APPROXIMATION FORMULAS FOR THE INCOMPLETE BETA-FUNCTION

by

Niels Christian Andersen

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

June 1962

Blacksburg, Virginia

TABLE OF CONTENTS

I. INTRODUCTION

This chapter presents historical notes on the Incomplete Beta-Function pertaining to both its importance and methods of evaluating it. This leads to the desirability of finding expressions to approximate the function and the use that could be made of such expressions.

1.1 Review of the Literature and Associated Research

Methods for evaluating the partial area, up to the point θ, under the skew curve:

$$y = y_0 \, \theta^{p-1} (1 - \theta)^{q-1} \quad , \quad 0 \leq \theta \leq 1 \qquad (1)$$

have long been a problem to mathematicians. This area can be represented by the ratio of the Incomplete Beta-Function to the Complete Beta Function, i.e.,

$$I_\theta(p,q) = \frac{\int_0^\theta \theta^{p-1}(1 - \theta)^{q-1} d\theta}{\int_0^1 \theta^{p-1}(1 - \theta)^{q-1} d\theta} \quad , \quad 0 \leq \theta \leq 1 \, . \qquad (2)$$

Using Pearson's (1934) notation, $I_\theta(p,q)$ will, for clarification, be referred to as the Probability Integral.

The earliest known need for evaluating $I_\theta(p,q)$ was realized by Bayes (1763) as a direct result of his famous theorem in probability. Bayes was able, for small integer values of p and q, to expand the integrand in (2), integrate term by term, and thus successfully approximate the integral. Bayes was entirely unsuccessful for large p and q and little work was done to develop other methods until Karl Pearson showed that its successful solution or approximation was an important part of his analysis of skew frequency.

With the importance of evaluating $I_\theta(p,q)$ having increased, Wishart (1927) found approximations for large p and q through an extension of Bayes' work. His approximations were extremely difficult to use and they required the tabulation of numerous coefficients used in his formulas.

Pearson (1924) showed that the sum of the first p terms of the binomial expansion $(a + b)^n$ could be found by a simple transformation to $I_\theta(p,q)$. He also proved that the sum of n terms of a hypergeometric series could be approximated by the partial area under the curve (1). This was done by equating the first four moments of $I_\theta(p,q)$ and the hypergeometric distribution.

Thus, the importance of finding methods of readily and accurately evaluating the Probability Integral had been more than re-doubled. This led Pearson to undertake extensive investigations and these, along with earlier work, were put forth by Soper (1921). These studies led to other attempts with Müller (1930) having some degree of success in his application of continued fractions to the integral. However, due to the inaccuracies and/or the laboriousness of all these methods, Pearson (1934) found it necessary to compute and publish tables.

Pearson's tables give the values of the Complete Beta-Function and $I_\theta(p,q)$ in the following increments and ranges of θ, p and q :

$$\theta = .01, .02, .03, \ldots., .98, .99, 1.00$$

for p and $q \leq 11$

$$p,q = .5, 1.0, 1.5, \ldots., 10.0, 10.5, 11.0$$

and for p and $q \geq 11$

$$p,q = 12, 13, \ldots., 49, 50$$

with the restriction $p \geq q$. For desired values with $p < q$, use must be made of the relationship

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

With such relatively large increments in the range of θ (particularly for large p and q) and the difficulty in handling non-integer values of p and q (other than shown above), it becomes apparent that a large amount of interpolation would be required. This interpolation is extremely tedious since complicated and different methods are needed over the various ranges to achieve the desired accuracy.

Since the work of Pearson, the uses of $I_\theta(p,q)$ have greatly increased. The most noteworthy of these was made by Snedecor (1934). It is possible by a simple transformation to calculate percentage points of Snedecor's F distribution from $I_\theta(p,q)$, where F is the F statistic with probability element:

$$\frac{\Gamma(\frac{n_1+n_2}{2})\,(\frac{n_1}{n_2}F)^{\frac{n_1}{2}-1}}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})\,(1+\frac{n_1}{n_2}F)^{\frac{n_1+n_2}{2}}}\,d\,\frac{n_1}{n_2}F \quad , F > 0$$

with degrees of freedom n_1 and n_2. It can be readily shown, that

$$Pr(F > F_0) = I_\theta(p,q)$$

where

$$\theta = \frac{p}{p + q F_0} \quad , \quad p = \frac{n_1}{2} \quad \text{and} \quad q = \frac{n_2}{2} \quad .$$

Merrington and Thompson (1943) were the first to tabulate the F distribution. Their tables, and the numerous tables that followed, were computed basically by transforming Pearson's tables. All of these tables were tabulated only for various percentage points such as $Pr(F > F_0) = .50, .25, .10, .05$, etc. Thus, if an experimenter desired the exact probability, the existing F tables would not normally be useful since they would only allow him to state whether or not the calculated statistic was significant. However, since the simple transformation to $I_\theta(p,q)$ can be readily performed, this is not crucial. The problem lies in the before mentioned inadequacies of the increments of the tables since, more often than not, it would be necessary to interpolate.

1.2 The Need and Purpose of Approximation Formulas

If the research worker is able to compute the exact probabilities from several experiments testing the same null hypothesis, it is possible to pool these experiments. This

would be done using Pearson's (1933) $P\lambda$ criterion for random sampling and therefore make it possible to increase the usefulness of the experiments. It would be possible to combine several experiments conducted by different methods but testing the same null hypothesis. This pooling of experiments would be particularly advantageous in cases where it is only possible to perform small experiments or where results existed from previous research.

Therefore, it can readily be seen that there are many cases where it would be necessary to know the value of $I_\theta(p,q)$. This and the difficulty of interpolation in Pearson's tables are the primary reasons for finding approximation formulas for the Probability Integral. These formulas would enable the research worker to have a simple and accurate method of calculating $I_\theta(p,q)$ for given θ, p, and q regardless of their values. One other important reason is that Pearson's tables are very often not readily available to the research worker.

The usefulness of such expressions for various distributions was shown by James P. Ray (1961) in his development of approximation formulas to the cumulative normal and

t distributions. His work has been largely responsible for this attempt to find approximation expressions to $I_\theta(p,q)$ in that he showed the feasibility of such a project. As he points out, if the research worker had readily available such expressions for the major statistical distributions, then more accurate evaluation would be available and the reporting of non-significant results more practical.

These approximation formulas, to be useful, would have to provide a simple and accurate means of evaluating $I_\theta(p,q)$. It is obvious that normally these two requirements will, in a sense, be acting against each other and for both to be satisfied will necessitate a certain amount of division on the ranges of θ, p and q. This immediately brings forth another problem in that the greater the number of expressions, the greater complexity in their use and therefore the less their value. Thus, one very quickly sees the need for simplicity of computation conflicting with the imperative need for accuracy and both will tend to create further division of the ranges, which must be kept at a minimum.

II. DISCUSSION OF THE PROCEDURE FOR DEVELOPING APPROXIMATIONS

In this chapter two methods for developing approximation formulas are discussed and an outline is presented of the procedure to be followed. There is also an explanation of the basic programs used for the IBM 650 digital computer.

2.1 Method Proposed for Developing Approximation Formulas

As pointed out in Chapter I, approximation formulas of $I_\theta(p,q)$ would be very useful to the research worker. In form, these expressions would be very similar to those developed by Hastings (1954), however, it was felt that the least squares method was more desirable for the original expression as explained below. Hastings defined the best fit to a set of data to be that form which gives the smallest maximum deviation between the actual value and its approximate (predicted) value, i.e., the best fit is obtained when

$$\text{Maximum of } |(\text{actual value}) - (\text{approximate value})|$$

is minimized. The least squares method defined the best fit as that form for which the sum of the deviations squared is minimized, i.e., the best fit is obtained when

$$\Sigma[(\text{actual value}) - (\text{approximate value})]^2$$

is minimized. It was decided that we should first find an approximation formula by the least squares method and then, by weighting observations of maximum deviation, achieve the desired accuracy of the fit. Thus, in a sense, it was decided to employ a combination of the two methods (or properties) described above.

The method of investigation was to first find a functional form of some type of expansion that could give us some concept of the desired expression. Then by using this as a guide and by means of the IBM 650 computer and the programs available devise expressions necessary to fit $I_\theta(p,q)$ as functions of θ, p and q.

2.2 Programs for the IBM 650 Digital Computer

The Incomplete Beta Subroutine (Library #6.6.010.1) evaluates $I_\theta(p,q)$ for given values of θ, p, and q within the following ranges:

$$0.000001 \leq p,q \leq 9999.999999$$

$$.000000 \leq \theta \leq 0001.000000 \quad .$$

This program was written at V.P.I. by Dr. Rolf Bargmann and calculates the Probability Integral to six places.

The Revised General Multiple Regression System (Library #6.2.008.1,2,3,4,5) was written at North Carolina State College and is divided into five parts. The division of the program not only greatly increases the size of problems that can be handled but also provides great flexibility. This program finds the least squares linear relationships of the p dependent variables, y_i, (i=1,...,p) and the n independent variables, x_j, (j=1,..., n) where

$$y_i = b_{0i} + b_{1i}x_1 + b_{2i}x_2 + \ldots + b_{ni}x_n \quad .$$

Part I unpacks and performs simple transformations on previously punched data. Part II computes the sums and the uncorrected sums of squares and cross products of the p dependent and n independent variables. The input is in single precision and must be scaled so that the output will have not more than 10 digits for the sums or more than 20 digits for the sums of squares and cross products.

Part III corrects for the mean, if desired, and converts the normal equation matrix to single precision floating point and then inverts this matrix. The solutions (regression coefficients) to the equations are calculated with the provision that variables may be deleted as desired. It is

possible to sum or pool two or more matrices of the same size for inversion. Part IV performs the same operations as Part III for larger matrices, but it is not self-restoring - i.e., it must be re-loaded for each matrix.

Part V computes the sum of squares due to regression and error, and their mean squares. It also computes R^2, the variance of each regression coefficient and the related t^2. The above values are computed for each dependent variable. It is also possible to compute predicted values $\hat{y}_i$ and deviations from the observations $(y_i - \hat{y}_i)$. After examining the output of Part V, it is possible to make an intelligent guess as to which variables can be deleted with the smallest effect on the fit. These variables can be deleted by simply returning to Part III.

The Least Squares Polynomial Program (Library #6.0.006.1) was written at the Oklahoma State University Computing Center. This program finds equations of polynomials of best fit by the least squares method. It fits equations of degree one thru four and for each equation computes the means, the total, regression and error sums of squares and their associated mean square. It also computes R and the sum of the deviations $\Sigma(y_i - \hat{y}_i)$ together with the F test for

goodness of fit and, if desired, the predicted values. While the data is entered in single precision, all internal computations are performed in double precision.

The Successive Partial Correlations and Partial Regression Coefficients Program (Library #6.2.006.1) was also written at North Carolina State College. This program computes the correlation coefficients between each of the variables and their associated regression coefficients. Thus it is possible to study the output for simple (i.e. - linear) regression and pick the independent variable that has the highest correlation with the dependent variable. Then, including this independent variable as a predictor, the process is repeated and the correlations again examined. This method is repeated until the desired number of independent variables are included as predictors of the dependent variable.

The output of Part III (6.2.000.3) can be used for input and it is possible to pool several matrices prior to the first cycle of computation. This program is also extremely helpful in the case that a matrix is near singular due to high correlation between independent variables. In this case, it is an easy matter to recognize the variables to delete.

The Double Precision Matrix Inversion Subroutine (Library #5.2.009) was developed by the IBM Corporation and by means of the Gaussian elimination method provides for both matrix inversion and solutions of simultaneous linear equations.

The input matrix is in floating point and can either be in single or double precision. Regardless of the input, all computations are performed in double precision and the output contains the solutions and the inverse matrix. If necessary, the program can be stopped and punched out at various stages of computation and re-started later. Although necessarily slow, this program was found extremely accurate in its computations.

III. DEVELOPMENT OF APPROXIMATION FORMULAS

In this chapter we show an expansion of $I_\theta(p,q)$ which is used as a guide in the development of approximation formulas. The second section deals with the investigations and the approximation expression derived.

3.1 Derivation of an Expansion of $I_\theta(p,q)$

After an extensive review of the literature, as outlined in Chapter I, it was decided that, owing to the complexity of existing methods of evaluation, no available form was applicable as a guide to find approximation formulas. Therefore several methods of integration of $I_\theta(p,q)$ were considered in order to obtain some idea of a satisfactory functional form. The best that was found for our purposes is the following method.

We expanded $(1-\theta)^{q-1}$ in a Maclaurin series, which gives:

$$(1-\theta)^{q-1} = 1 - (q-1)\theta + \frac{(q-1)(q-2)}{2!}\theta^2 - \ldots + (-1)^k \frac{(q-1)(q-2)\cdots(q-k)}{k!}\theta^k + \ldots \quad (4)$$

We note from Franklin (1940) that this series converges for $|\theta| < 1$ to $(1-\theta)^{q-1}$. Substituting (4) into $\beta_\theta(p,q)$, integrating and dividing by $\beta(p,q)$ we have

$$I_\theta(p,q) = \frac{\theta^p}{\beta(p,q)} \left\{ \frac{1}{p} - \frac{(q-1)}{(p+1)}\theta + \frac{(q-1)(q-2)}{2!(p+2)}\theta^2 \right. \tag{5}$$
$$\left. - \ldots + (-1)^k \frac{(q-1)(q-2)\cdots(q-k)}{p+k}\theta^k + \ldots \right\} .$$

A series expansion of this type has the convenience of converging fairly rapidly for θ small (say $\leq \frac{1}{2}$) and use could be made of the relationship

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

for larger θ .

3.2 Investigation of Various Expressions and Results

Since the numerical analysis had suggested that we try to fit an equation of the general form of (5), it was decided to first break this down into three parts. If possible, approximations could be found for θ^p and $\frac{1}{\beta(p,q)}$ in various ranges, then combine these with the third part, i.e.,

$$\left(\frac{1}{p} - \frac{(q-1)}{(p+1)}\theta + \frac{(q-1)(q-2)}{2!(p\ 2)}\theta^2 - \ldots \right) .$$

Finally, by means of the multiple regression program, delete all unnecessary terms to produce the desired expression.

It was found possible to fit θ^p over various ranges of p and θ in polynomials of $\frac{\theta}{p^2}$. Although the ranges of p and θ were as small as considered practical so as to not result in too many formulas, no expressions could be found by which it was possible to fit $\frac{1}{\beta(p,q)}$ over similar ranges. This attempt lead to two immediate results: 1) it was extremely difficult to fit the reciprocal of the Complete Beta-Function even for small ranges and 2) it appeared that this method, if feasible, would necessitate a prohibitive number of formulas.

Therefore it was decided to return to (5) and, again using it as a guide, investigate the entire function. Using the Incomplete Beta-Function Subroutine to generate the various observations on $I_\theta(p,q)$ and examining the lower ranges on θ, it was found possible to fit $I_\theta(p,q)$ for $.00 \leq \theta \leq .25$. This was done by holding p and q fixed and using the multiple regression program for a 6^{th} degree polynomial of θ, i.e.,

$$Y_{i\theta|p,q} = b_{0i} + b_{1i}\theta + b_{2i}\theta^2 + b_{3i}\theta^3 + b_{4i}\theta^4 + b_{5i}\theta^5 + b_{6i}\theta^6 \qquad (6)$$

$\theta = .01, .02, \ldots, .24, .25$

where $Y_{i\theta|p,q}$ is the i^{th} observation on $I_\theta(p,q)$ varying θ within the range and holding p and q fixed. With the restriction of $p \geq q$ and fitting $I_\theta(p,q)$ for all values $\geq .0004$, we found that p and q were never greater than 20 and we imposed the lower limit of $p,q \geq 2$. For all possible integer values of p and q, with the above restrictions, it was found that (6) was a good fit and that the deviations, i.e., $(Y_i - \hat{Y}_i)$ were less than $\pm .00003$.

We now attempted to fit the regression coefficients obtained from (6) as functions of p (or q) holding q (or p) fixed, i.e.

$$b_{ji\theta p|q} = f_1(p) \quad \ldots\ q \text{ fixed}$$

or

$$b_{ji\theta q|p} = f_2(q) \quad \ldots\ p \text{ fixed}$$

$j = 0, 1, 2, 3, 4, 5, 6$.

Various functional forms were attempted by means of the O.S.U. Polynomial Program but no relationships were found with the desired accuracy.

It was decided that the inability to perform the above was mainly due to the rapid changes that occurred to $I_\theta(p,q)$ in the lower ranges of θ. Thus it was felt that if the above procedure was repeated in a smoother area of

the surface, it might be possible to determine the necessary functional relationships and, after fitting the curve in that area, return to the lower ranges of θ.

Changing the range on θ to $.30 \leq \theta \leq .40$, it was seen that a 6^{th} degree polynomial would again give the desired accuracy. Repeating the procedure given above, (6) was fitted by the multiple regression program for all integer values of p and q within the range $10 \leq p, q \leq 20$. Note that the restriction $p \geq q$ has been dropped. It was found possible to fit the regression coefficients in a 4^{th} degree polynomial in $\frac{1}{p}$ holding q fixed. That is, for $10 \leq p$, $q \leq 20$

$$b_{ji\theta p|q} = c_{0i} + c_{1i}(\tfrac{1}{p}) + c_{2i}(\tfrac{1}{p})^2 + c_{3i}(\tfrac{1}{p})^3 + c_{4i}(\tfrac{1}{p})^4$$

$j = 0, 1, 2, 3, 4, 5, 6$

$\theta = .30, .31, \ldots, .40$ $\quad$ q fixed

$p = 10, 11, \ldots, 20$.

Since this area was considered too small to be useful, the range was extended on p and q to $4 \leq p, q \leq 48$. The multiple regression for equation (6) was run for all integer combinations of p and q that were multiples of 4. This again produced a satisfactory fit for θ, but upon

attempting to fit these regression coefficients as a polynomial in $\frac{1}{p}$ holding q constant, we found that the coefficients changed too rapidly and erratically to be fit. Upon investigating it was found that the earlier restriction on p and q, namely $10 \leq p,q \leq 20$, had enabled us to fit p. Consequently we were forced to re-evaluate our attempts.

The difficulty was found to lie in the multiple regression program or, more precisely, in the normal equation matrix inversion routine. Although our fit was good, i.e., $(Y_i - \hat{Y}_i) \leq \pm .00003$, there existed near singularity in the matrix due to the high correlation between the independent variables. Since the inversion routine, which performs all computation in single precision, was very inaccurate for these models, it was not possible to fit the regression coefficients because their variations were extremely large and erratic.

Further examination, using the O.S.U. Polynomial Program (which performs computations in double precision), showed that this was our problem. Continuing the investigation with this program, it was seen that for various ranges of

p and q it was possible to fit $I_\theta(p,q)$ as a polynomial in p (or q) holding θ and q (or p) fixed. Thus for the range $1 \leq p \leq 5$, $5 \leq q \leq 10$ and $.01 \leq \theta \leq .35$ (using similar notation to (6))

$$Y_{iq|\theta p} = d_{0i} + d_{1i}q + d_{2i}q^2 + d_{3i}q^3 + d_{4i}q^4 \qquad (7)$$

$q = 5, 6, \ldots, 10$ θ and p fixed

was found to be an accurate approximation. For $\theta \geq .35$ and over the same ranges of p and q, it was found that a satisfactory fit could be obtained by

$$Y_{ip|\theta q} = e_{0i} + e_{1i}p + e_{2i}p^2 + e_{3i}p^3 + e_{4i}p^4 \qquad (8)$$

$p = 1, 1.5, 2.0, \ldots, 5.0$ θ and q fixed .

Within these ranges on p and q it was indicated by the O.S.U. Polynomial Program that (6) was an adequate approximation holding p and q fixed. We now proceeded by trying to fit the product of (6) times (7) and (6) times (8). In the first case, for θ between .00 and .35, we were able to fit the product of the two functions. Recognizing the previously mentioned inaccuracies of the multiple regression inversion routine, the matrix was first examined by means of the Successive Partial Correlation Program and

the more highly correlated independent variables were deleted. Again using this program, we were able to "build up" the expression, by successively including the independent variables as predictors of the dependent variable, to a model of 10 variables.

Examining this formula by means of the Multiple Regression Program, it was found that we had an excellent fit. Thus, for p fixed, we had that

$$Y_{i\theta q|p} = f_{0i} + f_{1i}\theta + f_{2i}q\theta + f_{3i}q\theta^2 + f_{4i}q^2\theta^2 + f_{5i}q^2\theta^2 + f_{6i}q^3\theta^2 + f_{7i}q^3\theta^3 + f_{8i}q^3\theta^4 + f_{9i}q^4\theta^4 + f_{10i}q^4\theta^5 \qquad (9)$$

$$\theta = .05, .10, \ldots, .35$$

$$q = 5, 6, \ldots, 10$$

produced deviations $(Y_i - \hat{Y}_i) \leq \pm .00002$.

This process was unsuccessfully repeated on the product of (6) and (8). It was felt that this was caused by the larger values of θ which resulted in less change in the magnitude of the powers of θ. This increased the correlation of the independent variables which pushed the matrix beyond the accuracy of the programs we were using.

Investigating the region that we had used for (9), it was found possible to approximate $I_\theta(p,q)$, holding θ and q constant, in a polynomial of the reciprocal powers of p. Using this as a guide we were able to approximate the Probability Integral, after decreasing the range on θ, by combining (9) with a 6^{th} degree polynomial of $\frac{1}{p}$. After deleting terms and weighting the observations of maximum deviation, we obtained the following approximation formula:

$$I_\theta(p,q) \doteq \frac{c_1 + c_2\theta + c_3q\theta^2 + c_4q^2\theta^3 + c_5q^3\theta^4 + c_6q^4\theta^5}{1.0 + c_7p + c_8p^2 + c_9p^4 + c_{10}p^5 + c_{11}p^6} \qquad (10)$$

$$.025 \leq \theta \leq .200$$

$$1.0 \leq p \leq 5.0$$

$$5.0 \leq q \leq 10.0$$

where

$c_1 = .000443$	$c_7 = -1.905371$
$c_2 = .038396$	$c_8 = 1.012196$
$c_3 = -.020517$	$c_9 = -.142526$
$c_4 = .028508$	$c_{10} = .038866$
$c_5 = -.016338$	$c_{11} = -.003146$
$c_6 = .003310$.	

The differences between the approximate values and the actual values used for fitting do not exceed $\pm$ .000098, and the value when rounded to four decimal places does not differ by more than one unit in the fourth decimal place.

See Appendix A for a tabulation of the actual values minus the approximate values.

IV. COMPUTATIONAL PROCEDURES AND EXAMPLES

In this chapter we will illustrate the procedure, using a desk calculator, to compute $I_\theta(p,q)$ from the approximation derived (10) and, through the use of an example, compute the Probability Integral. An example is also given showing a method of pooling several experiments by Pearson's (1933) $P\lambda$ criterion.

4.1 Outline of Computational Procedures

To approximate $I_\theta(p,q)$ for any values of p, q and θ within the ranges

$$.025 \leq \theta \leq .200$$

$$1.0 \leq p \leq 5.0$$

$$5.0 \leq q \leq 10.0 \quad ,$$

and following the expression developed in (10), the steps involved using a desk calculator are as follows:

(1) Enter (- .003146)

(2) Multiply (1) by p

(3) Add (.038,866) to (2)

(4) Multiply (3) by p

(5) Subtract (.142526) from (4)

(6) Multiply (5) by p

(7) Multiply (6) by p

(8) Add (1.012196) to (7)

(9) Multiply (8) by p

(10) Subtract (1.905371) from (9)

(11) Multiply (10) by p

(12) Add 1.0 to (11) and retain this value

(13) Enter q

(14) Multiply (13) by θ and retain

(15) Multiply (14) by (.003310)

(16) Subtract (.016338) from (15)

(17) Multiply (16) by (14)

(18) Add (.028508) to (17)

(19) Multiply (18) by (14)

(20) Subtract (.020517) from (19)

(21) Multiply (20) by (14)

(22) Add (.038396) to (21)

(23) Multiply (22) by θ

(24) Add (.000443) to (23)

(25) Divide (24) by (12) to give the desired approximate value of $I_{\theta}(p,q)$.

Sufficient accuracy will be maintained by carrying p, q, θ and all computations to six decimals.

4.2 Example of Computing Procedure

Suppose laboratory A had performed a One-Way Classification Analysis to test the mean of 6 treatments with unequal observations in each block. The following Analysis of Variance table was computed:

ANOVA TABLE - LABORATORY A

Source	d.f.	S.S.	M.S.	F
Treatment	5	432.00	86.40	2.01
Error	15	644.85	42.99	
Total	20	1,076.85		

The calculated F statistic is reported only as non-significant at the 5% level. Thus, the only conclusion that can be drawn is that the treatments are similar. Since there were a relatively small number of observations, the research worker might be interested in learning the exact significance level to ascertain whether or not continued research was necessary.

Therefore, if the research worker desired the exact probability, it could be computed, using the transformations described in Chapter II, as follows:

$$p = \frac{5}{2} = 2.5 \qquad q = \frac{15}{2} = 7.5$$

$$\theta = \frac{2.5}{2.5 + 7.5(2.01)} = .142,248$$

(1) (-.003,146)

(2) (-.003,146) x (2.5) = -.007,865

(3) (-.007,865) + (.038,866) = .031,001

(4) (.031,001) x (2.5) = .077,503

(5) (.077,503) - (.142,526) = -.165,023

(6) (-.065,023) x (2.5) = -.162,558

(7) (-.162,558) x (2.5) = -.406,395

(8) (-.406,395) + (1.012,196) = .605,801

(9) (.605,801) x (2.5) = 1.514,503

(10) (1.514,503) - (1.905,371) = -.390,868

(11) (-.390,868) x (2.5) = -.977,170

(12) (-.977,170) + (1.0) = .022,830

(13) (7.5)

(14) (7.5) x (.142,248) = 1.066,860 = q θ

(15) (1.066,860) x (.003,310) = .003,531

(16) (.003,531) - (.016,338) = -.012,807

(17) (1.066,860) x (-.012,807) = -.013,663

(18) (-.013,663) + (.028,508) = .014,845

(19) (.014,845) x (1.066,860) = .015,838

(20) (.015,838) - (.020,517) = -.004,679

(21) (-.004,679) x (1.066,860) = -.004,992

(22) (-.004,679) + (.038,396) = .033,404

(23) (.033,404) x (.142,248) = .004,752

(24) (.004,752) + (.000,443) = .005,195

(25) .005,195 ÷ .022,830 = .227,551

$$I_{.142,248}(2.5,\ 7.5) \doteq .227,551$$

Actual value of $I_{.142,248}(2.5,\ 7.5) = .227,618$

Difference = .000,067 .

Therefore the experimenter would know that the calculated F statistic was significant at the 22.75% level. He might feel that this warranted further research, which could be performed by pooling the above results with those of other experiments testing the same hypothesis.

It should also be noted in the above example, that we calculated the exact probability from the F statistic, which is normally the method of presenting such results. However, if we knew beforehand that we desired the exact level of

significance, this could be computed from the Treatment and Error sums of squares, i.e.,

$\frac{\text{Treatment SS}}{\text{Treatment SS + Error SS}}$ is distributed as $I_\theta(p,q)$.

This procedure eliminates the calculation of the mean squares and the F statistic.

4.3 Example of Pooling Several Experiments

Suppose the research worker had available the following Analysis of Variance tables testing the same hypothesis as tested in the previous example (4.2).

ANOVA TABLE - LABORATORY B

Source	d.f.	S.S.	M.S.	F
Treatments	5	44.80	8.96	3.18
Error	20	56.40	2.82	
Total	25	101.20		

ANOVA TABLE - LABORATORY C

Source	d.f.	S.S.	M.S.	F
Treatments	5	591.65	118.33	2.92
Error	12	486.24	40.52	
Total	17	1,077.89		

The F statistics calculated from both analyses were found to be non-significant at the 5% level, as was the case in the results from laboratory A. Following the same procedure outlined in 4.2, we computed the probabilities, which, along with the probability from laboratory A are:

Laboratory	$Pr(F \geq F_0)$
A	.228
B	.106
C	.294

Combining these results by means of Pearson's $P\lambda$ criterion, we have:

$$P\lambda = -2 \sum_{i=1}^{3} \log_e P_i$$

[where $P_i = Pr(F > F_0)$; $i = 1, 2, 3$]

$$= -2 \log_e 10 \sum_{i=1}^{3} \log_{10} P_i$$

$$= -2(2.3026)(\bar{1}.35698 + \bar{1}.02653 + \bar{1}.24229)$$

$$= -4.6052(-6.3742)$$

$$= 29.35 \quad .$$

$P\lambda$ is distributed as a χ^2 with 2 x 3 degrees of freedom. Since a χ^2 equal to 29.35 with 6 degrees of freedom is significant at the 5% level, we reject the null hypothesis. Thus, on the basis of combined results, we conclude that the treatments are not all the same.

V. CONCLUSIONS AND SUGGESTIONS FOR ADDITIONAL RESEARCH

The desirability of finding approximation formulas for $I_\theta(p,q)$ is obviously not decreased by the author's difficulty in fitting the function over a larger area. This thesis, other than the area fitted, has pointed out some of the problems that must be solved before adequate approximations over a large range can be developed.

It is very strongly felt that the method of first fitting $I_\theta(p,q)$ as a function of one variable, then fitting these regression coefficients for a second variable and finally fitting the second set of regression coefficients as a function of the third variable is not feasible, with the equipment available, for two reasons. These are 1) this method would require a matrix inversion routine of such accuracy that the time involved would be prohibitive and 2) the resulting approximating expressions would be too large and tedious to compute. The 6^{th} degree polynomial in θ was found adequate for all ranges investigated and, using an inversion routine of higher precision, it is felt that attempts should be made to fit these regression coefficients

as expressions in p and q together. As pointed out above it is not considered practical to fit these coefficients as functions of one variable alone.

The author feels that the degree of success obtained in fitting the areas that he did was due to the method of fitting a ratio of expressions. This means that if two of the variables can be fitted, it is very possible that an expression can be found for the third variable and a final formula developed as the ratio of the two forms.

A third method to attempt would be to fit two of the variables as a polynomial of their quotient or product. For $1 \leq p \leq 5$, $5 \leq q \leq 10$ and $\theta \geq .20$, it was found possible to adequately fit $I_\theta(p,q)$ in a 10^{th} degree polynomial of $\frac{q}{p}$. A polynomial of this form had been attempted at various ranges, but was not successful until the Double Precision Matrix Inversion Program was used. Other combinations could be tried over various ranges such as $q\,\theta$ and $\frac{\theta}{p}$. Then the remaining variable could be fitted either as a function of the regression coefficients or as a ratio of expressions.

Therefore, it can be summed up that there are several definite approaches to be tried within this general method.

Two things would be of prime importance, with the first necessitating the second, which are 1) a high precision matrix inversion routine and 2) the availability of a high speed digital computer.

VI. BIBLIOGRAPHY

Bayes, Thomas (1763), "An Essay Towards Solving a Problem in the Doctrine of Chances", Phil. Trans., Vol. LIII, pg. 370 et seq.

Franklin, P. (1940), Treatise on Advanced Calculus, J. Wiley and Sons, Inc., New York.

Hastings, C. Jr. (1954), Approximations for Digital Computers, Princeton University Press, Princeton, New Jersey.

Merrington, Maxine and Thompson, Catherine M. (1943), "Tables of Percentage Points of the Inverted Beta (F) Distribution", Biometrika, Vol. 33, pg. 73-78.

Müller, J. H. (1930), "The Application of Continued Fractions", Biometrika, Vol. 22, pg. 284-297.

Pearson, Karl (1924), "Relationship of the Incomplete Beta-Function to the Binomial, $(a+b)^n$ ", Biometrika, Vol. 16, pg. 202-203.

Pearson, Karl (1933), "General Criterion for Random Sampling", Biometrika, Vol. 25, pg. 379-410.

Pearson, Karl (1934), Tables of the Incomplete Beta-Function, Cambridge University Press, London.

Ray, James P. (1961), Approximations to the Cumulative Normal and t Distributions, V.P.I., Blacksburg, Va.

Snedecor, G. W. (1934), Calculation and Interpretation of Analysis of Variance and Covariance, Collegiate Press Inc., Ames, Iowa.

Soper, Herbert E. (1921), Tracts for Computers, No. VII, Cambridge University Press, London.

Wishart, John (1927), "Approximate Quadrature of Certain Skew Curves," Biometrika, Vol. XIX, pg. 1-38.

VII. ACKNOWLEDGEMENTS

The author would like to express his sincere gratitude to Professor David C. Hurst for the guidance during the course of this study. Without his advice and assistance, this thesis would not have been written.

Appreciation is extended to Professor Whitney L. Johnson who kindly consented to read this thesis in manuscript form. The author is also deeply indebted to Dr. Boyd Harshbarger for his advice, encouragement and counsel.

To the author is indebted for the preparation of the final typewritten copies.

The author also wishes to express his gratitude for the financial assistance received from a National Institutes of Health fellowship.

Gratitude is extended to the author's wife for her assistance, encouragement and patience.

The vita has been removed from the scanned document

APPENDIX A

Table of [actual value - approximate value] x [10^6]

θ	p	q					
		5	6	7	8	9	10
.025	1	19	-07	00	04	-11	-16
	2	20	18	11	-02	02	-28
	3	18	42	-08	19	-01	-25
	4	13	16	02	14	07	14
	5	32	21	00	-04	15	-07
.050	1	46	-05	-09	-06	-17	-25
	2	36	12	-08	-09	-35	-57
	3	27	38	04	-13	-45	-54
	4	37	36	-12	-01	-26	-31
	5	41	16	-00	02	-18	-41
.075	1	-01	-16	-29	-56	-25	-44
	2	11	-14	-08	-34	-54	-17
	3	08	-33	-02	-32	-42	-53
	4	25	-46	-30	-52	-52	-48
	5	-22	-05	-08	-56	-49	-32
.100	1	07	-16	-48	-10	-26	-13
	2	-15	-32	-61	-43	-30	-02
	3	-16	-42	-59	-51	-11	-01
	4	-05	-55	-09	-27	-39	-45
	5	-21	-23	-56	-43	-20	-11
.125	1	-26	-28	-01	19	32	02
	2	-24	-38	-09	-15	-14	13
	3	-32	-20	12	-24	-01	13
	4	-22	-21	-01	-32	21	-21
	5	-24	-39	-15	-46	06	11

Appendix A - Table (continued)

θ	p	q					
		5	6	7	8	9	10
.150	1	01	-27	28	32	46	18
	2	-31	01	-08	-03	06	-12
	3	-44	00	07	33	25	29
	4	-44	-38	17	10	00	-21
	5	-48	-09	10	18	07	-01
.175	1	-10	12	38	40	-06	-17
	2	-27	12	43	16	-14	-22
	3	15	35	21	44	05	24
	4	-24	04	30	02	03	-15
	5	-19	32	20	-10	14	-03
.200	1	25	41	40	16	15	70
	2	22	25	17	01	-14	44
	3	19	26	29	22	28	96
	4	30	05	13	27	-36	83
	5	-04	58	17	11	00	98

ABSTRACT

The Incomplete Beta-Function is one of the most widely used statistical distributions, either directly or by means of simple transformations to other distributions. It is very often useful to calculate its value and this can be done by Pearson's (1934) tables. However, Pearson's tables have rather large increments in the three variables and interpolation is often required, which is at best tedious and time consuming.

It was felt that this problem could best be solved by approximation formulas. Using an IBM 650 digital computer and various programs available, attempts were made to fit the Incomplete Beta-Function. An adequate expression was derived for limited ranges on the three variables and some of the problems exposed that must be solved before similar formulas can be developed for a larger area.