APPROXIMATIONS TO THE BINOMIAL

by

MYRTLE ANNA BRUCE

B. A., Carson Newman College, Jefferson City, Tennessee, 1965

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

W J Conover
Major Professor

TABLE OF CONTENTS

1. INTRODUCTION

The binomial distribution function,

$$B(s;n,p) = \sum_{i=0}^{s} \frac{n!}{i!(n-i)!} p^{i} q^{n-i} \qquad s \geqslant 0$$

is difficult to tabulate directly because the tabulation involves three variables: p, n, and s. Abraham DeMoivre (1730) was the first to consider the problem of approximating the binomial probabilities using the normal distribution. The Laplace-DeMoivre result involves term by term normal approximation to the binomial probabilities. A proof for the DeMoivre-Laplace Limit is given by Feller (1950).

Today there are many different approximations for the binomial, but the accompanying problem appears to be that of determining the error involved when each approximation is used. Five different approximations are presented in this paper. Peizer and Pratt (1966) developed a new normal approximation, z_1 or the more refined z_2, which appears to give a smaller error than Camp's (1951) cube root, Pinkham's (1957) square root, or the usual normal approximation. The Poisson Gram-Charlier approximation performs best if it is used only if $n \geqslant 5$ and $np \leqslant 0.8$ (Raff, 1956). Finally Govindarajulu's (1963) approximation can be made as accurate as one desires because the error term can be made as small as desired.

Let P be a binomial tail probability. Then define $\emptyset$ to be the standard normal left or right tail function according as P is a left or right tail probability. A value z is sought such that

$$\emptyset_1(z) = \int_{-\infty}^{z} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} dx$$

is approximately equal to P if P is a left tail probability

or

$$\phi_2(z) = \int_z^{\infty} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} dx$$

is approximately equal to P if P is a right tail probability.

Let the parameters of the binomial be n and p. Let s be the number of successes. Then, if P is a right tail probability, let

$$S = s - .5$$

and

$$T = n - S$$

The following example will be given at the end of each section with the exception of the section on Govindarajulu's approximation, where the calculations are difficult and the error can be made as small as one wishes by continuing the expansion to as many terms as needed.

What is the approximate probability for 4 or more successes under the binomial distribution with parameters $n = 5$ and $p = .25$?

The probability can be computed exactly by evaluating the following terms of the binomial expansion:

$$\binom{5}{4} (.25)^4 (.75)^1 + \binom{5}{5} (.25)^5 (.75)^0 = .015625$$

Before presenting the five approximations considered in this paper, the example is computed using $z_o' = \frac{s - np}{(npq)^{\frac{1}{2}}}$ and $z_o = \frac{S - np}{(npq)^{\frac{1}{2}}}$. The reason for looking at these two ordinary approximations is to show how they compare with the approximations presented in this paper.

The case for $z_o' = \frac{s - np}{(npq)^{\frac{1}{2}}}$ results in $z_o' = \frac{4 - 1.25}{\sqrt{.9375}} = \frac{2.75}{.9682} = 2.8403$

because $np = 5(.25) = 1.25$ and $npq = 5(.25)(.75) = .9375$.

The desired approximation to the right tail probability is

$$\emptyset_2(z_o') = .0023 .$$

The second approximation $z_o = \frac{S - np}{(npq)^{\frac{1}{2}}}$ involves replacing s with S where $S = s - .5$ or $S = 4 - .5 = 3.5$. Thus

$$z_o = \frac{S - np}{(npq)^{\frac{1}{2}}} = \frac{3.50 - 1.25}{\sqrt{.9375}} = \frac{2.25}{.9682} = 2.3239 .$$

The desired approximation to the right tail probability is

$$\emptyset_2(z_o) = .0101 .$$

2. PEIZER AND PRATT APPROXIMATION

In order to find the probability P, the function $\emptyset_1(z)$, or $\emptyset_2(z)$, may be used; but then the proper value of z still remains undetermined. The unknown value of z, which depends on the values of n, p, and s, may be estimated several different ways. The following method presents z_1 and z_2, two estimates of the unknown z. The quantities z_o, z_3, and z_4 are introduced to show the different stages in the development of z_1 and z_2.

Peizer and Pratt (1966) considered

$$(2.1) \qquad z_i = \frac{d_i}{S-np}\left\{\frac{2}{1 + (6n)^{-1}}\left(S \ln \frac{S}{np} + T \ln \frac{T}{nq}\right)\right\}^{\frac{1}{2}} ; \qquad i = 1, 2,$$

or for computational purposes,

$$(2.2) \qquad z_i = d_i \left\{ \frac{1 + q\,g\,(\frac{S}{np}) + p\,g\,(\frac{T}{nq})}{(n + \frac{1}{6})\,pq} \right\}^{\frac{1}{2}} ,$$

as their approximation for the unknown quantity z, where d_i has the value

$$d_1 = S + \frac{1}{6} - (n + \frac{1}{3})\,p ,$$

or

$$d_2 = d_1 + .02\,(\frac{q}{S + .5} - \frac{p}{T + .5} + \frac{q - .5}{n + 1}) ,$$

and g can be obtained either from a table or by direct calculation as explained later.

They started their search for a normal approximation to the binomial probabilities by looking at $z_o = (S - np)/\sigma$ where $\sigma^2 = npq$, and stated that one reason why z_o is inaccurate is that it misplaces the median. The median and the mean in a normal distribution are the same, but this is not the case in the binomial distribution unless $p = \frac{1}{2}$. Therefore using z_o, one estimates the mean accurately, but not the median. After studying the median numerically and asympototically, they found the following modification:

$$(2.3) \qquad z_3 = d_1/\sigma = [S - np + (q - p)/6]/\sigma$$

The modification, z_3, gives better accuracy than z_o near $z = 0$. However, z_3 is not necessarily as accurate as z_o when the value of z is not near zero.

The next step is to try to find an approximation for z that gives greater accuracy than z_o in the tails of the distribution as well as around

$z = 0$. To accomplish this, they chose to observe the graph of z/z_3. This relationship appears to be approximately

(2.4) $$\frac{z}{z_3} \cong f(p,D)$$

when $D = (S - np)/\sigma^2 = \frac{z_o}{\sigma}$ and f is a gently varying function. Other arguments could be used, for example, p and S/np. However, D shows the symmetry aspect of the situation better.

In order to tabulate f, and thus use $z_3 f$ as an approximation to z, it is desirable to make z/z_3 more nearly a function of p and D alone rather than to make z_3 closer to z. An asymptotic calculation of the central term or terms of the binomial distributions leads to replacing n by $n + \frac{1}{6}$ in the denominator of z_3. The same result could be found by examining the graph of z/z_3 for $p = \frac{1}{2}$ (Pratt, 1966).

Since $$z_1 = d_1 \left\{ \frac{1 + q\, g\, (\frac{S}{np}) + p\, g\, (\frac{T}{nq})}{(n + \frac{1}{6})\, pq} \right\}^{\frac{1}{2}},$$

it can be seen that z_1 can be expressed as

(2.5) $$z_4 f(p,D) \quad \text{where} \quad z_4 = d_1 \left[(n + \tfrac{1}{6})\, pq\right]^{-\frac{1}{2}} \quad \text{and}$$

$$f = \left\{1 + q\, g\, (\tfrac{S}{np}) + p\, g\, (\tfrac{T}{nq})\right\}^{\frac{1}{2}}$$

By substituting the value of d_1 given in (2.3), z_4 may be written as

$(S + c_1 - np)/(n + c_2)\, pq$

for $c_1 = (q - p)/6$ and $c_2 = \frac{1}{6}$. Other choices for c_1 and c_2 lead to no better approximation (Pratt, 1966).

To make the approximation accurate in the extreme tails, define f to be the limit of z/z_4 as $\sigma \rightarrow \infty$ for p and D fixed. Thus one lets

$n \to \infty$ and looks at the limit as the number of success either gets larger or smaller for fixed p. These results can be obtained from results of Bahadur (1960). This makes the relative error in the approximation to z by $z_1 = z_4 f$ approach 0 as $\sigma \to \infty$ for fixed $D \neq 0$. Using this, f can be expressed in terms of familiar functions, in fact, in terms of a function of one variable;

$$(2.6) \qquad f = G^{\frac{1}{2}} = \left[(1 + q\, g\, (\frac{S}{np}) + p\, g\, (\frac{T}{nq}) \right]^{\frac{1}{2}}$$

where the function g is defined as

$$g(x) = \frac{1 - x^2 + 2x \ln x}{(1 - x)^2}, \quad x > 0, \quad x \neq 1$$

$$g(0) = 1, \quad g(1) = 0,$$

$$g(x) = -g(x^{-1})$$

Peizer and Pratt (1966) have tabulated the function $g(x)$.

The function G can be expressed with arguments p and D, which is the form specified, by using the relationship $D = \frac{S - np}{npq}$.

The relationship is solved for $\frac{S}{np}$ in the following manner:

$$D = \frac{S - np}{npq} = \frac{S}{npq} - \frac{np}{npq} = \frac{S}{npq} - \frac{1}{q}$$

or

$$qD = \frac{S}{np} - 1 \; ; \; \frac{S}{np} = 1 + qD \, .$$

By using the same technique, one obtains $\frac{T}{nq}$:

$$D = \frac{S - np}{npq} = \frac{S}{npq} - \frac{1}{q} = \frac{n - T}{npq} - \frac{1}{q} = \frac{1}{pq} - \frac{T}{npq} - \frac{1}{q}$$

or

$$\frac{T}{npq} = \frac{1}{pq} - \frac{1}{q} - D$$

$$\frac{T}{nq} = \frac{1}{q} - \frac{p}{q} - pD = \frac{(1 - p)}{q} - pD = 1 - pD .$$

When, $1 + qD$ is substituted for $\frac{S}{np}$ and $1 - pD$, for $\frac{T}{nq}$,

$$G = 1 + p\, g(1 - pD) + q\, g(1 + qD) .$$

In order to give greater accuracy near the median, Peizer and Pratt introduced z_2, a modification of their z_1.

The relationship between z_1 and z_2 can be seen below.

Let $z_i = (d_1 + c_i) \left\{ \frac{G}{(n + \frac{1}{6})pq} \right\}^{\frac{1}{2}}$ where $c_i = 0(\sigma^{-2})$.

For $i = 1$, let $c_1 = 0$.

For $i = 2$, let $c_2 = .02 \left(\frac{q}{S + .5} - \frac{p}{T + .5} + \frac{q - .5}{n + 1}\right)$.

The example presented in the introduction is calculated for z_1 and z_2 as follows:

Peizer and Pratt's approximation z_1 equals $d_1 \left\{ \frac{1 + q\, g(\frac{S}{np}) + p\, q(\frac{T}{np})}{(n + \frac{1}{6})\, pq} \right\}^{\frac{1}{2}}$

where $T = n - S$ or $5 - 3.5 = 1.5$, $\frac{S}{np} = \frac{3.5}{5(.25)} = 2.8$, $\frac{T}{nq} = \frac{1.5}{5(.75)} = .4$,

$\frac{S}{np} / \frac{T}{nq} = \frac{2.8}{.4} = 7$, $d_1 = S + \frac{1}{6} - (n + \frac{1}{3})p = 3.5 + \frac{1}{6} - (5 + \frac{1}{3})\,(.25) = 2.3333$,

and g is obtained from Peizer and Pratt's (1966) tabulation of the function g where $g(2.8) = -g(\frac{1}{2.8}) = -.3315$ and $g(.4) = .2971$.

These quantities substituted in (2.2) give

$$z_1 = 2.3333 \left\{ \frac{1 - (.75)(.3315) + .25(.2971)}{(5 + \frac{1}{6})(.25)(.75)} \right\}^{\frac{1}{2}}$$

$$= 2.3333 \left\{ \frac{.82565}{.96875} \right\}^{\frac{1}{2}} = 2.3333(.92319) = 2.1541 .$$

The desired approximation to the right tail probability is

$$\emptyset_2(z_1) = .01562 .$$

Peizer and Pratt's approximation z_2 equals $d_2 \left\{ \frac{1 + q\, g(\frac{S}{np}) + p\, g(\frac{T}{nq})}{(n + \frac{1}{6})pq} \right\}^{\frac{1}{2}}$ is a refinement of their z_1 where $d_2 = d_1 + .02 \left\{ \frac{q}{S + .5} - \frac{p}{T + .5} + \frac{q - .5}{n + 1} \right\}$

$$= 2.3333 + .02 \left(\frac{.75}{3.5 + .5} - \frac{.25}{1.5 + .5} + \frac{.25}{5 + 1} \right) = 2.3333 + .0021 = 2.3354 .$$

Thus $z_2 = 2.3354\ (.92319) = 2.1560$.

The desired approximation to the right tail probability is

$$\emptyset_2(z_2) = .01554 .$$

In this example z_1 was better than z_2, but usually z_2 is better.

Peizer and Pratt (1966) have graphed the relative error

$$(2.7) \qquad \epsilon = \pm \frac{P - \emptyset(z_2)}{\min\left\{\emptyset(z_2),\ 1 - \emptyset(z_2)\right\}}$$

of $\emptyset_2(z_2)$ against Sq/Tp for many values of S, n, and p.

By solving (2.7) for P,

$$P = \emptyset(z_2) \pm \epsilon \min\left\{\emptyset(z_2),\ 1 - \emptyset(z_2)\right\} ,$$

the approximation $\emptyset_2(z_2)$ can be corrected by using the value of ϵ found in the appropriate graph.

For z_2 in the example, the relative error, ϵ , is -.005. Then the corrected probability is

$$P = .01554 - (-.005)(.01554) = .01526 .$$

3. CUBE ROOT APPROXIMATION

Camp (1951) gives the following approximation to the sum of the first (t + 1) terms of the binomial probability function where t equals the number of failures: Let the binomial probability function be generated by the terms of $(p+q)^n$. Let the sum $= S_{t+1} = p^n + np^{n-1}q + \ldots\ldots + \binom{n}{t}p^{n-t}q^t$. Then S_{t+1} is approximately equal to the probability that a unit normal deviate will exceed $z'_{1/3}$ where

$$(3.1) \qquad z'_{1/3} = \frac{\frac{1}{3}\left[\left(\frac{9s-1}{s}\right)\left(\frac{s}{t+1}\,\frac{q}{p}\right)^{1/3} - \frac{9t+8}{t+1}\right]}{\left[\frac{1}{s}\left(\frac{s}{t+1}\,\frac{q}{p}\right)^{2/3} + \frac{1}{t+1}\right]^{1/2}} , \qquad s = n - t$$

In an attempt to standardize the notation, Camp's approximation is re-written in terms of the notation used by Peizer and Pratt.

Let $s = b = S + \frac{1}{2}$ and $t + 1 = a = T + \frac{1}{2}$.

Then

$$(3.2) \qquad z'_{1/3} = \frac{\left[\left(9 - \frac{1}{b}\right)\left(\frac{bq}{ap}\right)^{1/3} - 9 + \frac{1}{a}\right]}{3\left[\frac{1}{b}\left(\frac{bq}{ap}\right)^{2/3} + \frac{1}{a}\right]^{1/2}} .$$

Camp (1951) gives the following example which shows errors of approximation for the special case; $n = 8$, $p = 0.80$:

$(.8 + .2)^8$

t	$S_t + 1$ Approx.	True	Error
0	.166	.168	- .002
1	.505	.503	.002
2	.801	.797	.004
3	.943	.944	- .001
5	.999	.999	.000

While an error of .005 appears to be small, when P is actually .01, the relative error would be 50 per cent.

The cube root approximation (Pratt, 1966) belongs to the family of approximations coming from root transformations of chi square random variables. It will be shown that one can go from the chi square to the F distribution, to the incomplete beta, and hence to the binomial distribution. Let X^2 have a chi square distribution with v degrees of freedom. Then treat $(X^2/v)^m$ $(m > 0)$ as normally distributed by the Central Limit Theorem. That is

$$(3.3) \qquad \frac{(X^2/v)^m - E[(X^2/v)^m]}{\sqrt{\mathrm{Var}\ (X^2/v)^m}} \sim N(0,\ 1)\ .$$

One can find the mean and variance plus the third central moment and the median since Wilson and Hilferty (1931) have computed these quantities for $(X^2/v)^m$.

$$(3.4) \qquad \text{Median} \;=\; 1 - \frac{2m}{3v} + \frac{2m}{9v^2}\left(m - \frac{29}{45}\right) + \ldots$$

$$(3.5) \qquad E(X^2/v)^m \;=\; 1 + \frac{m}{v}(m - 1) + \frac{4m}{3v^2}(m - 1)(m - 2) + \ldots$$

(3.6) $$\text{Var}\ (X^2/v)^m = \frac{2m}{v} + \frac{2m^2}{v^2}(m-1)(3m-1) + \ldots$$

(3.7) $$\text{Third central moment of}\ (X^2/v)^m = \frac{4}{v^2}(3m-1) + \ldots$$

Then one can estimate the probability of the chi square random variable in the following way:

(3.8) $$\alpha \cong P\left(\frac{(X^2/v)^m - E[(X^2/v)^m]}{\sqrt{\text{Var}\ (X^2/v)^m}} \leq z_\alpha\right)$$

$$\alpha \cong P\left[(X^2/v)^m \leq y'_\alpha = E[(X^2/v)^m] + \sqrt{\text{Var}\ (X^2/v)^m}\ z_\alpha\right]$$

Since the mean and variance can be found, one can solve for z_α and then use the normal probability tables to obtain an approximate value for α .

Because the F distribution is a ratio of two independent chi squares divided by their degrees of freedom, one can extend the previous discussion to obtain a normal approximation for the F distribution.

Let X_1^2 and X_2^2 be two independent chi square random variables. Then

$$P\left[\frac{X_1^2}{v_1}\bigg/\frac{X_2^2}{v_2} \leq w\right] = P\left[\left(\frac{X_1^2/v_1}{X_2^2/v_2}\right)^m \leq w^m\right]$$

$$= P\left[\left(\frac{X_1^2}{v_1}\right)^m - w^m\left(\frac{X_2^2}{v_2}\right)^m \leq 0\right], \qquad w \geqslant 0$$

The first quality follows from the fact that chi square has only positive values.

The quantity $(X_1^2/v_1)^m - w^m\ (X_2^2/v_2)^m$ approaches normality because a linear combination of independent normal variables is normal with mean of the difference between the means of $(X_1^2/v_1)^m$ and $(X_2^2/v_2)^m$ and variance of the sum of the variances of the previous quantities. Thus the normal approximation can be applied, leading to a family of normal approximations for the F distribution.

Wilks (1962) gives a transformation for going to an incomplete beta from the F distribution. Therefore one can obtain a normal approximation for the beta distribution. Feller (1950) shows that the binomial can be obtained from the incomplete beta. In like manner, one can obtain a normal approximation for the binomial.

Since one wants $(X^2/v)^m$ to approach normality as fast as possible, one chooses a value of m that will make the distribution as symmetric as possible. Wilson and Hilferty (1931) picked $m = \frac{1}{3}$. For this value the leading term of the third central moment, or the measure of skewness, drops out. Then $(X^2/v)^{1/3}$ is treated as a normal variate with mean $(1 - 2/9v)$ and variance of $2/9v$, using only the first two terms of the expansion of the mean and variance. The median and the mean agree through the terms of order $\frac{1}{v}$ when $m = \frac{1}{3}$. This is only true when m has this value.

Paulson (1942) used the method previously described in obtaining a cube root approximation for the F distribution from the Wilson-Hilferty cube root approximation. Camp (1951) extended Paulson's work to obtain an approximation to the incomplete beta. Thus he was able to arrive at a normal approximation to the binomial distribution.

The example presented in the introduction is calculated using Camp's cube root approximation $z'_{1/3}$.

$$z'_{1/3} = \frac{\left[(9 - \frac{1}{b})(\frac{bq}{ap})^{1/3} - 9 + \frac{1}{a}\right]}{3\left[\frac{1}{b}(\frac{bq}{ap})^{2/3} + \frac{1}{a}\right]^{1/2}} = \frac{\left[(9 - \frac{1}{4})(\frac{4(.75)}{2(.25)})^{1/3} - 9 + \frac{1}{2}\right]}{3\left[\frac{1}{4}(\frac{4(.75)}{2(.25)})^{2/3} + \frac{1}{2}\right]^{1/2}}$$

$$= \frac{\frac{35}{4}(6)^{1/3} - \frac{17}{2}}{3\left[\frac{1}{4}(6)^{2/3} + \frac{1}{2}\right]^{1/2}} = \frac{(8.75)(1.82) - 8.5}{3\left[\frac{1}{4}(3.31) + \frac{1}{2}\right]^{1/2}} = 2.1484$$

because $b = S + .5 = 3.5 + .5 = 4$ and $a = T + .5 = 1.5 + .5 = 2$.

The desired approximation to the right tail probability is

$$\emptyset_2(z'_{1/3}) = .0159 .$$

4. SQUARE ROOT APPROXIMATION

The unknown value z can be estimated by using the square root approximation $z'_{\frac{1}{2}}$ suggested by Pinkham (1957), where

$$(4.1) \qquad z'_{\frac{1}{2}} = [(4b - 1)q]^{\frac{1}{2}} - [(4a - 1)p]^{\frac{1}{2}}$$

and as in the previous section,

$$a = T + \tfrac{1}{2}$$

and

$$b = S + \tfrac{1}{2} .$$

The same argument is used here as was used in the cube root approximation.

Choose $m = \frac{1}{2}$. Then treat $(X^2)^{\frac{1}{2}}$ as a normally distributed variate with mean $v^{\frac{1}{2}}$ and variance $\frac{1}{4}$. This leads to Freeman and Tukey's (1949, 1950) approximation;

(4.2) $z''_{\frac{1}{2}} = 2[(bq)^{\frac{1}{2}} - (ap)^{\frac{1}{2}}]$

Pinkham (1957) studied the first four moments of a transformed beta variable and decided to modify $z''_{\frac{1}{2}}$ with Fisher's modification of the mean leading to $z'_{\frac{1}{2}}$.

Although $z'_{\frac{1}{2}}$ is not as accurate as $z'_{1/3}$, it is useful for variance stabilization according to Pratt (1966). The variance $(X^2)^m$ is asymptotically constant only for $m = \frac{1}{2}$.

The example using Pinkham's square root approximation $z'_{\frac{1}{2}} = [(4b - 1)q]^{\frac{1}{2}} - [(4a - 1)p]^{\frac{1}{2}}$ results in $[(4(4) - 1).75]^{\frac{1}{2}} - [(4(2) - 1).25]^{\frac{1}{2}}$
$= [(15)\ (.75)]^{\frac{1}{2}} - [7(.25)]^{\frac{1}{2}} = [11.25]^{\frac{1}{2}} - [1.75]^{\frac{1}{2}}$
$= 3.3541 - 1.3229 = 2.0312$
The desired approximation to the right tail probability is

$$\emptyset_2(z'_{\frac{1}{2}}) = .0211 .$$

5. THE ABSOLUTE AND RELATIVE ERROR OF z_o, z_1, z_2, $z'_{1/3}$, $z'_{1/2}$

Peizer and Pratt (1966) have graphed the ratio of $\frac{\text{true left tail}}{\text{approx. left tail}} \times \frac{\text{approx. right tail}}{\text{true right tail}}$ against $\frac{Sq}{Tp}$ for the approximations previously mentioned with the exception of z'_o. The ratio $\frac{\text{true left tail}}{\text{approx. left tail}} \times \frac{\text{approx. right tail}}{\text{true right tail}}$ may be expressed as the rate of true and approximate P/1-P where P + 1-P equals the total area. These graphs appear to show that z_2 is the best approximation followed by z_1, the cube root, the square root, and last, the ordinary normal approximation. At a fixed value of Sq/Tp, other than one, with S increasing, z_1

and z_2 are the only approximations that appear to remain relatively accurate.

Pratt (1966) confirmed the conclusion drawn from the graphs in the previous paragraph by asymptotic theory.

He considered the case where $\sigma^2 \to \infty$, z is bounded, and p, S, and T are varying. Then z is not equal to $-\infty$ or $+\infty$, i.e., P is bounded away from 0 and 1. Then the order of magnitude of error in P for the different approximations is as follows:

$$z_1 \text{ or } z_2 \text{ ---------- } \frac{1}{\sigma^3}$$

$$z'_{1/3} \text{ ---------- } \frac{1}{\sigma^2}$$

$$z_o \text{ or } z'_m, m \neq 1/3 \text{ ---- } \frac{1}{\sigma}$$

For z_1 this means that $P - \phi(z) = O(1/\sigma^3)$ and only in special cases is this error of smaller order. For a review of the notation $o(\cdot)$ and $O(\cdot)$ see Lukacs (1960). The exception in the binomial case is when $p = \frac{1}{2}$ and then the error is $\frac{1}{\sigma^4}$, $\frac{1}{\sigma^2}$, $\frac{1}{\sigma^2}$ respectively.

Now consider $\sigma^2 \to \infty$ and $|z| \to \infty$ as p, S, and T vary. Now $|z| \to \infty$ but $z = o(\sigma)$, i.e. P approaches 0 or 1 and $D = z_o/\sigma = \frac{S - np}{npq} \to 0$. Then the order of magnitude of error in P is

$$z_1 \text{ or } z_2 \text{ ---------- } \frac{1}{\sigma} D^2$$

$$z'_{1/3} \text{ ---------- } zD^2$$

$$z_o \text{ or } z'_m, m \neq 1/3 \text{ ---- } zD\ .$$

In each approximation the error in z is of smaller order than z itself, but only for z_1 or z_2 does the error always approach 0 as $\sigma \rightarrow \infty$.

The relative error is defined as either (true - approx.) / true, or (true - approx.) / approx. since the error in z is of smaller order than z itself and therefore of the same order whichever is used.

Order of magnitude of error relative in z is as follows:

$$z_1 \text{ or } z_2 \text{ ---------- } \frac{1}{\sigma^2} D^2$$

$$z'_{1/3} \text{ ---------- } D^2$$

$$z_o \text{ or } z'_m, m \neq 1/3 \text{ ---- } D$$

Next look at the tail probability $P = \emptyset(z)$ where P is close to 0 or 1 since $|z| \rightarrow \infty$. The ln (min P, 1-P) is considered because as $P \rightarrow 0$, the error in P is hard to interpret except as relative to the true or approximate P. However, these errors are of different orders of magnitude. The error when considered as error in ln P avoids this problem. For $|z| \rightarrow \infty$, one uses error in ln P when P is near 0 and error in ln (1 - P) when P is near 1. The order of magnitude of errors in ln (min [P, 1 - P]) for the different approximations is as follows:

$$z_1 \text{ or } z_2 \text{ ---------- } D^3$$

$$z'_{1/3} \text{ ---------- } z^2 D^2$$

$$z_o \text{ or } z'_m, m \neq 1/3 \text{ ---- } z^2 D$$

The relative error in min (P, 1-P) will approach 0 if

$$z = o(\sigma) \text{ when } z_1 \text{ or } z_2 \text{ is used.}$$

$$z = o(\sigma^{1/2}) \text{ when } z'_{1/3} \text{ is used.}$$

$$z = o(\sigma^{1/3}) \text{ when } z_o \text{ or } z'_m, m \neq 1/3 \text{ is used.}$$

Both z_1 and z_2 are nearer to standard normal than z_o in the extreme tails. They are a strictly decreasing function of $p = 1-q$ and approach $+\infty$ at $p = 0$ and $-\infty$ at $p = 1$ for fixed S and T.

The root transformations are strictly decreasing, but do not approach $+\infty$ at $p = 0$ and $-\infty$ at $p = 1$. Instead, they have the range

$$-\frac{\mu_T}{\sigma_T} < z'_m < \frac{\mu_S}{\sigma_S}$$

provided μ_S, μ_T, σ_S^2, σ_T^2 are positive (Pratt, 1966).

6. POISSON GRAM-CHARLIER APPROXIMATION

The Poisson Gram-Charlier approximation is based on the fact that the binomial distribution can be approximated by the Poisson distribution when p is small, and is defined by Raff (1956) as follows:

$$B(s;n,p) = P(s, np) + \tfrac{1}{2}p(s - np)\Delta P(s, np)$$

where s has integral values from 0 to n inclusive

$$P(s, np) = e^{-np} \sum_{i=0}^{s} \frac{np^i}{i!},$$

and

$$\Delta P(s,\ np) \;=\; e^{-np}\,\frac{np^{s}}{s!}\ .$$

This approximation is the first two terms of what Rietz (1927) calls a Gram-Charlier Type B series. The remaining terms may be left out since the series converges very rapidly. This series is an extension of the Poisson exponential approximation and the development can be found in Rietz (1927).

Raff (1956) discusses maximum error, the largest possible error which can arise in estimating any sum of consective binomial terms with the specified parameters. In this approximation the maximum error decreases as p becomes smaller. The size of n does not affect the decrease for the approximation has the property that its maximum errors are practically independent of n. The maximum error can be kept below .005 if the approximation is used only if $n \geqslant 5$ and $np \leqslant 0.8$. He also states that the same maximum error can be had with Camp's approximation if it is used only if $n \quad 5$ and $np \quad 0.8$. While an error of .005 appears to be small, when P is actually .01, the relative error would be 50 per cent.

The example presented in the introduction does not fit the criterion for using the Poisson Gram-Charlier approximation, but the problem will be presented for comparison purposes. The procedure will be to find the probability of having 0, 1, 2, or 3 successes and then to subtract this quantity from 1.

The Poisson Gram-Charlier approximation

$$B(s;n,p) \;=\; P(s,np) + \tfrac{1}{2}p(s-np)\Delta P(s,np)$$

where s goes from 0 to 3, $p = \frac{1}{4}$, $np = \frac{5}{4}$.

The quantity $P(s, np) = e^{-np} \sum_{i=0}^{s} \frac{np^i}{i!} = e^{-\frac{5}{4}} \sum_{i=0}^{3} \frac{(\frac{5}{4})^i}{i!}$

$$= .286505 \left\{ 1 + \frac{5}{4} + \frac{25}{32} + \frac{125}{384} \right\} = .286505 \{3.3568\} = .96174 .$$

The next step is to calculate $\frac{1}{2}p(s-np)\Delta P(s,np) = \frac{1}{2}p(s-np)\, e^{-np} \frac{np^s}{s!}$

$$= \frac{1}{2}\left(\frac{1}{4}\right) \left(3 - \frac{5}{4}\right) e^{-\frac{5}{4}} \frac{(\frac{5}{4})^3}{3!} = \frac{1}{8}\left(\frac{7}{4}\right) (.286503) \left(\frac{125}{384}\right) = .02040 .$$

Therefore $B(s;n,p) = .96174 + .02040 = .98214$.

The desired probability of either 4 or 5 successes out of 5 trials is $1 - .98214 = .01786$.

7. GOVINDARAJULU APPROXIMATION

Govindarajulu uses the fact that the binomial random variable S_n can be expressed as a sum of n independently identically distributed Bernoulli random variables which take the values 1 and 0 with probabilities p and q respectively to develop his approximation. The binomial random variable can also be expressed as a lattice random variable of the form $c + sh$ where $c = 0$, $h = 1$, and $s = 0, 1, \ldots$ By putting these two results together, Esseen's theorem (1945) can be used to find the probability that the number of successes equals some s where $s = 0$ to n.

Let $f_n(s) = P(S_n = s) = \binom{n}{s} p^s q^{n-s}$

$$x = \frac{(s - np)}{(npq)^{\frac{1}{2}}} ,$$

$$\phi_1^{(s)}(x) = \frac{d^s}{dx^s} \phi(x) \qquad s = 1, 2, \ldots$$

where $\emptyset$ denotes the standard normal density function.

Then

$$f_n(s) = (npq)^{-\frac{1}{2}} [\emptyset(x) - \frac{(q-p)}{6(npq)^{\frac{1}{2}}} \emptyset^{(3)}(x) + \frac{1}{24npq} \{(1-6pq)\emptyset^{(4)}(x)$$

$$+ \frac{(q-p)^2}{3} \emptyset^{(6)}(x)\} - \frac{(q-p)}{24(npq)^{3/2}} \{\frac{(1-12pq)}{5} \emptyset^{(5)}(x) + \frac{(1-6pq)}{6} \emptyset^{(7)}(x)$$

$$+ \frac{(q-p)^2}{54} \emptyset^{(9)}(x)\}] + o(n^{-2})$$ as given by Govindarajulu (1963).

To show this, consider the following theorem (Esseen, 1945) with $s = 5$ and X_j as a Bernoulli random variable with $q = 0$ and $p = 1$:

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed lattice random variables taking values of the form $c + sh$, $s = 0, + 1, \ldots,$ where h is the maximum span of the distribution. Let $S_n = X_1 + X_2 + \ldots + X_n$. Also, let μ and σ respectively denote the mean and standard deviation of X_i. If X_i has finite absolute moment of the order $s (s \geq 3)$ inclusive, then for all k

$$P(S_n = nc + hs) = \frac{h}{\sigma n^{\frac{1}{2}}} \{\emptyset(x) + \sum_{j=1}^{s-2} n^{-\frac{j}{2}} P_j(-\emptyset(x))\} + o(n^{-(s-1)/2})$$

where

$$x = (nc + hs - nu)/\sigma n^{\frac{1}{2}}$$ and P_j is a polynomial as defined below.

If $F_n(x)$ is the cumulative distribution function of a standardized sum of n independent and identically distributed random variables having variance σ^2 and cumulants $\{\lambda_r \sigma^r ; r \geq 3\}$ and $\emptyset$ is the standard normal c.d.f. then

$$F_n(x) = \phi(x) + \sum_{j=1}^{\infty} n^{-j/2} P_j(-\phi(x))$$

where $P_j(-w)$ is a polynomial of degree $3j$ with coefficients depending on the cumulants of the component random variables of orders 3 through $j + 2$. They are defined by

$$\exp\left\{\sum_{j=1}^{\infty} \frac{\lambda_{j+2}}{(j+2)!} (-w)^{j+2} n^{-j/2}\right\} = 1 + \sum_{1}^{\infty} P_j(-w)n^{-j/2}$$

where $P_j(-\phi)$ is calculated by replacing w^r by $\phi(r)$ in $P_j(-w)$,

$\phi^{(r)}(x) = \frac{d^r}{dx^r} \phi(x)$ (Edgeworth, 1956).

Let $c = 0$

$h = 1$

$s = 0, 1, 2, \ldots$

$p =$ mean of X_i

$pq =$ variance of X_i

Using the previous conditions,

$$x = \frac{(n(0) + s - np)}{(pq)^{\frac{1}{2}} (n)^{\frac{1}{2}}} = \frac{(s - np)}{(npq)^{\frac{1}{2}}}$$

$$P(S_n = s) = \frac{1}{\sqrt{npq}}\left\{\phi(x) + \sum_{j=1}^{5-2} n^{-j/2} P_j(-\phi(x))\right\} + o(n^{-\frac{(s-1)}{2}})$$

$$= \frac{1}{\sqrt{npq}}\left\{\phi(x) + \sum_{j=1}^{3} n^{-j/2} P_j(-\phi(x))\right\} + o(n^{-2})$$

$$= \frac{1}{\sqrt{npq}}\left\{\phi(x) + n^{-\frac{1}{2}} P_1(-\phi(x) + n^{-1} P_2(-\phi(x))\right.$$

$$\left. + n^{-3/2} P_3(-\phi(x)]\right\} + o(n^{-2})$$

To evaluate the polynomials, consider the cumulants of the point binomial where $\sigma^2 = pq$, and $\mu = p$. Kendall and Stuart (1958) define the cumulants in terms of moments.

$$k_1 = p$$

$$k_2 = p - p^2$$

$$k_3 = p - 3p^2 + 2p^3$$

$$k_4 = p - 7p^2 + 12p^3 - 6p^4$$

$$k_5 = p + 15p^2 + 50p^3 - 60p^4 + 24p^5$$

In the problem the cumulants are defined as $k_i = \lambda_i \sigma^i$ so that

$$\lambda_3 = \frac{k_3}{\sigma^3}$$

$$= \frac{p - 3p^2 + 2p^3}{(pq)^{\frac{3}{2}}}$$

$$= \frac{p(1 - 3p + 2p^2)}{(pq)^{\frac{3}{2}}}$$

$$= \frac{p[(1 - 2p + p^2) - (p - p^2)]}{(pq)^{\frac{3}{2}}}$$

$$= \frac{p[q^2 - pq]}{(pq)^{\frac{3}{2}}}$$

$$= \frac{pq(q - p)}{(pq)^{\frac{3}{2}}}$$

$$= \frac{q - p}{(pq)^{\frac{1}{2}}}$$

$$P_1(-\phi(x)) = (6)^{-1} \lambda_3 (-\phi(x))^3$$

$$= (6)^{-1} \frac{q - p}{(pq)^{\frac{1}{2}}} (-\phi(x))^3$$

Similarily P_2 and P_3 may be calculated

$$P_2(-\phi(x)) = (24)^{-1} \lambda_4 \phi^4(x) + (72)^{-1} \lambda_3^2 \phi^6(x)$$

$$P_3(-\phi(x)) = (120)^{-1}(-1) \lambda_5 \phi^5(x) + (144)^{-1}(-1) \lambda_3 \lambda_4 \phi^7(x)$$

$$+ 6^{-4}(-1) \lambda_3^3 \phi^9(x)$$

Then

$$f_n(s) = (npq)^{-\frac{1}{2}} [\phi(x) - \frac{(q - p)}{6(npq)^{\frac{1}{2}}} \phi^{(3)}(x) +$$

$$\frac{1}{24npq} \left\{(1-6pq) \phi^{(4)}(x) + \frac{(q - p)^2}{3} \phi^{(6)}(x)\right\}$$

$$- \frac{(q - p)}{24(npq)^{\frac{3}{2}}} \left\{ \frac{(1-12pq)}{5} \phi^{(5)}(x) + \frac{(1-6pq)}{6} \phi^{(7)}(x) + \right.$$

$$\left. \frac{(q - p)^2}{54} \phi^{(9)}(x)\right\}] + o(n^{-2})$$

The approximation for the cumulative binomial distribution function can be found by summing the terms of $f_n(s)$.

Because the Bernoulli random variables have cumulants existing for all orders, the error term can be made as small as one wishes by continuing the expansion to as many terms as needed.

CONCLUSION

The result of Govindarajulu approximation is that the error term can be made as small as one wishes by continuing the expansion to as many terms as needed. However the mathematics involved in doing this becomes more complicated as more terms are calculated in order to reduce the error.

The Poisson Gram-Charlier approximation performs best if it is used only if $n \geq 5$ and $np \leq 0.8$. In this range the maximum error can be kept below .005, but with an error of .005 and $P = .01$, the relative error would be 50 per cent.

Peizer and Pratt's approximations, z_1, or the more refined z_2, given their graph to calculate g, are easier to calculate than Camp's cube root approximation $z'_{1/3}$. Also it has been shown by asymptotic theory that Peizer and Pratt's approximations have less error than Camp's. Pinkham's square root approximation $z'_{\frac{1}{2}}$ is easier to calculate than either Peizer and Pratt's or Camp's, but also is less accurate.

The method that one chooses to use should be determined by the amount of accuracy needed and the time that can be allotted.

ACKNOWLEDGMENTS

I am very much indebted to Dr. W. J. Conover for his advice and guidance in the preparation of this report. I am also indebted to Dr. Marcel Neuts for his suggestion of the papers written by Dr. David Peizer, Dr. John W. Pratt, and Dr. Zakkula Govindarajulu.

Most of this report is based on the papers by Peizer, Pratt, and Govindarajulu. The following references were cited by them, but not sighted by the author of this report.

DeMoivre, A. (1730)

Edgeworth, F. Y. (1956)

Esseen, Carl Gustav (1945)

Freeman, Murray F. and Tukey, John W. (1949)
Memorandum Report 24

Pinkham, Roger (1957)

Wilson, Edwin B. and Hilferty, Margaret M. (1931).

REFERENCES

Bahadur, R. R. (1960). Some approximations to the binomial distribution. Ann. of Math. Statist. 31 43-54.

Camp, Burton H. (1951). Approximation to the point binomial. Ann. Math. Statist. 22 130-131.

DeMoivre, A. (1730). Miscellanea Analytica.

Edgeworth, F. Y. (1956). The law of error. Cambridge Philos. Trans. 20 36-66 and 113-141.

Esseen, Carl Gustav. (1945). Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. Acta Math., Uppsala. 77 1-125.

Feller, William. (1950). An Introduction to Probability Theory and Its Applications. John Wiley & Sons, Inc., New York.

Freeman, Murray F. and Tukey, John W. (1949). Transformations Related to the Angular and the Square-Root. Memorandum Report 24, Statistical Research Group, Princeton University.

Freeman, Murray F. and Tukey, John W. (1950). Transformations related to the angular and the square-root. Ann. Math. Statist. 21 607-611.

Govindarajulu, Zakkula. (1963). Normal approximations to the classical discrete distributions. Classical and Contagious Discrete Distributions. Pergamon Press, New York.

Kendall, Maurice and Stuart, Alan. (1958). The Advanced Theory of Statistics, Vol. I. Charles Griffin & Company Limited, London.

Lukacs, Eugene. (1960). Characteristic Functions. Hafner Publishing Company, New York.

Paulson, Edward. (1942). An approximate normalization of the analysis of variance distribution. Ann. Math. Statist. 13 233-235.

Peizer, David B. and Pratt, John W. (1966). Approximating the binomial, F, and commonly used related distributions, I. Technical Report No. 12 Contract Nonr 1866(37), NR-042-097.

Pinkham, Roger. (1957). The First Four Moments of a Transformed Beta Variable. Memorandum Report 61, Statistical Research Group, Princeton University.

Pratt, John W. (1966). Approximating the binomial, F, and commonly used related distributions, II. Technical Report No. 13 Contract Nonr 1866(37), Nr-024-097.

Raff, Morton S. (1956). On approximation the point binomial. J. of the Amer. Statist. Assoc. 51 293-303.

Rietz, Henry Lewis. (1927). Mathematical Statistics. The Open Court Publishing Company, Chicago.

Wilks, Samuel S. (1962). Mathematical Statistics. John Wiley & Sons, Inc., New York, London.

Wilson, Edwin B. and Hilferty, Margaret M. (1931). The distribution of chi-square. Proceedings of the National Academy of Sciences. 17 684-688.

APPROXIMATIONS TO THE BINOMIAL

by

MYRTLE ANNA BRUCE

B. A., Carson Newman College, Jefferson City, Tennessee, 1965

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

The binomial distribution is hard to tabulate directly since the tabulation involves three parameters. There are many different approximations for the binomial, but the problem of finding a satisfactory approximation appears to be that of determining the error involved when each is used. Five different approximations are presented in this paper.

Let P be a binomial tail probability. Then define $\emptyset$ to be the standard normal left or right tail function according as P is a left or right tail probability. A value z is sought such that

$$\emptyset_1(z) = \int_{-\infty}^{z} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} dx$$ is approximately equal to P if P is a left tail probability

or

$$\emptyset_2(z) = \int_{z}^{\infty} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} dx$$ is approximately equal to P if P is a right tail probability.

Peizer and Pratt have developed a new approximation for estimating z, z_1 or the more refined z_2, which appears to give a smaller error than Camp's cube root approximation for z, Pinkham's square root approximation for z, or the usual normal approximation.

The Poisson Gram-Charlier approximation is based on the fact that the binomial distribution can be estimated by the Poisson distribution and performs best if it is used only if $n \geq 5$ and $np \leq 0.8$.

Finally Govindarajulu's approximation is derived by using the fact that the binomial random variable S_n can be expressed as a sum of n independently identically distributed Bernoulli random variables and that these random variables are also lattice random variables. His approximation can be made as accurate as one desires since the error term can be made as small as desired.