

SOME CONTRIBUTIONS TO THE EVALUATION  
OF PEARSONIAN DISTRIBUTION FUNCTIONS

by

John Edward White

Thesis submitted to the Graduate Faculty of the  
Virginia Polytechnic Institute  
in candidacy for the degree of  
MASTER OF SCIENCE  
in  
Statistics

March 1960

Blacksburg, Virginia

## TABLE OF CONTENTS

	page
I. INTRODUCTION	4
1.1 General Remarks on Pearsonian Method	4
1.2 Notation	5
II. REVIEW OF LITERATURE	7
2.1 Pearson Distributions	7
2.2 Computation of Percentage Points of Pearson Distributions	7
III. COMPUTING PERCENTAGE POINTS	9
3.1 Objectives	9
3.2 Plan of Attack	9
3.3 Type I Pearson Distribution	10
3.4 Type II Pearson Distribution	14
3.5 Type VI Pearson Distribution	15
3.6 Type III Pearson Distribution	17
IV. INSTRUCTIONS FOR THE USE OF THE TABLES	19
V. NUMERICAL ILLUSTRATIONS	20
5.1 Comparison of the Wald-Brookner Series and Pearson Method in the Test of Independence	20
5.2 Approximating the Binomial Distribu- tion by a Pearson Distribution	21

5.3	Comparison of the Probabilities of the Extreme Order Statistics when the Underlying Distribution is Assumed to be Standard Normal	24
5.4	Comparison of Pearson Method with the Exact c.d.f. of a J-Shaped Curve	25
VI.	APPENDIX A - TABLES OF PERCENTAGE POINTS	27
6.1	Discussion of the Tables	27
	Upper 5% points	29
	Lower 5% points	32
	Upper 2.5% points	35
	Lower 2.5% points	38
	Upper 1% points	41
	Lower 1% points	44
	Upper 0.5% points	47
	Lower 0.5% points	50
VII.	APPENDIX B - THE INCOMPLETE BETA FUNCTION	53
VIII.	APPENDIX C - THE IBM 650 PROGRAM FOR THE PEARSON DISTRIBUTION	57
IX.	BIBLIOGRAPHY	62
X.	ACKNOWLEDGEMENTS	64
XI.	VITA	65

## I. INTRODUCTION

### 1.1 General Remarks on Pearsonian Method

The fitting of Pearson curves [7] by the method of moments, that is, fitting a curve by using only the first four moments about the mean, can be divided into two categories:

- 1) First, as a method of approximating theoretical distribution functions;
- 2) Secondly, to obtain an expression to represent the underlying distribution when the moments were obtained from a sample.

In many instances, the Pearson system has been used to approximate distribution functions and to obtain percentage points when only a few moments were known. In several cases, comparisons with exact distributions which were derived later showed the satisfactory closeness of the approximation [6]. This occurred in the case of the distribution of the range, where the table of percentage points based on Types I and VI curves were in existence some ten years earlier than those based upon the exact distribution. The Pearson approach has been proven to be adequate for the multivariate likelihood-ratio statistic developed by Wilks.

The second method has been criticized, for in many cases the calculated moments from a sample do not lead to the best

estimates of the population parameters. Although criticized, in practical experience this method has proved very useful.

It is also illustrated in a later section of this paper, that the Pearson approach is quite accurate as an approximation to the test of independence, usually obtained by the Wald-Brookner Series.

Thus, percentage points of the Pearson curves, tabulated according to  $\beta_1$  and  $\beta_2$ , have provided very useful approximations, and may be even more useful now that greater accuracy has been attained.

## 1.2 Notation

It will be helpful to list the notation specific to this paper:

<u>Symbol</u>	<u>Meaning</u>
$\beta_1$	$\frac{\mu_3^2}{\mu_2^3}$ , where $\mu_i$ represents the $i$ th population moment around the mean.
$\beta_2$	$\frac{\mu_4}{\mu_2^2}$ .
$\alpha$	a given probability level.
$X_\alpha$	percentage point, expressed in standardized measure, $[(x - \mu)/\sigma]$ .

$I_x(p,q)$  Incomplete Beta function, with argument  $x$ , and parameters,  $p, q$ .

$B_x(p,q)$   $\int_0^x y^{p-1}(1-y)^{q-1}dy$  (same as  $I_x$  except for constant).

c.d.f Cumulative distribution function.

## II. REVIEW OF THE LITERATURE

### 2.1 Pearson Distributions

A system of frequency-curves, defined by the solutions of the differential equation

$$(2.1.1) \quad \frac{dy}{dx} = \frac{y(x - a)}{b_0 + b_1x + b_2x^2 + \dots},$$

was presented by Karl Pearson, and hence has been known as the Pearson Distributions. For solutions of the above differential equation and the derivation of the Pearson curves, see Kendall [3], or a more detailed discussion may be found in Elderton [2]. Elderton gives the derivations of the three main types of Pearson curves and also the important transition curves (i.e., boundaries between the three main types).

### 2.2 Computation of Percentage Points of the Pearson Distribution

Pearson [4] tabled percentage points of the Pearson Distributions to a limited degree. At that time,  $\beta_1$  and  $\beta_2$  were chosen so as to cover the range of the Type I curves likely to be met within the applications of Baye's Theorem, but the fact was recognized that the tables could have been used in other problems, too. Pearson obtained percentage points by approximations rather than by evaluation of the

function, which, however, would have been very difficult due to the non-existence, at the time, of tables of the incomplete Beta function.

Pearson and Merrington [6] extended the tables to their present status as given in Table 42 [5]. The problem was attacked by them in a method similar to the one used by Pearson [4] in order to make use of the earlier tables. This table, which was constructed with the help of Pearson's Tables of the Incomplete Beta Function [8], is considerably more extensive than the former. However the J and U-shaped curves were not included since they would have required extrapolation of the Incomplete Beta Tables. Thus, the tables were bounded above by points lying on the Type IX line, i.e. points lying on the curve

$$y = y_0(1 + x/a)^m,$$

which is the boundary between the unimodal and the J-shaped curves.

Exact percentage points were calculated, whenever possible, and then various interpolation techniques were used to obtain the remaining points. Again, it was the non-existence of tables for the Incomplete Beta Function for fractional degrees of freedom which made the computational work laborious.

### III. PERCENTAGE POINTS OF THE PEARSON DISTRIBUTIONS

#### 3.1 Objectives

It has been suggested that a more extensive table of percentage points for the Pearson Distributions would be helpful in statistical work. For example, the Army Ballistic Missile Agency has listed this project as one of the research projects to be considered for their tracking problems.

The purpose of this paper is to present a more extensive and a more accurate table than those now available. The ranges of  $\beta_1$  and  $\beta_2$  have been extended so as to have  $0 \leq \beta_1 \leq 1.8$  and  $1.2 \leq \beta_2 \leq 6.6$ . This extension also allows coverage of the J and U-shaped curves, up to the limit of convergence of our method, which is almost at the limit of frequency distributions.

It is now possible to obtain the probability for a given percentage point, by the Pearson Program for the IBM 650. (See Appendix C). This is a more general result than the actual percentage points, as will be demonstrated in a few examples.

#### 3.2 Plan of Attack

The derivation of any particular type of Pearson Distribution may be found in Elderton [2] or Kendall [3]. The plan followed here was to take the main types of Pearson

Distributions and break them down into a form more accessible for high speed computing.

Generally, to obtain the percentage points, a solution for  $x_\alpha$  must be obtained from the integral equation

$$(3.2.1) \quad \alpha = \int_{l_2}^{x_\alpha} y \, dx$$

where,  $y = f(x)$ , is a particular Pearson Distribution,  $l_2$  is the lower limit of  $x$ , and  $\alpha$  is a given probability level. Then in the body of the tables are the values of

$$(3.2.2) \quad X_\alpha = (x_\alpha - \mu)/\sigma$$

for  $\alpha = 0.005, 0.01, 0.025, 0.05, 0.95, 0.975, 0.99,$  and  $0.995$ .

One main type of the Pearson Distributions, Type IV, will not be considered in this paper due to the complexity of the function and lack of time. This work will be continued at a later date.

### 3.3 Type I, Pearson Distribution

If in equation (2.1.1) we apply a transformation of the origin to the mode we have, after eliminating all terms of order greater than  $x^3$ ,

$$(3.3.1) \quad \frac{dy}{dx} = \frac{y(x-a)}{B_0 + B_1(x-a) + B_2(x-a)^2},$$

or

$$(3.3.2) \quad \frac{d}{dx} (\int^n y) = \frac{X}{B_0 + B_1 X + B_2 X^2} .$$

If the roots,  $\alpha_1$  and  $\alpha_2$ , of the denominator on the right-hand side of (3.3.2) are considered to be real and of opposite sign then

$$\begin{aligned} \frac{d}{dx} (\int^n y) &= \frac{X}{B_2(X + \alpha_1)(X - \alpha_2)}, \quad \alpha_1, \alpha_2 > 0 \\ &= \frac{\alpha_1}{B_2(\alpha_1 + \alpha_2)} \frac{1}{X + \alpha_1} + \frac{\alpha_2}{B_2(\alpha_1 + \alpha_2)} \frac{1}{X - \alpha_2}, \end{aligned}$$

giving

$$(3.3.3) \quad y = k(X + \alpha_1)^{\frac{\alpha_1}{B_2(\alpha_1 + \alpha_2)}} (X - \alpha_2)^{\frac{\alpha_2}{B_2(\alpha_1 + \alpha_2)}} .$$

This may be written in the form

$$(3.3.4) \quad y = k(1 + x/a_1)^{m_1} (1 - x/a_2)^{m_2}, \quad -a_1 \leq x \leq a_2$$

where  $m_1 a_2 = m_2 a_1$ .

This is the form in which the Type I distribution is usually expressed. It covers also the J and U-shaped curves, i.e., if  $m_1$  or  $m_2$  is negative, the curve is J-shaped, and if both  $m_1$  and  $m_2$  are negative, the curve is U-shaped.

To obtain the constants for the above curve, moments about the line  $x = -a_1$  are obtained;  $k$  is found by integrating over the range of  $x$ .

After changing the origin to get moments about the mean and then setting  $\beta_1 = \mu_3^2/\mu_2^3$ , and  $\beta_2 = \mu_4/\mu_2^2$ , expressions for the parameters can be obtained in terms of  $\beta_1$  and  $\beta_2$ . (For complete detail of this procedure, see Elderton [2]).

These expressions for the parameters are

$$k = \frac{1}{a_1+a_2} \frac{m_1^{m_1} m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}} \frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)},$$

$$r = 6(\beta_2 - \beta_1 - 1)/(6 + 3\beta_1 - 2\beta_2),$$

$$a_1+a_2 = \frac{1}{2} \sqrt{\mu_2} [\beta_1(r+2)^2 + 16(r+1)]^{\frac{1}{2}},$$

(3.3.5)

and the  $m$ 's are expressible as

$$m_1, m_2 = \frac{1}{2} \left[ (r-2) \pm r(r+2) \sqrt{\frac{\beta_1}{\beta_1(r+2)^2 + 16(r+1)}} \right]$$

where  $m_2$  is the positive root if  $\mu_3 > 0$ .

If we employ the method described earlier, to construct percentage points of this distribution, we have

$$(3.3.6) \quad \alpha = k \int_{-a_1}^{x_\alpha} (1 + x/a_1)^{m_1} (1 - x/a_2)^{m_2} dx.$$

Substituting  $x = (a_1 + a_2)y - a_1$  and the value of  $k$  given in (3.3.5) we have

$$(3.3.7) \quad \alpha = \frac{1}{B(m_1+1, m_2+1)} \int_0^{\frac{x_\alpha + a_1}{a_1 + a_2}} y^{m_1} (1-y)^{m_2} dy,$$

or

$$(3.3.8) \quad \alpha = \frac{1}{B(m_1+1, m_2+1)} \int_0^\theta y^{m_1} (1-y)^{m_2} dy$$

$$\text{where } \theta = \frac{x_\alpha + a_1}{a_1 + a_2}.$$

As was stated earlier we must obtain a solution for  $x_\alpha$  in the integral equation, here, equation (3.3.6), or equivalently, obtain a solution for  $\theta$  in equation (3.3.8).

To solve (3.3.8), let

$$(3.3.9) \quad f(\theta) = \frac{1}{B(m_1+1, m_2+1)} \int_0^\theta y^{m_1} (1-y)^{m_2} dy,$$

then we have the equation

$$(3.3.10) \quad \alpha - f(\theta) = 0. \quad *)$$

Then, by using the Newton iteration technique, approximations to  $\theta$  can be obtained to any desired accuracy, i.e.,

$$\theta_{i+1} = \theta_i - \frac{\alpha - f(\theta_i)}{f'(\theta_i)}.$$

Hence, given  $\theta$  as a solution to (3.3.8) we may directly evaluate the percentage point. As defined before,

$$X_\alpha = \frac{x_\alpha - \mu}{\sigma}; \quad \text{but } \theta = \frac{x_\alpha + a_1}{a_1 + a_2},$$

---

\*) For complete explanation of how  $f(\theta)$  was evaluated, see Appendix B.

therefore,

$$\begin{aligned} X_{\alpha} &= \frac{(a_1+a_2)\theta}{\sigma} - \frac{a_1}{\sigma} - \frac{\mu}{\sigma} \\ &= \frac{(a_1+a_2)\theta}{\sigma} - \frac{(a_1+a_2)(m_1+1)}{r\sigma}, \end{aligned}$$

since 
$$\mu = \frac{(a_1+a_2)(m_1+1)}{r} - a_1.$$

Using the value of  $(a_1+a_2)$  given in (3.3.5) we have

$$(3.3.11) \quad X_{\alpha} = \frac{\sqrt{\beta_1(r+2)^2+16(r+1)}}{2} \left( \theta - \frac{m_1+1}{r} \right).$$

Hence, given  $\beta_1$  and  $\beta_2$  and a solution of equation (3.3.8), percentage points may be computed.

In the actual computing program, the skewness was taken to be negative, i.e.,  $m_1$  in (3.3.5) was taken as the positive square root. Hence, to obtain points as given in the tables,  $\alpha$  was replaced by  $1 - \alpha$ , and signs were changed on  $X_{\alpha}$ , the percentage points; this procedure was adopted because it was slightly faster on the computer.

#### 3.4 Type II, Pearson Distribution

If in (3.3.4), we consider  $m_1 = m_2$ , or, equivalently, if we let  $\beta_1 = 0$ , we will have a symmetrical distribution, or Type II, a "transition type" Pearson curve. This is computed in the same manner as the Type I distribution, by merely setting the degrees of freedom equal in the incomplete Beta function.

### 3.5 Type VI, Pearson Distribution

This distribution, considered by Pearson as the third main type is, in many aspects, quite similar to the Type I distribution. Here, the factorising of (3.3.2) is identical to the method employed in the Type I distribution, except that the roots of the denominator on the right hand side are real and have the same sign.

In a manner similar to that used in deriving Type I, we obtain

$$(3.5.1) \quad y = y_0(x - a)^{q_2} x^{-q_1}, \quad q_1 > q_2 - 1.$$

The range of the curve is from  $a$  to  $\infty$ , except when  $\mu_3 < 0$ ; then  $a$  is negative and the range is from  $-\infty$  to  $a$ .

The parameters are derived by the method of moments, as they were in the Type I distribution. In this case,

$$r = 6(\beta_2 - \beta_1 - 1)/(6 + 3\beta_1 - 2\beta_2), \quad r < 0.$$

$$a = \frac{1}{2} \sqrt{\mu_2} \sqrt{\beta_1(r+2)^2 + 16(r+1)},$$

$$(3.5.2) \quad \begin{matrix} q_2 \\ -q_1 \end{matrix} = \frac{r-2}{2} \pm \frac{r(r+2)}{2} \sqrt{\frac{\beta_1}{\beta_1(r+2)^2 + 16(r+1)}},$$

and

$$y_0 = \frac{a^{q_1 - q_2 - 1} \Gamma(q_1)}{\Gamma(q_1 - q_2 - 1) \Gamma(q_2 + 1)}.$$

To construct the percentage points, assume  $\mu_3 > 0$ , then from (3.2.1), we have,

$$\begin{aligned}
 \alpha &= y_0 \int_a^{x_\alpha} (x-a)^{q_2} x^{-q_1} dx. \\
 (3.5.3) \quad &= y_0 a^{q_2-q_1} \int_a^{x_\alpha} \left(\frac{x}{a} - 1\right)^{q_2} \left(\frac{x}{a}\right)^{q_1} dx.
 \end{aligned}$$

Substituting  $\frac{1}{z} = \frac{x}{a}$  and expressing the result as an incomplete Beta function, we have

$$(3.5.4) \quad \alpha = 1 - \frac{1}{B(q_2+1, q_1-q_2-1)} \int_0^\theta (1-z)^{q_2} z^{q_1-q_2-2} dz,$$

where  $\theta = a/x_\alpha$ , and can be obtained from the integral equation.

The standardized variate,  $X_\alpha = \frac{x_\alpha - \mu}{\sigma}$ , expressed in terms of  $\theta$  is

$$\begin{aligned}
 (3.5.6) \quad X_\alpha &= \frac{a/\theta - \mu}{\sigma} \\
 &= \frac{a}{\sigma} \left( \frac{1}{\theta} - \frac{\mu}{a} \right) \\
 &= \frac{1}{2} \sqrt{\beta_1 (r+2)^2 + 16(r+1)} \left( \frac{1}{\theta} + \frac{q_1-1}{r} \right),
 \end{aligned}$$

since  $\mu = \frac{a(1-q_1)}{r}$ . This can be seen from the Type I distribution, for it is obvious that  $m_1 = -q_1$  and  $a = a_1 + a_2$ ; the  $r$ 's being algebraically identical. As before, we employ the Newton method to obtain a close approximation of  $\theta$  from equation (3.5.4). [See Appendix B.]

### 3.6 Type III, Pearson Distribution

For the few remaining points that lie on the Type III curve, i.e., points that satisfy the relation  $2\beta_2 - 3\beta_1 - 6 = 0$ , the following simple scheme has been used.

This distribution occurs if  $b_2 = 0$  in (2.1.1), hence, the curve is expressible as

$$(3.6.1) \quad y = y_0 e^{-px/a} (1 + x/a)^p, \quad -a \leq x \leq \infty.$$

The parameters, obtained by taking moments about the point  $x = -a$ , are

$$y_0 = \frac{1}{a} \frac{p^{p+1}}{e^p \Gamma(p+1)},$$

$$p = \frac{4}{\beta_1} - 1,$$

$$a = \frac{2\mu_2^2}{\mu_3} - \frac{\mu_3}{2\mu_2}.$$

(See Elderton [2], page 95 for further detail.)

If the substitution

$$z = \frac{2p}{a} (x + a).$$

is used in (3.6.1), we can write

$$(3.6.2) \quad y = \frac{1}{2^{p+1} \Gamma(p+1)} e^{-z/2} z^p.$$

Now, let  $p = \frac{n-2}{2}$ , and hence we have the familiar  $\chi^2$  distribution,

$$(3.6.3) \quad y = \frac{1}{2^{n/2} \Gamma(n/2)} z^{n/2 - 1} e^{-z/2} .$$

To compute percentage points, a solution for  $\delta$  from the equation

$$(3.6.4) \quad \alpha = \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^{\delta} z^{n/2 - 1} e^{-z/2} dz,$$

can be obtained by interpolating in Table 8 of [5]. Having  $\delta$  (i.e.,  $\frac{2p}{a}(a + x_{\alpha})$ , in terms of the former variables) we can immediately obtain the percentage points, in a manner similar to those previously used. For,

$$\mu = \frac{a(p+1)}{p} - a,$$

and

$$\sigma = \frac{a \sqrt{p+1}}{p} ,$$

hence it can easily be shown that

$$(3.6.5) \quad x_{\alpha} = \frac{1}{2 \sqrt{p+1}} [\delta - 2(p+1)].$$

## IV. INSTRUCTIONS FOR THE USE OF THE TABLES

The tables are presented as in Table 42 [5], assuming  $\mu_3 > 0$ , i.e., the distributions are assumed to be positively skewed (long tail at right). Of course the upper percentage points, ( $\alpha > 0.50$ ) are positive and the lower percentage points, ( $\alpha < 0.50$ ) are negative.

If  $\mu_3 < 0$ , the roles of the tables must be interchanged. That is to say, if  $\mu_3 < 0$  and the lower percentage points are desired, i.e.,  $\alpha < 0.50$ , obtain the value desired from the tabled upper percentage points, attaching a negative sign; and if  $\mu_3 < 0$  and the upper percentage points are desired, i.e.,  $\alpha > 0.50$ , then read the desired value from the tabled lower percentage points, attaching a positive sign.

For clarity, suppose  $\mu_3 < 0$ , and the lower 5% point, i.e.,  $\alpha = 0.05$ , corresponding to  $\beta_1 = 0.05$  and  $\beta_2 = 2.0$  is desired. From the table of upper 5% points we obtain, after attaching a negative sign, -1.7168. The upper 5% point, i.e.,  $\alpha = 0.95$ , corresponding to  $\beta_1 = 0.05$  and  $\beta_2 = 2.0$  is 1.4746, from the table of lower 5% points.

## V. NUMERICAL ILLUSTRATIONS

### 5.1 Comparison of the Wald-Brookner Series and the Pearson Method in the Test of Independence

This example, suggested by Bargmann and contained in [1], is a typical study illustrating the test of independence. The determinant of the sample correlation matrix,  $|R|$ , was found, by Bargmann, to be 0.3225. The parameters, namely,  $\beta_1$ ,  $\beta_2$ ,  $\mu$ , and  $\mu_2$  were found to be, respectively, 0.075145, 2.750390, 0.640553, and 0.015255.

For a Pearson approximation,  $X_\alpha = -2.575708$  was obtained by (3.2.2), where in this case,  $x_\alpha = |R|$ .

Since it is well known from the distributions of  $|R|$  that  $\mu_3 < 0$ , the roles of the upper and lower tables must be interchanged. From the Table of Upper 1% Points we find that

$$\Pr(|R| \leq 0.3225) \approx 0.01.$$

The actual probability obtained by using the IBM 650 Program was 0.006908.

As is well known, the true probability may be computed from the Wald-Brookner Series, to any desired degree of accuracy. For six place accuracy, in this case, we require:

$$\begin{aligned} \Pr(-m \log |R| \leq c) &= \Pr(\chi_f^2 \leq c) \\ &+ \frac{v}{m^2} [\Pr(\chi_{f+4}^2 \leq c) - \Pr(\chi_f^2 \leq c)], \end{aligned}$$

where

$$f = \frac{1}{2} p(p-1),$$

$$m = N - \frac{2p + 11}{6},$$

and 
$$v = \frac{p(p-1)}{288} (2p^2 - 2p - 13).$$

In this case

$$f = 10,$$

$$m = 21.5,$$

and 
$$v = 1.875.$$

Using Table 7 of [5] to compute the probabilities, we find,

$$\Pr(-m \log |R| \leq c) = 0.00700.$$

Obviously, the difference in the probabilities is only 0.00009, thus, in this case, the Pearson approach seems to be quite adequate.

## 5.2 Approximating the Binomial Distribution by a Pearson Distribution

It is well known that the normal distribution provides a good approximation to the binomial if  $np > 5$  when  $p \leq \frac{1}{2}$ , and  $nq > 5$  when  $p > \frac{1}{2}$ . Now, let us try the Pearson approximation for  $n = 6$  and  $p = 1/3$  and compare this with the normal approximation, and with the true results computed directly from the binomial distribution.

The moments in this case, after applying Sheppard's corrections, are

$$\begin{aligned}\mu &= 2.000000, \\ \mu_2 &= 1.250000, \\ \mu_3 &= 0.444444\dots, \\ \mu_4 &= 4.2513888\dots\end{aligned}$$

Hence,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.101136$$

and

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 2.720889.$$

To approximate the binomial we use the usual procedure, i.e. to find  $\Pr(X \leq 2)$  we use the approximation

$$\int_{-\infty}^{2.5} f(x) dx,$$

where,  $f(x)$ , in one case, represents the normal distribution, and in the other case, the appropriate Pearson Distribution.

The standardized variates  $X_\alpha = \frac{x_\alpha - \mu}{\sigma}$  for  $x_\alpha = 2.5$  and 4.5 are, respectively, 0.447214 and 2.236068.

The following table illustrates the results.

Distribution \ $x_\alpha$	2.5	4.5
Binomial	0.6804	0.9822
Pearson	0.6811	0.9823
Normal	0.6726	0.9874

It is seen that the Pearson approximation is somewhat closer than the normal in this case.

For a second illustration, assume  $n = 6$ , and  $p = 1/10$ . We would expect the approximation in this case to be less accurate, since  $p$  is fairly small.

The corrected moments are,

$$\mu = 0.600000\dots ,$$

$$\mu_2 = 0.456666\dots ,$$

$$\mu_3 = 0.432000\dots ,$$

$$\mu_4 = 0.882366\dots ,$$

which yields  $\beta_1 = 1.959610$  and  $\beta_2 = 4.231073$ .

For  $x_\alpha = 1.5, 2.5$  the standardized variate  $X_\alpha$  is 1.331812, and 2.811603, respectively. The results are

Distribution \ $x_\alpha$	1.5	2.5
Binomial	0.8857	0.9842
Pearson	0.8761	0.9803
Normal	0.9085	0.9975

As was mentioned earlier, the results are not striking, although the Pearson approximation is considerably better than the normal.

5.3 Comparison of the Probabilities of the Extreme Order Statistic When the Underlying Distribution is Assumed to be Standard Normal

This example, suggested by H. A. David, compares the Pearson approximation to the correct results obtained from the normal distribution. Ruben [9] calculated and tabulated moments,  $\beta_1$ , and  $\beta_2$  for extreme order statistics. Suppose we take a random sample of  $n$  observations from  $N(0, 1)$ , the the c.d.f. of the largest observation is

$$(5.3.1) \quad \Pr(X \leq x) = [\Phi(x)]^n,$$

where  $\Phi(x)$  is the standard normal c.d.f.

As a first comparison, take  $n = 20$ , and let us choose  $x$  values so as to obtain probabilities close to 0.95 and 0.995. These  $x$  values are, respectively, 2.80 and 3.48. For a Pearson approximation, we need  $X_\alpha = \frac{x - \mu}{\sigma}$ . From Ruben's tables [9],  $\beta_1 = 0.468,546$ ,  $\beta_2 = 3.525,068$ ,  $\mu = 1.867,475,060$ , and  $\sigma = 0.525,068,2$ , hence, for the above values of  $x$ ,  $X_\alpha = 1.776,007$  and  $3.071,077$ .

The results, tabulated below, were very favorable, even on the extreme end of the distribution.

x		
Method \	2.80	3.48
Normal	0.950,120	0.994,998
Pearson	0.950,632	0.995,049

For a second comparison, take  $n = 50$  and again choose  $x$  values to yield the approximate probabilities, 0.95 and 0.995. The  $x$  values are 3.07 and 3.72; and the corresponding  $X_\alpha$ 's are 1.767,529 and 3.167,039, respectively. By the same procedure as before, with  $\mu = 2.249,073,631$ ,  $\sigma = 464,448,5$ ,  $\beta_1 = 0.356,448$ , and  $\beta_2 = 3.643,728$ , we obtain

Method \ x	3.07	3.72
Normal	0.947,876	0.995,032
Pearson	0.948,380	0.995,102

These few examples present further illustration of the fact that the Pearson approximations are very adequate for many practical purposes, and may be regarded a justifiable procedure if the true distribution is not known. The purpose of this paper was the preparation of improved tables which may lead to more extensive comparative studies of this kind.

#### 5.4 Comparison of Pearson Method with the Exact c.d.f. of a J-Shaped Curve

Let  $x$  and  $y$  be independent random variables each rectangularly distributed,  $R(0,1)$ . Then the distribution of  $u = xy$  can be expressed as

$$f(u) = -\ln u, \quad 0 < u < 1.$$

To obtain moments about the origin, it can easily be shown that

$$E(u^r) = \frac{1}{(r+1)^2}$$

Hence, the moments about the mean are

$$\mu = 0.250,000,000,0 ,$$

$$\mu_2 = 0.048,611,111,1 ,$$

$$\mu_3 = 0.010,416,666,7 ,$$

and  $\mu_4 = 0.007,447,916,7 .$

Then

$$\beta_1 = \mu_3/\mu_2^3 = 0.944,606,$$

and  $\beta_2 = \mu_4/\mu_2^2 = 3.151,837.$

From Table 43 of [5] we see that this combination of  $\beta$ 's leads to a J-shaped distribution.

Choose  $x = 0.30$  and  $0.75$ , then

$$P(x \leq 0.30) = - \int_0^{0.30} f_n u \, du = 0.661,19$$

and  $P(x \leq 0.75) = - \int_0^{0.75} f_n u \, du = 0.965,76.$

Corresponding to the above  $u$  values, the values of  $(u - \mu)/\sigma$  are, respectively,  $0.226,779$  and  $2.267,787$ .

From the Pearson Program we obtain probabilities  $0.660,12$  and  $0.966,16$ . Hence, the Pearson method, in this case, is also satisfactory in this region if a high degree of accuracy is not required.

## VI. APPENDIX A - TABLES OF PERCENTAGE POINTS

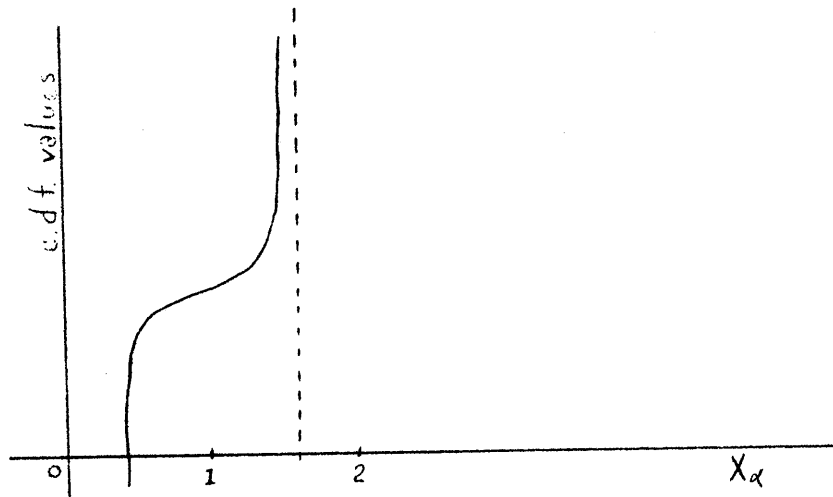
6.1 Discussion of the Tables

There are several important points to be noted here. None are detrimental to the tables but are mentioned merely to explain certain peculiarities that occur.

There are certain values in the tables that remain the same in two or more tables, (e.g., check  $\beta_1 = 0.00$  and  $\beta_2 = 1.2$ ). This occurrence is due to the fact that the value of  $X_\alpha$  in the body of the table satisfies the more stringent of the two probability levels. Changing it by just one unit in the sixth decimal, however, would put this value into the acceptance region of the less stringent one. In some of these cases, the Newton iteration did not converge, unless the first trial value of  $X_\alpha$  was chosen to correspond to a probability below the desired one (on the lower tail, and conversely on the upper tail). The following procedure was then adopted. A value of  $X_\alpha$  was chosen which fell below the range of the distribution. In the program, the lowest (to six decimal places)  $X_\alpha$  which just got inside the range was determined. If the associated probability was below the desired one, a Newton iteration was used to improve  $X_\alpha$  which, in this case, converged. If, however, this lowest permissible value of  $X_\alpha$  exceeded the desired probability level it was recorded uncorrected; for our program did not justify correction in the seventh place and, as was obvious, correction in the sixth

place by even one unit would get  $X_\alpha$  outside the range of the distribution.

On the following drawing, which illustrates the c.d.f. of a U-shaped distribution, let one unit represent a difference of  $10^{-6}$  in  $X_\alpha$ . It will then be seen that the first permissible value may be larger than desired, however, reduction of  $X_\alpha$  by one unit in the sixth decimal would lead to a non-permissible value.



In certain tables, there are values closer to the limit of all frequency distributions than in others. This was due to the impossibility of getting within the range with the six decimal accuracy used in the computational work.

Numerous error and accuracy studies indicated no errors greater than one unit in the fourth decimal place.

















































## VII. APPENDIX B - THE INCOMPLETE BETA FUNCTION

In computing such equations as (3.3.8) a generalized IBM 650 program was needed. Presented here is the method used to express this incomplete Beta integral in computational form.

Consider the function

$$B_{\theta}(m+1, n+1) = \int_0^{\theta} x^m (1-x)^n dx,$$

where  $m \geq n > 1$ .

Integrating by parts, letting  $u = (1-x)^n$ ,  $dv = x^m dx$ , we have

$$B_{\theta}(m+1, n+1) = \frac{1}{m+1} \theta^{m+1} (1-\theta)^n + \frac{n}{m+1} B_{\theta}(m+2, n).$$

Continuing the integration by parts, we obtain the general expression

$$\begin{aligned} (1) \quad B_{\theta}(m+1, n+1) &= \frac{1}{m+1} \theta^{m+1} (1-\theta)^n + \frac{n}{(m+1)(m+2)} \theta^{m+2} (1-\theta)^{n-1} \\ &+ \dots + \frac{n(n-1)\dots(n-[n]+2)}{(m+1)(m+2)\dots(m+[n])} \theta^{m+[n]} (1-\theta)^{n-[n]+1} \\ &+ \frac{n(n-1)\dots(n-[n]+1)}{(m+1)(m+2)\dots(m+[n])} \int_0^{\theta} x^{m+[n]} (1-x)^{\ell} dx, \end{aligned}$$

where  $\ell = n - [n]$ , such that  $0 \leq \ell < 1$ .

Now consider the residual integral,

$$\int_0^{\theta} x^{m+[n]} (1-x)^{\ell} dx,$$

which after applying the binomial expansion to  $(1-x)^\ell$  may be written

$$(2) \quad \int_0^{\theta} \sum_{j=0}^{\infty} \binom{\ell}{j} (-1)^j x^{m+[n]+j} dx$$

$$= \sum_{j=0}^{\infty} (-1)^j \binom{\ell}{j} \frac{\theta^{m+[n]+j+1}}{m+[n]+j+1} .$$

The combinatorial relationships may be rewritten as follows:

$$\binom{\ell}{0} = 1; \quad \binom{\ell}{1} = \ell; \quad \binom{\ell}{2} = \frac{\ell(\ell-1)}{2} = -\frac{\ell(1-\ell)}{2};$$

$$\binom{\ell}{3} = \frac{\ell(\ell-1)(\ell-2)}{6} = \frac{\ell(1-\ell)(2-\ell)}{6},$$

or in general

$$\binom{\ell}{j} = \frac{(-1)^{j+1} \ell(1-\ell)\dots(j-1-\ell)}{j!}, \quad j \geq 2.$$

Hence, we may write

$$(3) \quad \int_0^{\theta} x^{m+[n]} (1-x)^\ell dx = \frac{\theta^{m+[n]+1}}{m+[n]+1} - \frac{\ell}{m+[n]+2} \theta^{m+[n]+2}$$

$$- \sum_{j=2}^{\infty} \frac{\ell(1-\ell)\dots(j-\ell-1)}{j!} \frac{\theta^{m+[n]+j+1}}{m+[n]+j+1}$$

$$= \frac{\theta^{m+[n]+1}}{m+[n]+1} - \frac{\ell}{m+[n]+2} \theta^{m+[n]+2}$$

$$- \sum_{j=1}^{\infty} \frac{\ell(1-\ell)\dots(j-\ell)}{(j+2)!} \frac{\theta^{m+[n]+j+2}}{m+[n]+j+2} .$$

Dividing each term of (3) by  $B(m+1, n+1)$  and writing the result in terms of gamma-functions, we have,

$$(4) \quad I_{\theta}(m+1, n+1) = \sum_{v=1}^{[n]} f(\theta, v) + \frac{\Gamma(m+n+2)}{\Gamma(m+[n]+1)\Gamma(n-[n]+1)} \left[ \frac{\theta^{m+[n]+1}}{m+[n]+1} - \ell \frac{\theta^{m+[n]+2}}{m+[n]+2} - \sum_{j=1}^{\infty} \phi(\theta, j) \right]$$

where

$$f(\theta, v) = \frac{\Gamma(m+n+2)}{\Gamma(m+v+1)\Gamma(n-v+2)} \theta^{m+v} (1-\theta)^{n-v+1},$$

$$\phi(\theta, j) = \frac{\ell(1-\ell)\dots(j-\ell)}{\Gamma(j+2)} \frac{\theta^{m+[n]+j+2}}{m+[n]+j+2},$$

and  $\ell$  is defined as before.

A similar expression was obtained for the case when  $-1 < \ell < 0$ . Obviously if  $0 < n < 1$ , the first summation was set equal to zero and just the latter terms were used.

It is easily seen that  $\sum_{j=1}^{\infty} \phi(\theta, j)$  does not converge rapidly if  $\theta$  becomes large. For values of  $\theta > 0.85$  the relationship

$$I_{\theta}(p, q) = 1 - I_{1-\theta}(q, p)$$

was used, thus making the parameter one which converges very rapidly.

The Incomplete-Beta Subroutine, available from the Virginia Polytechnic Institute Computing Center, program

number 6.6.010.1, is quite general with the following limits on the parameters;

$$0.000,001 \leq p \leq 9999.999,999,$$

$$0.000,001 \leq q \leq 9999.999,999,$$

and, of course,

$$0 \leq \theta \leq 1,$$

where  $p$  and  $q$  are defined as usual in the incomplete Beta function, i.e.,

$$I_{\theta}(p,q) = \frac{\int_0^{\theta} x^{p-1}(1-x)^{q-1} dx}{\int_0^1 x^{p-1}(1-x)^{q-1} dx} .$$

VIII. APPENDIX C - THE IBM 650 PROGRAM  
FOR PEARSON DISTRIBUTIONS

(1) General Remarks

The general Pearson program will compute either percentage points for a given probability or the probability associated with a given point, for  $\beta_1$  and  $\beta_2$  in Type I, Type VI, or Type II regions.\* The ranges of  $\beta_1$  and  $\beta_2$  are as those of table 43 of [5], i.e.,  $0 \leq \beta_1 \leq 1.8$  and  $0 \leq \beta_2 \leq 8.00$ . Although the program will work for values outside the ranges, the accuracy may be in doubt.

(2) Data Input

The program is in fixed point, with a four-six input and output, i.e., input is of the form xxx.xxxxx.

The input is of the following general arrangement.

<u>word</u>	<u>variable</u>
1	$\beta_1$
2	$\beta_2$
3	probability level
4	$X_\alpha$ (guessed value of the percentage point for probability $\alpha$ .)

---

\*Type IV will be added at a later date.

The input for Type I will be explained in more detail below. To obtain percentage points in the form as those given in the tables, word 3 must contain  $1-\alpha^*$ . If upper percentage points are desired, the guess,  $X_\alpha$ , must be read in negatively.  $X_\alpha$  is read in positive form for lower percentage points.

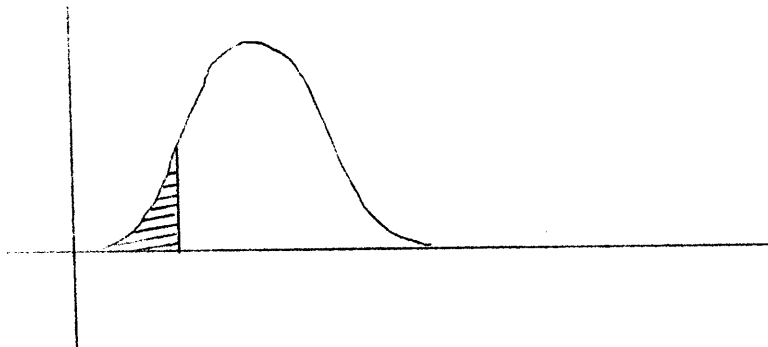
The Type VI distribution, however, follows standard notation. In this case word 3 must contain  $\alpha$  and word 4 the guessed percentage point, which is negatively for lower percentage points and positive otherwise.

For clarity, suppose it is desired to have the upper 5% point for  $\beta_1 = 0.10$  and  $\beta_2 = 2.6$ . This is a Type I distribution, hence, the input would read;

<u>word</u>	<u>variable</u>
1	0000.100,000
2	0002.600,000
3	0000.950,000
4	-0001.760,000 .

---

\*  $\alpha$  is considered as the shaded area in the following graph.



For an example of the Type VI input, assume the lower 2.5% point for  $\beta_1 = 1.80$  and  $\beta_2 = 6.60$  is desired. Hence, the card would read;

<u>word</u>	<u>variable</u>
1	0001.800,000
2	0006.600,000
3	0000.025,000
4	-0001.400,000.

If it is desired to have the probability associated with a given percentage point, then word 3 should contain zeros. Word 4 contains the given percentage point with algebraic sign attached according to the methods discussed above.

### (3) Output

The output has the same general form as the input except that word five contains the final  $X_\alpha$  (fourth decimal accurate). In obtaining the probabilities, word 3 of the output contains the probability associated with the  $X_\alpha$  that appears in word 4.

### (4) Machine Operation Instructions

The following is a list of instructions for the computer:

- (a) Use the standard 80-80 board.
- (b) Set console switches to 70 1951 xxxx.
- (c) Set programmed switch to stop.
- (d) Set overflow on sense.
- (e) Set error on stop.

- (f) Order of the cards
- (i) Pearson deck (blue)
  - (ii) Transfer card (red)
  - (iii) Data.

There are several program stops in this program. They are:

<u>Stops</u>	<u>Cause</u>
01 0000 1900	- Degrees of freedom become negative. Push start; will read next cd.
01 2000 1900	- $X_{\alpha}$ (for Type I distribution) is a bad guess. You may continue but this should not happen more than ten times in succession. If more than five seconds elapse between these stops, this indicates that the Newton Iteration technique fails to converge. At this point the yellow card should be added before the transfer card. This causes the computer to approach the desired results by shorter intervals. If one should get the same program stop, after about five seconds has elapsed since the previous stops, then one must add the extension (white deck), which then approaches the root by a different technique.

<u>Stops</u>	<u>Cause</u>
01 1966 0822	- $X_{\alpha}$ (for Type VI) is a bad guess. This is very unlikely if the guess was anywhere near the corrected result. Consult tables for possible error
-- ---- 7000	- Degrees of freedom for incomplete Beta becomes too large. Definitely STOP. Manual transfer to 1900 to read next card.
01 1998 1900	- The correction from the Newton Iteration technique becomes too large. Push start to continue to the next card.
-- ---- 4444	- A point has been chosen in Type IV region. <u>STOP</u> . Manual transfer to 1900.

(5) Comments

The program requires approximately one minute per value, points in the Type VI requiring more time than those points in Type I.

It should be noted that considerable computing time can be saved if the  $X_{\alpha}$  guess is close to the true value. Very good estimates of these values may be obtained from table 42 [5] or from tables presented here.

The program, number 6.6.011.1, is available from the Virginia Polytechnic Institute Computing Center, along with the instruction write-up.

## IX. BIBLIOGRAPHY

- [1] Bargmann, R. E. "A Study of Independence and Dependence in Multivariate Normal Analysis." Inst. of Statistics, University of North Carolina, Mimeo Series, No. 186, (1957).
- [2] Elderton, W.P. Frequency Curves and Correlation. Harren Press, (1953).
- [3] Kendall, M.G. and Stuart, A. The Advanced Theory of Statistics, Volume I. Charles Griffin and Company Limited, (1958).
- [4] Pearson, E. S. "Bayes Theorem, In the Light of Experimental Sampling." Biometrika, Volume 17, (1925), pp. 436 - 442.
- [5] Pearson, E.S., and Hartley, H.O. Biometrika Tables for Statisticians, Volume I. Cambridge University Press, (1956).
- [6] Pearson, E.S., and Merrington, M. "Tables of the 5% and 0.5% Points of Pearson Curves (with Argument  $\beta_1$  and  $\beta_2$ ) Expressed in Standard Measure." Biometrika, Volume 38, (1951), pp. 4 - 10.
- [7] Pearson, K. Early Statistical Papers, Cambridge University Press, (1956).
- [8] Pearson, K. Tables of the Incomplete Beta Function, Cambridge University Press, (1956).

- [9] Ruben, H. "On the Moments of Order Statistics in Samples from Normal Populations." Biometrika, Volume 41, (1954).

## X. ACKNOWLEDGEMENTS

The author wishes to express his deep appreciation to Professor Rolf E. Bargmann for his generous assistance, beneficial advice, and encouragement during the preparation of this thesis. He is especially grateful for the time and effort expended by Professor Bargmann in order to provide him with a workable incomplete-Beta program.

The author also thanks Professor Rudolf J. Freund for his assistance during the computational work involving the IBM computers.

The assistance of \_\_\_\_\_ in preparing the final manuscript for presentation is gratefully acknowledged.

This study was supported in part by a contract from the National Institutes of Health.

**The vita has been removed from  
the scanned document**

## ABSTRACT

This paper represents a report of the construction of extended tables of percentage points of the Pearson system of distributions. It consists of two parts:

- (1) The evaluation of the c.d.f. of any member of the Pearson system (except the so-called Type IV distribution) for a given value of  $X$  (standardized, i.e.,  $(x - \mu)/\sigma$ ) and for a given pair of  $\beta_1$  and  $\beta_2$ ; and
- (2) The determination of a percentage point  $X_\alpha$ , associated with a given level  $\alpha = 0.05, 0.025, 0.01, 0.005, 0.995, 0.99, 0.975, 0.95$ , for given values of  $\beta_1$  and  $\beta_2$ . The latter have been tabulated for  $\beta_1 = 0.00, 0.01, 0.03, 0.05, 0.10, 0.15, 0.20(.10)1.80$  and  $\beta_2 = 1.2(.2)6.6$ .

Through the use of high speed computers it was possible to expand the tables from their former status in accuracy as well as increasing the ranges of  $\beta_1$  and  $\beta_2$ .

Numerical illustrations are given which, in all cases studied, show close agreement of the Pearson approximation with the exact distribution.

The numerical analysis methods used for the evaluation of incomplete Beta Functions with continuous degrees of freedom is also described, since the latter is a key distribution used in computing the percentage points.

At the Virginia Polytechnic Institute Computing Center,  
the following programs are now available:

- 1) General Incomplete Beta,
- 2) c.d.f. of Pearson distributions,
- 3) Percentage points of Pearson distributions.

Instructions for the use of the last two programs are included  
in this report.