is sometimes substituted for the words "at least." Of course we can express all of them in terms of the $p[(\nu^r)]$'s or of the $p((\nu^r))$'s. However elegant formulas such as in the ordinary theory seem to be lacking.

Finally, we may also consider conditions of existence for the $p[(\nu^r)]$'s and the $p((\nu^r))$'s. For the former system the conditions are that they be all non-negative and that their sum be 1. For the latter system, the conditions are given by (4′), viz. for every $(\nu^r) \epsilon (\nu^\lambda)$,

$$\sum_{(\alpha') \epsilon (\nu^{\lambda-r})} \mu((\alpha'))p((\nu^r) + (\alpha)) \geqq 0.$$

These conditions are necessary and sufficient since (3) and (4) are equivalent.

ON THE MECHANICS OF CLASSIFICATION

By Carl F. Kossack

University of Oregon

1. Introduction. Wald[1] has recently determined the distribution of the statistic U to be used in the classification of an observation, z_i $(i = 1, 2, \cdots, p)$, as coming from one of two populations. He also determined the critical region which is most powerful for such a classification. It is the purpose of this paper to show how such a classification statistic under the assumption of large sampling can be applied in an actual problem and to present a systematic approach to the necessary computations.

The data used in this demonstration are those which were obtained from the A.S.T.P. pre-engineering trainees assigned to the University of Oregon. The problem considered is that of classifying a trainee as to whether he will do unsatisfactory or satisfactory work[2] in the first term mathematics course (Intermediate Algebra). The variables used in the classification are: (1) A Mathematics Placement Test Score. This is the score obtained by the trainee on a fifty-minute elementary mathematics test (including elementary algebra). The test was given to each trainee on the day that he arrived on the campus. (2) A High School Mathematics Score. A trainee's high school mathematics record was made into a score by giving 1 point to students who had had no high school algebra, 2 points to students with an F in first-year, high-school algbra and no second-year algebra, 3 points for a D, $\cdots$, 10 points for an average grade of A in first- and second-year algebra. (3) The Army General Classification Test Score. An individual needed a score of 115 or better in order to be assigned to the A.S.T.P. These data were obtained for 305 trainees along with the actual

[1] Abraham Wald, "On a statistical problem arising in the classification of an individual into one of two groups," *Annals of Math. Stat.*, Vol. 15, (1944), No. 2.

[2] Unsatisfactory work was defined as a grade of F or D in the course (failure or the lowest passing grade).

grade made by them in the algebra course. Trainees who had had college work were not included in the study.

2. Steps in the Computation of U and the Critical Region. Let

π_1 be the population of individuals who do unsatisfactory work in their first-term mathematics course.

π_2 be the population of individuals who do satisfactory work.

N_1 and N_2 = respectively the number of observed individuals in π_1 and π_2.

$x_{1\alpha}$ and $y_{1\alpha}$ = respectively the Mathematics Placement Test Score for the αth individual observed in π_1 and π_2.

$x_{2\alpha}$ and $y_{2\alpha}$ = respectively the High School Mathematics Score.

$x_{3\alpha}$ and $y_{3\alpha}$ = respectively the Army General Classification Test Score.

Step 1. *Computation of Summations*

$N_1 = 96$	$N_2 = 209$
$\sum_\alpha x_{1\alpha} = 3570$	$\sum_\alpha y_{1\alpha} = 11450$
$\Sigma x_{2\alpha} = 547$	$\Sigma y_{2\alpha} = 1567$
$\Sigma x_{3\alpha} = 11745$	$\Sigma y_{3\alpha} = 26684$
$\Sigma x_{1\alpha}^2 = 145476$	$\Sigma y_{1\alpha}^2 = 672452$
$\Sigma x_{2\alpha}^2 = 3509$	$\Sigma y_{2\alpha}^2 = 12577$
$\Sigma x_{3\alpha}^2 = 1439559$	$\Sigma y_{3\alpha}^2 = 3421996$
$\Sigma x_{1\alpha}x_{2\alpha} = 21012$	$\Sigma y_{1\alpha}y_{2\alpha} = 88774$
$\Sigma x_{1\alpha}x_{3\alpha} = 436964$	$\Sigma y_{1\alpha}y_{3\alpha} = 1469302$
$\Sigma x_{2\alpha}x_{3\alpha} = 66731$	$\Sigma y_{2\alpha}y_{3\alpha} = 200150$
$\Sigma(x_{1\alpha} - \bar{x}_1)^2 = 12716.625$	$\Sigma(y_{1\alpha} - \bar{y}_1)^2 = 45167.311$
$\Sigma(x_{2\alpha} - \bar{x}_2)^2 = 392.240$	$\Sigma(y_{2\alpha} - \bar{y}_2)^2 = 828.249$
$\Sigma(x_{3\alpha} - \bar{x}_3)^2 = 2631.656$	$\Sigma(y_{3\alpha} - \bar{y}_3)^2 = 15125.876$
$\Sigma(x_{1\alpha} - \bar{x}_1)(x_{2\alpha} - \bar{x}_2) = 670.438$	$\Sigma(y_{1\alpha} - \bar{y}_1)(y_{2\alpha} - \bar{y}_2) = 2926.392$
$\Sigma(x_{1\alpha} - \bar{x}_1)(x_{3\alpha} - \bar{x}_3) = 196.812$	$\Sigma(y_{1\alpha} - \bar{y}_1)(y_{3\alpha} - \bar{y}_3) = 7427.359$
$\Sigma(x_{2\alpha} - \bar{x}_2)(x_{3\alpha} - \bar{x}_3) = -191.031$	$\Sigma(y_{2\alpha} - \bar{y}_2)(y_{3\alpha} - \bar{y}_3) = 83.837$

Step 2. *Computation of Statistics.*

$\bar{x}_1 = 37.188$	$\bar{y}_1 = 54.785$
$\bar{x}_2 = 5.6979$	$\bar{y}_2 = 7.4976$
$\bar{x}_3 = 122.3438$	$\bar{y}_3 = 127.6746$

$$s_{ij} = \frac{\Sigma(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) + \Sigma(y_{i\alpha} - y_i)(y_{j\alpha} - y_j)}{N_1 + N_2 - 2}$$

$s_{11} = 191.04$	$s_{12} = 11.871$
$s_{22} = 4.0280$	$s_{13} = 25.162$
$s_{33} = 58.606$	$s_{23} = -.35378$

Step 3. *Computation of Inverse Matrix* $|s^{ij}|$

$$|s_{ij}| = \begin{vmatrix} 191.04 & 11.871 & 25.162 \\ 11.871 & 4.0280 & -.35378 \\ 25.162 & -.35378 & 58.606 \end{vmatrix} = 34053$$

$$\begin{aligned} s^{11} &= .0069286 & s^{12} &= -.020692 \\ s^{22} &= .31019 & s^{13} &= -.0030996 \\ s^{33} &= .018459 & s^{23} &= .010756 \end{aligned}$$

Step 4. *Computation of the Classification Equation.*

$$\begin{aligned} U = {} & [s^{11}(\bar{y}_1 - \bar{x}_1) + s^{12}(\bar{y}_2 - \bar{x}_2) + s^{13}(\bar{y}_3 - \bar{x}_3)]\cdot z_1 \\ & + [s^{21}(\bar{y}_1 - \bar{x}_1) + s^{22}(\bar{y}_2 - \bar{x}_2) + s^{23}(\bar{y}_3 - \bar{x}_3)]\cdot z_2 \\ & + [s^{31}(\bar{y}_1 - \bar{x}_1) + s^{32}(\bar{y}_2 - \bar{x}_2) + s^{33}(\bar{y}_3 - \bar{x}_3)]\cdot z_3 \end{aligned}$$

where z_i plays the same role for individuals to be classified as $x_{i\alpha}$ and $y_{i\alpha}$ do for observed individuals.

$$U = .068160\, z_1 + .25147\, z_2 + .063215\, z_3$$

Step 5. *Computation of the Critical Region* (*assuming* $W_1 = W_2$)

$$\begin{aligned} \bar{a}_1 &= .068160\, \bar{x}_1 + .25147\, \bar{x}_2 + .063215\, \bar{x}_3 = 11.702 \\ \bar{a}_2 &= .068160\, \bar{y}_1 + .25147\, \bar{y}_2 + .063215\, \bar{y}_3 = 13.691 \\ \tfrac{1}{2}(\bar{a}_1 + \bar{a}_2) &= 12.696 \end{aligned}$$

Therefore,

For $U \leq 12.696$ classify the individual as coming from π_1 population.
For $U > 12.696$ classify the individual as coming from π_2 population.

Step 6. *Computation of the Efficiency of Classification.*

$$\begin{aligned} \bar{\sigma}^2 = {} & s^{11}(\bar{y}_1 - \bar{x}_1)(\bar{y}_1 - \bar{x}_1) + s^{12}(\bar{y}_1 - \bar{x}_1)(\bar{y}_2 - \bar{x}_2) + s^{13}(\bar{y}_1 - \bar{x}_1)(\bar{y}_3 - \bar{x}_3) \\ & + s^{21}(\bar{y}_2 - \bar{x}_2)(\bar{y}_1 - \bar{x}_1) + s^{22}(\bar{y}_2 - \bar{x}_2)(\bar{y}_2 - \bar{x}_2) + s^{23}(\bar{y}_2 - \bar{x}_2)(\bar{y}_3 - \bar{x}_3) \\ & + s^{31}(\bar{y}_3 - \bar{x}_3)(\bar{y}_1 - \bar{x}_1) + s^{32}(\bar{y}_3 - \bar{x}_3)(\bar{y}_2 - \bar{x}_2) + s^{33}(\bar{y}_3 - \bar{x}_3)(\bar{y}_3 - \bar{x}_3) \\ = {} & 1.5764. \end{aligned}$$

$$\frac{\bar{a}_2 - \bar{a}_1}{2\bar{\sigma}} = .792$$

$$P_1 = 1 - P_2 = \frac{1}{\sqrt{2\pi}} \int_{.792}^{\infty} e^{-t^2/2} = .2062$$

where P_1 is the probability of making an error of Type I, that is, of classifying an individual as one who will do satisfactory work when he actually does unsatisfactory work; and $1 - P_2$ is the probability of making an error of Type II,

that is, of classifying a student as one who will do unsatisfactory work when he actually does satisfactory work.

3. Conclusions. In using the above classification equation to classify the 305 trainees used in this study, 21 errors of Type I were made or 22.9 percent, while 50 errors of Type II were made or 23.9 percent. These percentages seem reasonably close to the expected 20.6 percent.

NOTE ON AN IDENTITY IN THE INCOMPLETE BETA FUNCTION

By T. A. Bancroft

Iowa State College

Since the incomplete beta function has proved of some importance in statistics, it would appear that any additional information concerning its properties might at some time prove useful. In a paper by the author, [1], two identities in the incomplete beta function were incidentally obtained. They are as follows:

$$(1) \qquad (p+q)I_x(p,q) = pI_x(p+1,q) + qI_x(p,q+1)$$

and

$$(2) \quad (p+q+1)^{[2]}I_x(p,q) = (p+1)^{[2]}I_x(p+2,q) + 2pqI_x(p+1,q+1) + (p+1)^{[2]}I_x(p,q+2),$$

where the incomplete beta function $I_x(p,q) = \dfrac{B_x(p,q)}{B(p,q)}$, etc., and $(p+1)^{[2]}$, etc. refer to the standard factorial notation.

Written in the above form these two identities suggest a possible general identity to which they belong as special cases. The third special case suggested is:

$$\begin{aligned}(3) \quad (p+q+2)^{[3]}I_x(p,q) &= (p+2)^{[3]}I_x(p+3,q) \\ &+ 3(p+1)^{[2]}qI_x(p+2,q+1) + 3p(q+1)^{[2]}I_x(p+1,q+2) \\ &+ (q+2)^{[3]}I_x(p,q+3).\end{aligned}$$

The general formula suggested is

$$(4) \quad (p+q+n-1)^{[n]}I_x(p,q) = \sum_{r=0}^{n}\binom{n}{r}(p+n-r-1)^{[n-r]}(q+r-1)^{[r]}I_x(p+n-r,q+r).$$

To prove the general formula we write (4) as

$$(5) \quad (p+q+n-1)^{[n]}I_x(p,q) = \sum_{r=0}^{n}\binom{n}{r}(p+n-r-1)^{[n-r]} \cdot (q+r-1)^{[r]}\frac{B_x(p+n-r,q+r)}{B(p+n-r,q+r)}.$$