. The straight line separating areas A and C and areas B and C represents another contour line. These contour lines represent minimum values of x ensuring three decimal place accuracy of $B(x;n,p)$ when using the two and three-term series. Thus, for any (n,p) point in area A one can determine a minimum value for x which gives a series approximation of $B(x;n,p)$ with an accuracy of three decimal places. For instance, for (n,p) equal $(200,0.03)$ the two-term series gives an approximation of desired accuracy if $x \geq 14$ and the three-term series if $x \geq 12$. Naturally one would use the shortest series giving the desired accuracy.



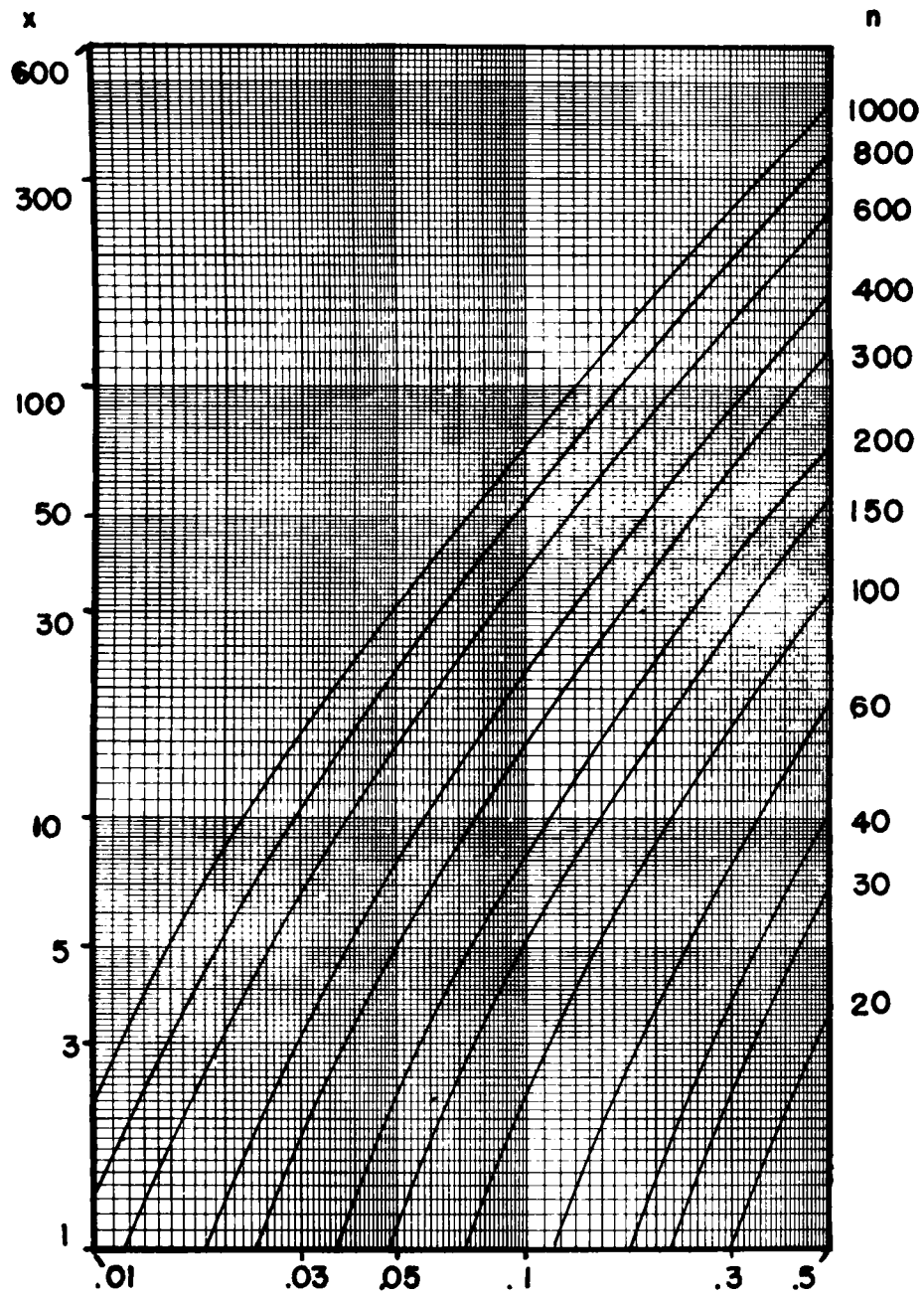
PROCEDURES MAP

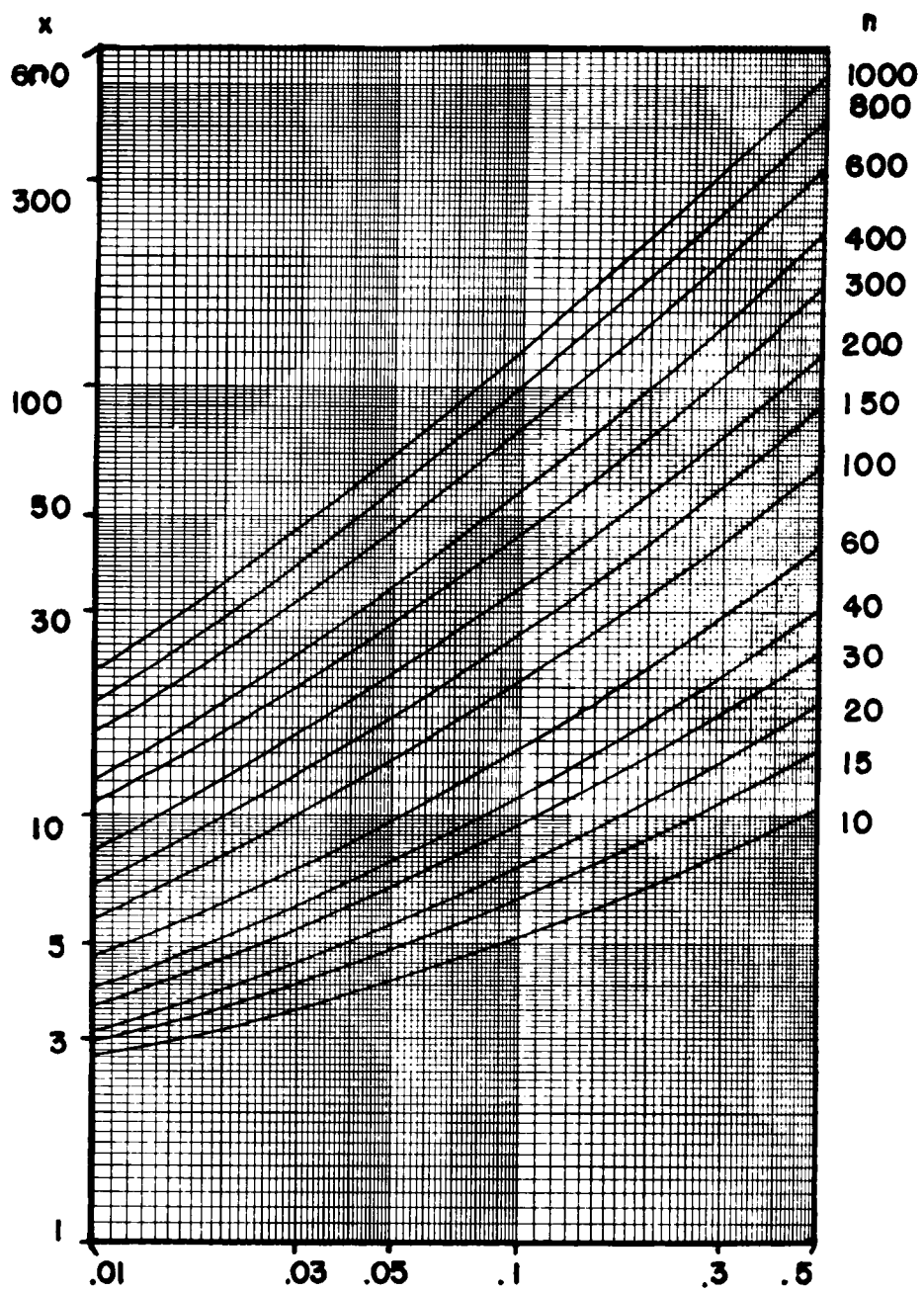
FIG. 2

The gradient across the contour lines is continuous though not linear. One can interpolate for values of x between contour lines if sufficient care is used. Assuming linearity of gradient for interpolation provides a slight safety factor to counterbalance possible interpolation error. For instance, for the example just used with (n,p) equal (200,0.03), one can use the two-term series if $x \geq 13$ rather than 14 and use the three-term series with $x \geq 10$ rather than 12.

In area A if x is not large enough to use either the two or three-term series, then B(x;n,p) must be computed directly. One should consider the values of x on the percentage point graph, Figure 3, when deciding whether to sum b(x;n,p) terms for values of x which are > x or to sum those b(x;n,p) terms for values of x which are $\leq$ x. Probably five or less b(x;n,p) terms past the term for the percentage point value of x must be included in the sum to obtain three decimal place accuracy.

As an example, let (n,p) equal (40,0.3) and x=14. At this (n,p) point x is not large enough to permit use of even the three-term series; it must be $\geq$ 15. Therefore, B(14;40,0.3) must be computed directly. The values of x on the Percentage Point Chart for this (n,p) point are 4 and 22. Obviously it is more efficient to sum b(x;n,p) terms for values of x which are greater than 14 than to sum terms for lower values of x; it requires





PERCENTAGE POINT CHART

FIG. 3

computation of eight b(x;n,p) terms plus overrun rather than eleven plus overrun. The sum of the b(x;n,p) terms for x > 14 is then subtracted from one to obtain B(14;40,0.3). When percentage point values of x are used to determine which b(x;n,p) terms to sum, a maximum of twelve terms plus overrun will be required anywhere in area A to compute B(x;n,p) directly.

In area B the three-term series produces an approximation of B(x;n,p) within three decimal place accuracy for any value of x for which 0.001 < B(x;n,p) < 0.999. The values of x satisfying this requirement can be determined from the percentage point graph in Figure 3. Therefore, in area B the two-term series is used if x is large enough and, if not, the three-term series is used.

In area C only one term of the series is used if x is sufficiently large as determined by the contour lines. If x is not large enough, the two-term series is used. In area D the one-term series produces the approximation to three decimal place accuracy for values of x in the range shown on Figure 3.

The percentage point graph shown in Figure 3 enables one to determine the range of x for each (n,p) point which produces B(x;n,p) such that 0.001 < B(x;n,p) < 0.999. Actually, the two sets of contours are plotted from values of x such that

$$B(x;n,p) < 0.0001 < B(x+1;n,p)$$

for one set and for the other set

$$B(x-1;n,p) < 0.999 < B(x;n,p) .$$

Computing B(x;n,p) using approximating series

The following example problems demonstrate the computing procedures.

Example 1.--area A, two terms

$$n=100,\ p=0.1,\ x=20$$

This (n,p) point on Figure 2 (page 44) is in area A where one can use the two-term series if $x \geq 18$ and the three-term series if $x \geq 15$. Thus, one can use either series and will naturally use the two-term series.

$$z = \frac{x + 0.5 - np}{\sqrt{np(1-p)}} = 3.4$$

$$T_1(z) = \int_0^z N(z)dx = 0.9998$$

$$T_2(z) = -\frac{1}{n^{1/2}} \frac{1-2p}{6(p-p^2)^{1/2}} (z^2-1) \frac{e^{\frac{-z^2}{2}}}{\sqrt{2\pi}} = -0.00035$$

$$T_1(z) + T_2(z) = 0.99945$$

Actual $B(x;n,p)$ = 0.99919

Difference = 0.00026

Example 2.--area A, two terms

n=100, p=0.1, x=18

z = 2.8333

$T_1(z)$ = 0.9977

$T_2(z)$ = -0.00225

$T_1(z) + T_2(z)$ = 0.99545

Actual $B(x;n,p)$ = 0.99542

Difference = 0.00003

Example 3.--area A, two terms

n=1000, p=0.01, x=18

z = 2.7015

$T_1(z)$ = 0.9965

$T_2(z)$ = -0.00356

$T_1(z) + T_2(z)$ = 0.99294

Actual $B(x;n,p)$ = 0.99310

Difference = 0.00016

Example 4.--area A, three terms

n=1000, p=0.01, x=15

TEXAS TECHNOLOGICAL COLLEGE
LUBBOCK, TEXAS
LIBRARY

z	= 1.7480
$T_1(z)$	= 0.9597
$T_2(z)$	= -0.00923
$T_3(z)$	= 0.00128
$T_1(z) + T_2(z) + T_3(z)$	= 0.95175
Actual $B(x;n,p)$	= 0.95213
Difference	= 0.00038

Example 5.--area B, two terms

$n=200,\ p=0.1,\ x=24$

z	= 1.0607
$T_1(z)$	= 0.8556
$T_2(z)$	= -0.00089
$T_1(z) + T_2(z)$	= 0.85471
Actual $B(x;n,p)$	= 0.85511
Difference	= 0.00040

Example 6.--area B, three terms

$n=200,\ p=0.1,\ x=20$

z	= 0.117851
$T_1(z)$	= 0.5469
$T_2(z)$	= 0.01228
$T_3(z)$	= -0.00052

$T_1(z) + T_2(z) + T_3(z)$ = 0.55866

Actual B(x;n,p) = 0.55917

Difference = 0.00051

Example 7.--area C, two terms by interpolation

n=400, p=0.3, x=145

z = 2.7823

$T_1(z)$ = 0.9973

Actual B(x;n,p) = 0.99692

Difference = 0.00038

Example 8.--area C, two terms

n=400, p=0.3, x=100

z = -2.1276

$T_1(z)$ = 0.0167

$T_2(z)$ = -0.000106

$T_1(z) + T_2(z)$ = 0.01659

Actual B(x;n,p) = 0.01553

Difference = 0.00106

Example 9.--area D, one term

n=800, p=0.45, x=350

z = -0.6751

$T_1(z)$ = 0.25023

Actual $B(x;n,p)$ = 0.25001

Difference = 0.00022

Example 10.--In Example 2 with n = 100, p=0.1 and x=18 the value of x lies very close to a percentage point value for x on Figure 3 (page 46). Therefore, few individual terms would be required to compute this case directly. To compute this case directly it is best to sum individual terms of $b(x;n,p)$ for values of x which are greater than 18, then subtract the sum from one. Compute $b(19;n,p)$ directly, then for the remaining terms use the recurrence relation

$$b(r+1;n,p) = \frac{n-r}{r+1} \frac{p}{q} b(r;n,p)$$

$$b(19;100,0.1) = \frac{100!}{19!(100-19)!} (0.1)^{19} (0.9)^{81}$$

log 100!	= 157.97000
19 log 0.1	= 1.00000-20
81 log 0.9	= 6.29344-10
Sum	= 135.26344
-log 19!	= - 17.08509
-log 81!	= -120.76321
Sum	= -137.84830
Sum	= 7.41514-10
b(19;100,0.1)	= 0.0026010

$$b(20;n,p) = \frac{100-19}{19+1} \frac{0.1}{0.9} (0.002601)$$

$$= \frac{81}{20} \quad (0.111111)(0.002601)$$

$$= 0.001170$$

$$b(21;n,p) = \frac{80}{21} \quad (0.111111)(0.001170)$$

$$= 0.000495$$

$$b(22;n,p) = \frac{79}{22} \quad (0.111111)(0.000495)$$

$$= 0.000197$$

$$b(23;n,p) = \frac{78}{23} \quad (0.111111)(0.000197)$$

$$= 0.000074$$

Considering the rate of decrease in value of the individual terms, the remaining terms can probably be ignored. Summing and subtracting from one gives

$$1 - 0.004537 = 0.995463$$

which is within 0.00004 of the actual value of B(18;100,0.1). Only five individual terms are required which are computed quite rapidly on a desk calculator after the initial term is computed. Fortunately the logarithms of the factorials required for the initial term in this case are tabulated. If the values are too large to be found in tables or tables are not available, the values are obtained using Stirling's series approximation (2-10).

The development in this chapter shows the use of three or less terms of Wilks' asymptotic expansion in obtaining B(x;n,p) to an accuracy of three decimal places.

An algorithm based on this development and utilizing the Procedures Map and Percentage Point Chart is presented in the next chapter.

CHAPTER IV

SUMMARY AND RECOMMENDATIONS

Summary of Procedures for Obtaining the Cumulative Binomial Probability Distribution B(x;n,p) within Nominal Three Decimal Place Accuracy

The cumulative binomial

$$B(x;n,p) = \sum_{r=0}^{x} c_r^n \, p^r \, q^{n-r}$$

can be approximated within a nominal accuracy of three decimal places using one, two, or three terms of the asymptotic expansion of the distribution of a sample sum. The first term of the expansion is the cumulative normal

$$T_1(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

where

$$Z = \frac{x+\frac{1}{2}-np}{\sqrt{np(1-p)}} .$$

The second term is

$$T_2(z) = -K(z^2-1)\phi^{(1)}(z)$$

where

$$K = \frac{1}{n^{1/2}} \frac{1-2p}{6(p-p^2)^{1/2}} ,$$

and the third term is

$$T_3(z) = -(A+B)\phi^{(1)}z$$

where

$$A = \frac{1}{24n}\left(\frac{1}{p-p^2} - 6\right)(z^3-3z)$$

and

$$B = \frac{K^2}{2}(z^5-10z^3+15z) .$$

These procedures can be applied with confidence in the domain: $0 \leq n \leq 1000$, $0 \leq p \leq 1$ and $0 \leq x \leq 1000$. If a problem has $p > 0.5$, use the relation

$$B(x;n,p) = 1 - B(n-x+1;n,1-p) .$$

Enter the Percentage Point Chart in Figure 3 (page 46) to determine if the case is nontrivial; that is, if $0.001 \leq B(x;n,p) \leq 0.999$. Proceed only if the case is nontrivial.

Enter Procedures Map on Figure 2 (page 44) with values of x, n, and p. If in area A and $p < 0.007$, use

cumulative Poisson tables if available (9,pp.4,32). If Poisson tables are not available, use the two or three-term series according to the value of x. If x is too small to permit use of the three-term series, compute B(x;n,p) directly by summing individual binomial terms b(r;n,p). Use the values on the Percentage Point Chart (Figure 3, page 46) to determine whether to sum b(r;n,p) terms for values of r which are > x and subtract the sum from one or to sum b(r;n,p) terms for values of r which are $\leq$ x. Use the following recurrence relations to compute b(r;n,p) terms after computing as the initial term either b(x+1;n,p) or b(x;n,p):

$$b(r+1;n,p) = \frac{(n-r)\ p}{(r+1)\ q}\ b(r;n,p)$$

$$b(r-1;n,p) = \frac{r\ \ q}{(n-r+1)p}\ b(r;n,p)$$

Use the following relation when required

$$B(x;n,p) = 1 - \sum_{r=x+1}^{n} c_r^n p^r q^{n-r} \ .$$

If in area B on the Procedures Map, use the two-term series when x > 24; otherwise, use the three-term series.

If in area C use one term, the cumulative normal, when x is large enough. Otherwise, use the two-term

series. Only by interpolation between contour lines for minimum value of x will the use of one term be indicated in an appreciable proportion of area C. If in area D the small triangular area in the upper right of the map, use one term.

The one and two-term series can be used to cover the majority of the area not mapped in Figure 2 (page 44) which lies above the line for n=1000. The 0.001 limit for the two-term series is defined by the expression $np^{1.24} \geq 12.7$ and can be extended into the area of higher n. The two-term series obtains B(x;n,p) within three decimal place accuracy in the area to the right of this 0.001 limit (9,p.31). The one-term series obtains the B(x;n,p) to the right of the limit defined by the expression $np \approx 4000$ (9,p.33). Procedures are not prescribed for the area to the left of the 0.001 limit for the two-term series and above the n=1000 line. The x value contours on Figures 2 and 3 and the 0.001 limit of the three-term series on Figure 2 cannot be extrapolated above the n=1000 line with confidence.

Recommendations

It is recommended that the procedure developed in this thesis for approximating B(x;n,p) be made universal by extending the domain in which the procedure can be applied with confidence to include values of n exceeding 1000.

It is recommended that an attempt be made to find a more effective norming procedure to use with Wilks' asymptotic expansion. Traditional norming, using the factor of 1/2 to adjust a continuous distribution to reflect a more accurate representation of a discrete distribution, is used in this thesis for lack of something better rather than for its effectiveness.

It is further recommended that additional (n,p) points on the Procedures Map (Figure 2, page 44) be evaluated to improve the effectiveness of the procedure of this thesis for approximating B(x;n,p). The evaluation of (n,p) points is to provide minimum values of x ensuring three decimal place accuracy of the approximation of B(x;n,p) obtained with the one, two, and three-term series of Wilks' asymptotic expansion of the distribution of a sample sum. The additional values of x are the means of refining the position of the contour lines on the Procedures Map. The contour lines as shown include various amounts of safety factor ensuring three decimal place accuracy of the approximation of B(x;n,p), and thus the over-all procedure is not as efficient as it could possibly be.

REFERENCES

1. Beckenbach, Edwin F. (ed.). Modern Mathematics for the Engineer. New York: McGraw-Hill Book Company, Inc., 1961.

2. Bowker, Albert H. and Gerald J. Lieberman. Engineering Statistics. New Jersey: Prentice-Hall, Inc. 1959.

3. Feller, W. "On the Normal Approximation to the Binomial Distribution," The Annals of Mathematical Statistics. Vol. XVI, No. 4, December, 1945.

4. Fisher, Arne. The Mathematical Theory of Probabilities. New York: The Macmillan Company, 1926.

5. Harvard Computation Laboratory. Tables of the Cumulative Binomial Probability Distribution. Massachusetts: Harvard University Press (1955).

6. Heineman, E. Richard. College Algebra. New York: The Macmillan Company, 1947.

7. Hoel, Paul G. Introduction to Mathematical Statistics. New York: John Wiley & Sons, Inc., 1962.

8. Kendall, Maurice G. The Advanced Theory of Statics. London: Charles Griffin & Company Limited, 1947.

9. Smith, Ed Sinclair. Binomial, Normal and Poisson Probabilities. Bel Air, Maryland: Ed Sinclair Smith, 1953.

10. Spiegel, Murray R. Theory and Problems of Statistics. New York: Schaum Publishing Company, 1961.

11. Stanton, Ralph G. Numerical Methods for Science and Engineering. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1961.

12. Uspensky, J. V. Introduction to Mathematical Probability. New York: McGraw-Hill, 1937.

13. Wilks, Samuel S. Mathematical Statistics. New York: John Wiley and Sons, Inc., 1962.

APPENDIX

COMPUTER PROGRAM FOR EVALUATING (n,p) POINTS

INTRODUCTION

This is a FORTRAN II program written to be used on the IBM 1620 Electronic Computer Mod II. The program is developed specifically for the requirements of the investigation accomplished for this thesis.

PURPOSE

The purpose of this program is to determine the minimum values of x for a specific (n,p) point which ensure accuracy within 0.001 and within 0.002 of the approximation of B(x;n,p) obtained with the one, two, and three-term series of Wilks' asymptotic expansion of the distribution of a sample sum.

INPUT DATA

The evaluation of each (n,p) point requires a data card in the following format.

Columns	
1 - 5	The value of n with decimal point.
6 - 10	The value of p with decimal point.
11 - 15	The value of x without decimal point which is the maximum value of x which obtains $B_r(x;n,p) \geq 0.999$. This value is obtained from cumulative binomial

tables. $B_r(x;n,p)$, which is equal to 1- $B(x;n,p)$, is the form of the cumulative binomial tabulated in reference five.

16 - 20 The value of x without decimal point which is the minimum value of x which obtains $B_r(x;n,p) \leq 0.001$.

26 - 35 The initial value of $B_r(x;n,p) = B_o(x;n,p)$ from cumulative binomial tables for x equal to one plus the value in columns 16-20.

Examples of the data form of three input cards follow.

```
1000..1      73  131          .00069
100. .11      4   22          .00034
20.  .2       1   11          .00010
```

OUTPUT

The output for the sample input follows.

```
N=  1000  P=  .10   X     ERR    X     ERR       Z         B
N=   100  P=  .11   X     ERR    X     ERR       Z         B
                   17.5  .001                 2.0774   .024220
N=    40  P=  .20   X     ERR    X     ERR       Z         B
                   12.5  .001                 1.7787   .043237
                                3.5   .002   -1.7787   .971536
```

The output for n=1000, p=0.1 shows no values under the x's in the heading. This indicates that the series being evaluated obtains an approximation of $B(x;n,p)$ within an accuracy of 0.0001 for all values of x which obtain $0.001 \leq B(x;n,p) \leq 0.999$.

The output for n=100, p=0.11 shows x=17.5, error= 0.001, Z=2.0074, and B=0.024220. The x value is the maximum value of x for which the series approximation differs

from $B(x;n,p)$ by more than 0.001. Thus, $x+1=18.5$ is the minimum value of x for which the series approximation differs from $B(x;n,p)$ by less than 0.001. With the norming factor of 1/2 removed the minimum value of x becomes 18. The Z output is the value for $\frac{17.5-np}{\sqrt{np(1-p)}}$. The B output is value of $B_r(18;n,p)$ computed using a recurrence relation repeatedly with initial value of $B_o(x;n,p)$ from the input data. The B output is used to check the accuracy of the recurrence relation computations. No value is shown under the second x in the heading. This indicates that the approximation of $B(x;n,p)$ is within 0.002 for all values of x which obtain $0.001 \leq B(x;n,p) \leq 0.999$.

The output for $n=40$, $p=0.20$ is an example showing minimum values of x for accuracies of 0.001 and 0.002. Interpretation of values are similar to the above example.

SPECIAL OPERATING PROCEDURES

Card 9999 in the computer program is different for evaluation using one, two, and three-term series. The correct card must be inserted for each series.

The accuracy limits in the evaluation can easily be changed by changing the values on cards 5000 and 5001. The value for the more stringent accuracy limit must be on card 5000.

COMPUTER PROGRAM

```
    READ:  n;p; x VALUE FOR 0.1 AND 99.9 PERCENTAGE POINT
                  OF B(x;n,p); AND B(JJ+1;n,p)
   DIMENSION D(30)
 1 READ 2, AN, P, II, JJ, BT
 2 FORMAT ( 2F5.2, 2I5,5X,F10.5 )

          COMPUTE & STORE 0! THROUGH 30! FOR USE
             IN COMPUTING CUMULATIVE NORMAL

   D(1)=1.
   D(2)=1.
   DO 12 N=3, 30
   A=N
12 D(N)=D(N-1) * (A-1.)

                PUNCH n,p & COLUMN HEADINGS

   NN=AN
   PUNCH 4,NN,P
 4 FORMAT(2HN=I6,5X2HP=F5.3,5X1HX5X3HERR7X1HX5X3HERR12X1HZ10X1HB)
   PUNCH 3
 3 FORMAT(17X1HX5X5HE-LIM5X1HX5X5HE-LIM10X1HZ10X1HB )
     KKKK=0
     KKK=-1
5000 DM=.001
     DO 30 I=II, JJ
     KKK=KKK+1
     AI=JJ-KKK
     Y=AI +1.

                    COMPUTE b(JJ;n,p) = BI

     IF(KKK)1000,1000,5
1000 AAA=AN
     III=-2
1001 III=III+1
     AAA=AAA+1.
     FA=      1.+1. /(12. *AAA)+1. /(288. * AAA**2)
     FA=FA-139. /(51840. * AAA**3)-571. /(2488320. * AAA**4 )
     FA=LOGF(FA)
     FA=FA+(AAA   )*LOGF((AAA   )/2.7182818284)
     FA=FA+.5*LOGF(6.2831853072 /  AAA   )
     IF(III) 1005, 1010, 1020
1005 FN=FA
     AAA=Y
     GO TO 1001
1010 FY=FA
```

```
      AAA=(AN-Y)
      GOTO 1001
1020  FD=FA
      BI=FN-FD-FY+  Y*LOGF(P) + (AN-Y)*LOGF(1.-P)
      BI=EXPF(BI)

                    COMPUTE RECURRENCE BI,
                    b(r-1;n,p) = r/(n-r+1) b(r;n,p)

    5 AK=AI+1.
      BI=BI*AK* (1.-P) / (AN-AK+1.) / P

                    COMPUTE RECURRENCE BT
              BT = B(x;n,p) = B(r+1;n,p) + b(r;n,p)
      BT=BT+BI

                COMPUTE SERIES ENTERING ARGUMENT

      Y=AI-.5
    6 FORMAT (F7.1)
      X=Y
      X=(X-AN*P)/((AN*P*( 1.-P)) **.5 )

                    COMPUTE THIRD TERM OF SERIES

      XX=X
      H=P
      V= -((1.-3.*P+3.*P*P) / (P-P*P) - 3. ) *X*(X*X-3.) / 24.
      V=V-(1.-2.*P)**2*( (X*X-10.)*X*X+15. )*X / (72.*(P-P*P) )
      V=V /( 2.71828183 ** (.5*X*X) * 6.2831854 ** .5 *AN)

                    COMPUTE SECOND TERM OF SERIES

      U=-(1.-2.*P) * (X*X-1.) / (6. * (P-P*P) ** .5)
      U=U / (2.71828183 ** (.5*X*X) * 6.2831854 ** .5 *AN**.5  )

           FOR ONE-TERM SERIES THIS LINE IS 9999 T=0.0
           FOR TWO-TERM SERIES THIS LINE IS 9999 T=U
           FOR THREE-TERM SERIES THIS LINE IS 9999 T=U+V

9999  T=U+V

             COMPUTE CUMULATIVE NORMAL PN FROM 0 TO Z
             FOR ABSOLUTE VALUE OF STANDARD DEVIATE Z

      IF (ABSF(X) - 4.2 ) 20, 20, 18
   20 F=0.39894228
      X=ABSF (X)
      PN=0.0
      N=0
```

```
16 N=N+1
   A=N
   K=N-1
   J=2*N-1
   G=F*X**J  /  (  (2.*A-1.) * 2.**K )
   G=G/D(N)
   J=N+1
   S=(-1) **J
   IF (G-0.00001) 19,17,17
17 PN=PN+S*G
   GO TO 16
18 PN=.5
19 CONTINUE
```

SUM CUMULATIVE NORMAL FIRST TERM TO OTHER TERMS

```
   IF(XX) 21,21,22
21 T=T+.5-PN
   T=1.-T
   GO TO 23
22 T=T+.5+PN
   T=1.-T
```

OPERATOR MONITOR SWITCH

```
  23 CONTINUE
     IF(SENSE SWITCH 2)2000,2001
2000 PRINT 2002,XX,T
2002 FORMAT (2F15.5)
```

COMPUTE B(x;n,p) MINUS SERIES APPROXIMATION AND PUNCH ANSWER IF DIFFERENCE > 0.001 OTHERWISE CONTINUE ITERATIONS

```
2001 CONTINUE
     DIF=BT-T
     IF(KKKK)99,99,110
  99 IF(DM-.001) 100,100,110
 100 IF(ABSF(BT-T) -DM) 30, 30, 101
 101 PUNCH 102, Y,DM,XX,BT
 102 FORMAT(21X,F7.1,F7.3,19X,F12.4,F12.6)
5001 DM=.002
     KKKK=KKKK+1
```

PUNCH ANSWER IF DIFFERENCE > 0.002 OTHERWISE CONTINUE ITERATIONS

```
 110 IF(ABSF(BT-T) -DM) 30,30,111
 111 PUNCH 112,Y,DM,XX,BT
 112 FORMAT(34X,F10.1,F7.3,F15.4,F12.6)
     GO TO 31
```

```
30 CONTINUE
31 CONTINUE
   PUNCH 32
32 FORMAT (/ )
   IF (SENSE SWITCH 1) 33, 1
33 CONTINUE
   END
```