APPROXIMATE SOLUTIONS FOR MEANS AND VARIANCES IN A CERTAIN CLASS OF BOX PROBLEMS

By Philip J. McCarthy

Social Science Research Council

1. Summary. Consider n boxes, each box having an associated probability, p_i, $(\sum_i p_i = 1)$, and an associated integer, k_i. If balls are thrown one by one into these boxes, the probability being p_i that any one ball falls into the ith box, then the number of balls which must be thrown in order to obtain, *for the first time*, at least k_{i_1} balls in the i_1th box, at least k_{i_2} balls in the i_2th box, $\cdots$, and at least k_{i_s} balls in the i_sth box, is a random variable, $N_s[k_1(p_1), k_2(p_2), \cdots, k_n(p_n)]$. Here $i_1, i_2, \cdots, i_s$ represent the numbers of that set of s boxes, $(1 \leq s \leq n)$, which *first* satisfies the stated condition.

The distribution of $N_s[k_1(p_1), k_2(p_2), \cdots, k_n(p_n)]$ can be written down for any set of values assigned to n, s, the p_i's and the k_i's. However, for n greater than 2 the distribution assumes such an extremely complicated multinomial form that except for certain special cases even the mean of the distribution cannot be numerically evaluated without a prohibitive amount of labor.

This paper presents the exact moments of $N_1[k_1(p_1), k_2(p_2)]$ and $N_2[k_1(p_1), k_2(p_2)]$ in forms that readily lend themselves to computation and shows how these moments can be used to obtain approximate values for the mean and variance for certain situations where n is greater than two. These approximation formulae are given for

1. The mean and variance, for any n and any set of k_i's and p_i's when $s = 1$ or n.

2. The mean, for any n and $2 \leq s \leq n - 1$, when $p_i = 1/n$, $k_i = k$, $(i = 1, 2, \cdots, n)$.

Some indications are given concerning the error of the approximations, and the circumstances which lead to a minimum (and maximum) error. Curves have been prepared to show the mean for the two box case, the primary function of these curves being to assist in the application of the approximation formulae. Some problems where the results of this paper might be applicable are suggested in the Introduction.

2. Introduction. A box problem is defined when one is given a fixed number of boxes, a collection of balls (either finite or infinite), a set of rules governing the throwing of the balls into the boxes and a statement of the conditions which will bring the throwing to an end. The terminating conditions usually state either that a fixed number of balls will be thrown or that balls will be thrown until a particular distribution of balls in the boxes has been obtained. In the first of these, interest is centered on the possible distributions which can be ob-

tained, while in the latter the number of balls necessary to obtain a specified distribution is of primary interest.

This paper will be concerned with certain problems falling in the latter category. In the simplest case one is given two boxes with associated probabilities p_1 and p_2 and associated integers k_1 and k_2. Balls are thrown one by one into the two boxes, the probability being p_1 that any one ball goes in the first box and p_2 that it goes in the second box. This process is stopped when either k_1 balls fall in box 1 or k_2 balls in box 2, *whichever occurs first.* One is interested in the distribution of the number of balls necessary to terminate the throwing. This problem was stated in essentially this form by Laplace [4], but he contented himself with merely writing down the probability generating function.

Here the special case of two boxes will be treated in detail and the results will then be generalized to the n-box case. In all of these instances it is possible to write down exact expressions for the mean and variance of the number of balls required to achieve the stated distribution. However, in almost every case the resulting expressions are too complicated to be of any use when a numerical answer is desired. The principal portion of this paper will be devoted to obtaining approximate formulae from which numerical answers can be obtained for these problems. Some evaluation of the degree of approximation will be given in section 5, while curves to facilitate the computation will be given in section 6.

The statement of these problems in terms of boxes and balls may lead one to the belief that they have no other interpretation. Actually this is not the case, and a few illustrations of this point will now be given. For example, consider the curtailed single sampling plan used in acceptance sampling. A buyer receives a lot of articles. This lot will contain a certain proportion of defective items. The buyer wishes to determine on the basis of sampling whether to accept or reject the lot. His knowledge of his own situation will allow him to specify the largest proportion of defectives which he is ordinarily willing to accept and the risk he is willing to take of accepting a lot with a proportion defective larger than this critical proportion. On the basis of these two values it is possible to set up a sampling plan in which the buyer will take a sample of size n out of the lot, inspect it, and reject it if there are k_1 or more defectives in the sample. Of course once he has obtained k_1 defectives there is no need to inspect the remainder of the sample. The lot will then be automatically rejected. Similarly, once he has obtained $n-k_1$ non-defectives, he can accept the lot without inspecting the remainder of the items. The average number of items which he must inspect in order to reach a decision is given by the solution to the two box problem stated above. Box 1 will receive the defective items, the associated integer being k_1 and the associated probability being p_1, the true proportion of defectives in the lot. Box 2 will receive the non-defective items, the associated integer being $n-k_1$ and the associated probability being p_2, the true proportion of non-defectives in the lot.

Laplace [4] considered problems of this type as applied to games of chance.

Thus suppose there are two players A and B who participate in successive trials of a given event, the probability being p_1 that A wins on any one trial and p_2 that B wins. Then one can associate the integer k_1 with A and k_2 with B by saying that A wins the match if he wins k_1 trials before B wins k_2 trials and conversely. The analysis is exactly the same as for the two box problem. It is apparent that this same situation can be extended to any number of players.

Another possible interpretation is as a particular kind of random walk problem. Let a particle start at the origin of a system of rectangular coordinates and suffer successive positive unit displacements, the probability being p_1 that it moves one unit in the x-direction and p_2 that it moves one unit in the y-direction. Furthermore assume that it is absorbed if it ever reaches the line $x = k_1$ or the line $y = k_2$. Then the analysis of the above two box problem gives the mean number of displacements before it is absorbed. In the same manner, such a random walk problem can be stated for n dimensions. For n equal to three, there will be three planes and the particle will be absorbed when it reaches any one of the three.

Certain problems in public opinion polling may fit into this category of box problems, particularly if the above problem is rephrased so that one requires the mean number of trials to obtain at least k_1 balls in the first box and at least k_2 balls in the second box, *for the first time*. For example, suppose that one desires to sample from a population composed of two types of individuals, A and B. Let the population proportions of A and B be known and be denoted by p_1 and p_2. Then if one wishes to obtain at least k_1 individuals of type A and at least k_2 individuals of type B, the average number of persons who must be chosen in order to fulfill this condition is given by the analysis of the corresponding box problem. This is rather artificial when there are only two categories and $p_1 + p_2 = 1$. However, these restrictions will be removed in the course of the paper, and the problem will be considered for any number of types of individuals.

As a final example, consider one of the many bombing problems which arose during the course of war research. Suppose that a factory which is to be demolished has n vital units, the destruction of any one of which will destroy the usefulness of the factory. Let the probability be p_1 of hitting the first unit with a single bomb, p_2 the probability of hitting the second with a single bomb, etc., and assume that k_1 bomb hits will finish off the first unit, k_2, the second, etc. Then the mean number of bombs required will be given by the analysis for the corresponding box problem.

Corresponding interpretations are possible for the other problems which are to be considered in this paper. Some of these will be indicated as the analysis proceeds and it is to be hoped that others will occur to the reader.

As previously noted, this paper will be concerned with the distribution of balls necessary to terminate the throwing, assuming the p's are known. Another possible interpretation is to assume the p's unknown and to estimate them with the results of the ball throwing. Certain aspects of this problem for two boxes

have been considered by J. B. S. Haldane [3] and Girshick, Mosteller and Savage [2].

3. Solution for the two box case.

3.1. *Distribution and moments of the number of trials necessary to obtain either k_1 balls in the first box or k_2 balls in the second box.* This problem may be stated as follows: Suppose one is given two boxes with associated probabilities p_1 and p_2, and associated integers k_1 and k_2. For the present it will be assumed that $p_1 + p_2 = 1$, although this restriction will be removed later. Now let balls be thrown one by one into these two boxes, the probability being p_1 that a particular ball will fall in the first box and p_2 that it will fall in the second box. This process is stopped on the first ball which leaves either k_1 balls in the first box or k_2 balls in the second box. The number of balls, x, which is required to accomplish this is a random variable and we desire the moments of x. The probability that k_1 balls are obtained in the first box on the xth throw, $k_1 \leq x \leq k_1 + k_2 - 1$, before k_2 balls are obtained in the second box, is immediately seen to be

$$\left[\binom{x-1}{k_1-1} p_1^{k_1-1} p_2^{x-k_1}\right] \cdot p_1 = \binom{x-1}{k_1-1} p_1^{k_1} p_2^{x-k_1}. \tag{3.1}$$

Similar reasoning gives the probability that k_2 balls are obtained in the second box for the first time on the xth throw, $k_2 \leq x \leq k_1 + k_2 - 1$, as

$$\binom{x-1}{k_2-1} p_1^{x-k_2} p_2^{k_2}. \tag{3.2}$$

From (3.1) and (3.2), the hth moment of x, $E(x^h)$, is

$$\sum_{x=k_1}^{k_1+k_2-1} x^h \binom{x-1}{k_1-1} p_1^{k_1} p_2^{x-k_1} + \sum_{x=k_2}^{k_1+k_2-1} x^h \binom{x-1}{k_2-1} p_1^{x-k_2} p_2^{k_2} \tag{3.3}$$

However, it is inconvenient to consider (3.3) directly. A much simpler procedure is to determine the increasing factorial moments of x and then transform these into the ordinary moments. Thus the hth increasing factorial moment of x, $F_{h,1}[k_1(p_1), k_2(p_2)]$, is defined as $E[x(x+1)\cdots(x+h-1)]$. Then $F_{h,1}[\ ]$ is equal to

$$\sum_{x=k_1}^{k_1+k_2-1} \frac{(x+h-1)!}{(x-1)!} \binom{x-1}{k_1-1} p_1^{k_1} p_2^{x-k_1} + \sum_{x=k_2}^{k_1+k_2-1} \frac{(x+h-1)!}{(x-1)!} \binom{x-1}{k_2-1} p_1^{x-k_2} p_2^{k_2} \tag{3.4}$$

(3.4) can be transformed by means of the relationship

$$\sum_{j=0}^{a} \binom{k+j}{j} p_1^j = (1-p_1)^{-(k+1)} I_{1-p_1}(k+1, a+1), \tag{3.5}$$

where $I_x(p, q)$ is the Incomplete Beta-Function as tabulated by Karl Pearson [6], and the result is obtained that

$$F_{h,1}[k_1(p_1), k_2(p_2)] = \frac{k_1(k_1+1)\cdots(k_1+h-1)}{p_1^h} I_{p_1}(k_1+h, k_2) + \frac{k_2(k_2+1)\cdots(k_2+h-1)}{p_2^h} I_{p_2}(k_2+h, k_1). \tag{3.6}$$

The ordinary hth moment of x may be written in terms of $F_{1,1}[\ ]$, $F_{2,1}[\ ]$, $\cdots$, $F_{h,1}[\ ]$ as

$$E(x^h) = \sum_{i=1}^{h} F_{i,1}[\ ] \frac{\Delta^i 0^h}{i!} (-1)^{h+i}, \tag{3.7}$$

where $\Delta^i 0^h$ represents a difference of zero. Tabular values of $\Delta^i 0^h / i!$ are given by Fisher and Yates [1].

In particular, the mean and variance of x, which will receive the special designations $E_1[k_1(p_1), k_2(p_2)]$ and $\sigma_1^2[k_1(p_1), k_2(p_2)]$ respectively, are

$$\frac{k_1}{p_1} I_{p_1}(k_1+1, k_2) + \frac{k_2}{p_2} I_{p_2}(k_2+1, k_1) \tag{3.8}$$

and

$$\frac{k_1(k_1+1)}{p_1^2} I_{p_1}(k_1+2, k_2) + \frac{k_2(k_2+1)}{p_2^2} I_{p_2}(k_2+2, k_1) - E_1[k_1(p_1), k_2(p_2)] - \{E_1[k_1(p_1), k_2(p_2)]\}^2. \tag{3.9}$$

In the event the p's are equal and sum to one, $E_1[k_1(p_1), k_2(p_2)]$ will be abbreviated to $E_1[k_1, k_2]$, and finally, if both the p's and k's are equal, it will be written as $E_1[k^2]$. In this two box situation, the only other possibility is $E_2[k_1(p_1), k_2(p_2)]$, which will denote the expected number of balls required to obtain at least k_1 in the first box and at least k_2 in the second box, for the first time. This problem will be considered in section 3.2.

In order to facilitate the computation of mean values, both for the two box problem itself and for its application to problems involving a larger number of boxes, (3.8) has been graphed for various values of k_1, k_2, p_1 and p_2. A discussion of this procedure and the results obtained will be found in section 6.

There is one further result which will later prove useful. Consider the situation when there is only one box with p_1 and k_1, $p_1 < 1$. This is the same as having two boxes where the k_2 corresponding to the second box is infinite. In other words, one can terminate the throwing of balls only because of what happens to the first box, never because of anything that happens to the second box. In this case one obtains

$$E_1[k_1(p_1), \infty\ (p_2)] = \sum_{x=k_1}^{\infty} x \binom{x-1}{k_1-1} p_1^{k_1} p_2^{x-k_1} = \frac{k_1}{p_1}. \tag{3.10}$$

Similarly,

$$\sigma_1^2[k_1(p_1), \infty\ (p_2)] = \frac{k_1 p_2}{p_1^2}. \tag{3.11}$$

3.2. *Distribution and moments of the number of throws necessary to obtain at least k_1 balls in the first box and at least k_2 balls in the second box.* This problem may be stated as follows: Suppose there are two boxes with associated probabilities p_1 and p_2, and associated integers k_1 and k_2. As in 3.1, $p_1 + p_2 = 1$. Let balls be thrown into the boxes one by one, the probability being p_1 that a particular ball will fall in the first box and p_2 that it will fall in the second box. This process is stopped on the first ball which leaves at least k_1 in the first box and exactly k_2 in the second or at least k_2 in the second and exactly k_1 in the first. Again x is the number of balls required to accomplish this. As explained in 3.1, the mean value in this case will be written as $E_2[k_1(p_1), k_2(p_2)]$. The analysis follows through as in 3.1 and the mean number of trials is equal to

$$\sum_{x=k_1+k_2}^{\infty} x \binom{x-1}{k_1-1} p_1^{k_1} p_2^{x-k_1} + \sum_{x=k_1+k_2}^{\infty} x \binom{x-1}{k_2-1} p_1^{x-k_2} p_2^{k_2}. \tag{3.12}$$

Making use of (3.5), this can be written as

$$\frac{k_1}{p_1}[1 - I_{p_1}(k_1 + 1, k_2)] + \frac{k_2}{p_2}[1 - I_{p_2}(k_2 + 1, k_1)]. \tag{3.13}$$

Referring to (3.8) it is evident that

$$E_1[k_1(p_1), k_2(p_2)] + E_2[k_1(p_1), k_2(p_2)] = \frac{k_1}{p_1} + \frac{k_2}{p_2}. \tag{3.14}$$

The hth increasing factorial moment in this problem, denoted by $F_{h,2}[k_1(p_1), k_2(p_2)]$, is

$$\begin{aligned} &\frac{k_1(k_1+1)\cdots(k_1+h-1)}{p_1^h}[1 - I_{p_1}(k_1 + h, k_2)] \\ &\quad + \frac{k_2(k_2+1)\cdots(k_2+h-1)}{p_2^h}[1 - I_{p_2}(k_2 + h, k_1)]. \end{aligned} \tag{3.15}$$

Comparison of (3.15) with (3.6) gives the relationship

$$\begin{aligned} F_{h,1}[\ ] + F_{h,2}[\ ] = &\frac{k_1(k_1+1)\cdots(k_1+h-1)}{p_1^h} \\ &+ \frac{k_2(k_2+1)\cdots(k_2+h-1)}{p_2^h}. \end{aligned} \tag{3.16}$$

The ordinary moments of x can be computed from (3.15) by the use of (3.7). That is, formula (3.7) holds in this case if $F_{h,1}[\ ]$ is replaced by $F_{h,2}[\ ]$.

It can be easily shown by the use of the recursion relationship for the Incomplete Beta-Function,

$$I_x(p, q) = xI_x(p - 1, q) + (1 - x)I_x(p, q - 1),$$

that $F_{h,1}[\ ]$ and $F_{h,2}[\ ]$ satisfy the partial difference equation

$$\begin{aligned} F_{h,i}[k_1(p_1), k_2(p_2)] &= hF_{h-1,i}[k_1(p_1), k_2(p_2)] \\ &+ p_1F_{h,i}[(k_1 - 1)(p_1), k_2(p_2)] \\ &+ p_2F_{h,i}[k_1(p_1), (k_2 - 1)(p_2)], \end{aligned} \tag{3.17}$$

where $i = 1$ or 2. This equation can be used as an alternative way of obtaining many results, examples of which are (3.10) and (3.11). Certain of these applications have been discussed by McCarthy [5].

4. Solution for the n box case.

4.1. *Preliminary discussion.* The problems of this section, although direct generalizations of the two box cases, can perhaps be most easily stated and illustrated as applied to the behavior of a random particle. Suppose that we have a random particle which starts at the origin of n-dimensional rectangular coordinates and moves in unit steps along the positive coordinate axes. At any given point the probability will be taken as p_i that it moves in the x_i-direction. $\sum_{i=1}^{n} p_i$ is assumed to be one unless otherwise specified. Now consider the n hyperplanes, $x_i = k_i$, and assume that the particle will be absorbed if it passes through a *specified number*, say s, of these hyperplanes. Notice that we are interested only in the number of planes which it passes through, and not in the particular ones. For each s, $(s = 1, 2, \cdots, n)$, the number of moves which the particle makes before it is absorbed is a random variable, and in this section we will be concerned with the distribution of this random variable. The corresponding interpretations for boxes and balls is immediately obvious.

These problems are seen to be generalizations of the two box cases considered in section 3. Although it is always relatively easy to write down formal expressions for the quantities to be considered, the step from two boxes to three or more boxes produces expressions which are extremely difficult, or even impossible, to evaluate. In this section we shall develop approximate solutions which make use only of simple computations based on the solution for the two box case.

As an introduction to the contents of this section, we shall discuss briefly a box problem which is a special case of the general problem. Assume that there are n boxes with a probability of $1/n$ that any one ball will be thrown into a particular one of the n boxes. Then one can ask for the mean and variance of

the number of trials required to obtain s occupied boxes (i.e. $k_1 = k_2 = \cdots = k_n = 1$). Making use of (3.10) and (3.11), we obtain

$$
\begin{aligned}
E_1[1^n] &= 1 \\
E_2[1^n] &= 1 + E_1\left[1\left(\frac{n-1}{n}\right), \infty\left(\frac{1}{n}\right)\right] = 1 + \frac{n}{n-1} \\
E_3[1^n] &= 1 + \frac{n}{n-1} + E_1\left[1\left(\frac{n-2}{n}\right), \infty\left(\frac{2}{n}\right)\right] \\
&= 1 + \frac{n}{n-1} + \frac{n}{n-2} \qquad (4.1) \\
&\vdots \\
E_s[1^n] &= 1 + \frac{n}{n-1} + \cdots + \frac{n}{n-s+1} = n\sum_{i=0}^{s-1}\frac{1}{n-i},
\end{aligned}
$$

and

$$
\begin{aligned}
\sigma_1^2[1^n] &= 0 \\
\sigma_2^2[1^n] &= 0 + \sigma_1^2\left[1\left(\frac{n-1}{n}\right), \infty\left(\frac{1}{n}\right)\right] = 0 + \frac{n}{(n-1)^2} \\
\sigma_3^2[1^n] &= 0 + \frac{n}{(n-1)^2} + \sigma_1^2\left[1\left(\frac{n-2}{n}\right), \infty\left(\frac{2}{n}\right)\right] \\
&= 0 + \frac{n}{(n-1)^2} + \frac{2n}{(n-2)^2} \qquad (4.2) \\
&\vdots \\
\sigma_s^2[1^n] &= 0 + \frac{n}{(n-1)^2} + \frac{2n}{(n-2)^2} \\
&\quad + \cdots + \frac{(s-1)n}{(n-s+1)^2} = n\sum_{i=1}^{s-1}\frac{i}{(n-i)^2}.
\end{aligned}
$$

The solution for this problem for $s = n$ is given in Uspensky [9], but a straightforward solution requires a great deal of formal manipulation. The step-by-step procedure used here is somewhat indicative of the methods to be used in the succeeding portions of this paper.

4.2. *Mean and variance of the number of trials required to obtain either k_1 balls in the first box, or k_2 in the second, $\cdots$, or k_{n-1} in the $(n-1)$st, the probability associated with the nth box being non-zero.* The mean number of trials in this

particular problem is represented by $E_1[k_1(p_1), \cdots, k_{n-1}(p_{n-1}), \infty\ (p_n)]$. The formal expression for this quantity is

$$\sum_{i=1}^{n-1} \sum_{j=k_i}^{\infty} j \frac{(j-1)!}{(k_i-1)!\,(j-k_i)!} p_i^{k_i} \times \sum \frac{(j-k_i)!}{r_1!\cdots r_{i-1}!\, r_{i+1}!\cdots r_n!} p_1^{r_1}\cdots p_{i-1}^{r_{i-1}} p_{i+1}^{r_{i+1}}\cdots p_n^{r_n}, \tag{4.3}$$

where the third sum is taken over all values of the r's such that

$$r_1 + \cdots + r_{i-1} + r_{i+1} + \cdots + r_n = j - k_i$$

and

$$r_1 < k_1, \cdots, r_{i-1} < k_{i-1}, r_{i+1} < k_{i+1}, \cdots, r_{n-1} < k_{n-1}.$$

This expression can be reduced by one dimension by the application of some of the results for two boxes. Consider for the moment only those balls going into the first $(n-1)$ boxes. Then the number of balls (conditional) which is necessary to obtain either k_1 in the first box, or k_2 in the second, $\cdots$, or k_{n-1} in the $(n-1)$st box is a random variable X which takes on values

$$k_1, k_1 + 1, \cdots, k_1 + k_2 + \cdots + k_{n-1} - (n-2)$$

with corresponding probabilities π_j, where with no loss of generality it is assumed that $k_1 \leq k_2 \leq \cdots \leq k_{n-1}$. π_j is given by a sum of $(n-1)$ multinomial expressions, the probability associated with the ith box now being $p_i \Big/ \left(\sum_{i=1}^{n-1} p_i\right)$, which will be designated by p'_i.

Under these circumstances it is apparent that

$$E_1[k_1(p_1), \cdots, k_{n-1}(p_{n-1}), \infty\ (p_n)] = \sum_j \pi_j E_1[x_j(p_1 + \cdots + p_{n-1}), \infty\ (p_n)]. \tag{4.4}$$

However, (3.10) can be applied to each term in (4.4), leading to

$$\frac{1}{(p_1 + p_2 + \cdots + p_{n-1})} \sum_j \pi_j x_j. \tag{4.5}$$

Now from the definition of π_j and x_j we have

$$E_1[k_1(p_1), \cdots, k_{n-1}(p_{n-1}), \infty\ (p_n)] = \frac{1}{(p_1 + p_2 + \cdots + p_{n-1})} E_1[k_1(p'_1), k_2(p'_2), \cdots, k_{n-1}(p'_{n-1})]. \tag{4.6}$$

Similarly, the application of (3.11) gives the result that

$$\sigma_1^2[k_1(p_1), \cdots, k_{n-1}(p_{n-1}), \infty\ (p_n)] = \frac{p_n}{(p_1 + p_2 + \cdots + p_{n-1})^2} E_1[k_1(p'_1), \cdots, k_{n-1}(p'_{n-1})]. \tag{4.7}$$

These results are of immediate importance for two reasons:

1. They indicate that by combining boxes and introducing a new random variable, certain problems can be simplified. This statement will be expanded and the principle applied repeatedly in the later portions of this paper.

2. With respect to the section on two boxes, they mean that the restriction $p_1 + p_2 = 1$ is not necessary for the solution of the problems. One can always assume that $p_3 (= 1 - p_1 - p_2)$ refers to a box which receives balls but which otherwise has no effect on the outcome of an experiment. In this paper it has been convenient to refer to such a box as having an infinite capacity.

4.3. *The mean value and variance of the number of trials required in a two box problem when one or both of the constants k_1 and k_2 are replaced by random variables.* The discussion in 4.2 has indicated that the idea of associating a random variable with a box instead of a single integer may sometimes lead to simplification. Here this procedure will be treated in more detail. Consider $E_1[k_1(p_1), k_2(p_2)]$ and assume that k_1 is replaced by a random variable X which can take on values $x_1, x_2, \cdots, x_t$ with corresponding probabilities $\pi_1, \cdots, \pi_i, \cdots, \pi_t$. Under these circumstances $E_1[\ ]$ itself becomes the random variable $E_1[X(p_1), k_2(p_2)]$, taking on values $E_1[x_i(p_1), k_2(p_2)]$, $(i = 1, 2, \cdots, t)$, with corresponding probabilities π_i. The mean value of this new random variable can be formally written down as

$$E(E_1[X(p_1), k_2(p_2)]) = \sum_{i=1}^{t} \pi_i E_1[x_i(p_1), k_2(p_2)]. \tag{4.8}$$

This expression can always be calculated from the probabilities π_i and (3.8) or from the curves given in section 6. However, in the applications which will arise later in this paper, this computation would be very time consuming. Instead, an approximation to (4.8) will now be derived which will prove to yield very good results, and which can be obtained by a simple reading on the above mentioned curves.

If X is regarded as a continuous variable, then $E_1[X(p_1), k_2(p_2)]$ is a continuous function of X, and, in fact, can be represented by a single curve similar to those appearing in section 6. Moreover, as is apparent from (3.8), repeated differentiation of $E_1[X(p_1), k_2(p_2)]$ yields continuous derivatives. Consequently, $E_1[X(p_1), k_2(p_2)]$ can be expanded in Taylor series about a, where $a = \sum_{i=1}^{t} \pi_i x_i$. This procedure gives

$$E(E_1[X(p_1), k_2(p_2)]) = \sum_{i=1}^{t} \pi_i \sum_{j=0}^{\infty} \frac{(x_i - a)^j}{j!} E_1^j[a(p_1), k_2(p_2)], \tag{4.9}$$

where $E_1^j[a(p_1), k_2(p_2)]$ represents the jth derivative of $E_1[X(p_1), k_2(p_2)]$ with respect to X evaluated at a. Interchanging the order of summation one obtains

$$\sum_{j=0}^{\infty} \frac{E_1^j[a(p_1), k_2(p_2)]}{j!} \sum_{i=1}^{t} \pi_i (x_i - a)^j. \tag{4.10}$$

The final result then becomes

$$E(E_1[X(p_1), k_2(p_2)]) = \sum_{j=0}^{\infty} \frac{E_1^j[a(p_1), k_2(p_2)]}{j!} \mu_j, \tag{4.11}$$

where μ_j is the jth moment of X about its mean, a. Thus to a first approximation

$$E(E_1[X(p_1), k_2(p_2)]) \simeq E_1[a(p_1), k_2(p_2)]. \tag{4.12}$$

It is of interest to note that if $E_1[X(p_1), k_2(p_2)]$ is linear in X then (4.12) is an exact expression since all derivatives except the first are zero. Furthermore, if $E_1[X(p_1), k_2(p_2)]$ is of the second degree in X, then only the second non-zero term on the right hand side of (4.11) needs to be added to (4.12) in order to make it exact. The former of these is the relation which gave an exact solution in 4.2.

It is important to realize that this analysis for $E(E_1[X(p_1), k_2(p_2)])$ can be immediately applied to $E(E_2[X(p_1), k_2(p_2)])$. For, by the use of (3.14) and (4.8), one obtains

$$E(E_2[X(p_1), k_2(p_2)]) = \frac{a}{p_1} + \frac{k_2}{p_2} - E(E_1[X(p_1), k_2(p_2)]). \tag{4.13}$$

The same analysis can be applied to $F_{h,1}[\ ]$ and the general result obtained that

$$E(F_{h,1}[X(p_1), k_2(p_2)]) \simeq F_{h,1}[a(p_1), k_2(p_2)]. \tag{4.14}$$

This immediately allows one to approximate the variance in the obvious manner.

It is of interest to consider briefly the situation when both k_1 and k_2 are replaced by random variables. Let k_1 be replaced by X_1 taking on values $x_{11}, x_{12}, \cdots, x_{1t}$ with probabilities $\pi_{11}, \pi_{12}, \cdots, \pi_{1t}$ and k_2 be replaced by X_2 taking on values $x_{21}, x_{22}, \cdots, x_{2s}$ with probabilities $\pi_{21}, \pi_{22}, \cdots, \pi_{2s}$. Then

$$E(E_1[X_1(p_1), X_2(p_2)]) = \sum_{i,j} \pi_{1i}\pi_{2j} E_1[x_{1i}(p_1), x_{2i}(p_2)], \tag{4.15}$$

where $i = 1, 2, \cdots, t$ and $j = 1, 2, \cdots, s$. Again applying Taylor series and expanding about $a = \sum_i \pi_{1i}x_{1i}$ and $b = \sum_j \pi_{2j}x_{2j}$, the result is obtained that

$$E(E_1[X_1(p_1), X_2(p_2)]) = \sum_{u,v=0}^{\infty} \frac{E_1^{uv}[a(p_1), b(p_2)]}{u!\, v!} \mu_{1u}\mu_{2v}, \tag{4.16}$$

where $E_1^{uv}[a(p_1), b(p_2)]$ is the uth partial derivative with respect to X_1 and the vth partial derivative with respect to X_2 of $E_1[X_1(p_1), X_2(p_2)]$ evaluated at $X_1 = a$, $X_2 = b$. This gives the approximate formula

$$E(E_1[X_1(p_1), X_2(p_2)]) \simeq E_1[a(p_1), b(p_2)]. \tag{4.17}$$

4.4. *Mean and variance of the number of trials required to obtain either* (at least) k_1 *balls in the first box, or* (at least) k_2 *balls in the second box,* $\cdots$ *, or* (and at least) k_n *balls in the* n*th box.* In accordance with previous notation the mean number of trials required is given by $E_1[k_1(p_1), k_2(p_2), \cdots, k_n(p_n)]$. The exact value of this quantity can be written down and it would be a complicated multinomial expression. The evaluation of such an expression would be extremely difficult, if not impossible, especially for large values of $k_1, k_2, \cdots, k_n$. In order to obtain an approximation to $E_1[\ ]$, repeated applications of (4.12) can be made and the resulting expression can be evaluated by means of the curves in section 6.

For convenience, consider $E_1[k_1(p_1), k_2(p_2), k_3(p_3), k_4(p_4)]$. The general result will then be apparent. Assume that the first three boxes form a single unit with probability $(p_1 + p_2 + p_3)$. Then the number of balls required to obtain either k_1 in the first, k_2 in the second or k_3 in the third, if all balls are going in these three boxes, is a random variable X. Consequently,

$$E_1[k_1(p_1), \cdots, k_4(p_4)] = E(E_1[X(p_1 + p_2 + p_3), k_4(p_4)]). \tag{4.18}$$

Applying (4.12),

$$E_1[k_1(p_1), \cdots, k_4(p_4)] \simeq E_1\left[E_1\left[k_1\left(\frac{p_1}{p_1 + p_2 + p_3}\right), k_2\left(\frac{p_2}{p_1 + p_2 + p_3}\right), k_3\left(\frac{p_3}{p_1 + p_2 + p_3}\right)\right](p_1 + p_2 + p_3), k_4(p_4)\right]. \tag{4.19}$$

Applying (4.12) once again the final approximation is

$$E_1[k_1(p_1), \cdots, k_4(p_4)] \simeq E_1\left[E_1\left[E_1\left[k_1\left(\frac{p_1}{p_1 + p_2}\right), k_2\left(\frac{p_2}{p_1 + p_2}\right)\right]\left(\frac{p_1 + p_2}{p_1 + p_2 + p_3}\right), k_3\left(\frac{p_3}{p_1 + p_2 + p_3}\right)\right](p_1 + p_2 + p_3), k_4(p_4)\right]. \tag{4.20}$$

Expression (4.20) can be translated into a course of procedure. One considers the first two boxes and computes

$$a_1 = E_1\left[k_1\left(\frac{p_1}{p_1 + p_2}\right), k_2\left(\frac{p_2}{p_1 + p_2}\right)\right].$$

It is then assumed that a_1 is a new number associated with a box with probability $(p_1 + p_2)$ and

$$a_2 = E_1\left[a_1\left(\frac{p_1 + p_2}{p_1 + p_2 + p_3}\right), k_3\left(\frac{p_3}{p_1 + p_2 + p_3}\right)\right].$$

Repeating this procedure again, one computes $a_3 = E_1[a_2(p_1 + p_2 + p_3), k_4(p_4)]$, and by (4.20) this is approximately equal to $E_1[k_1(p_1), \cdots, k_4(p_4)]$. This method of computation is seen to be completely general and one can apply it to any number of boxes. Each step consists of computing $E_1[\ ]$ for two boxes and consequently can be carried out with the curves of section 6. It is evident that the order in which the boxes are taken may have an important effect on the size of the error involved in using this step-by-step procedure. This problem will be considered in section 5.

It is of interest to note that one can also obtain another approximation for $E_1[k_1(p_1), k_2(p_2), k_3(p_3), k_4(p_4)]$. Suppose that the first two boxes are considered as one unit and the second two boxes as another unit. Then the number of balls which must fall in the first two boxes in order to obtain either k_1 in the first box or k_2 in the second is a random variable X_1. Similarly a random variable X_2 can be associated with the last two boxes. Accordingly

$$E_1[k_1(p_1), \cdots, k_4(p_4)] = E(E_1[X_1(p_1 + p_2), X_2(p_3 + p_4)]). \tag{4.21}$$

By use of (4.17), (4.21) can be written as

$$E_1[k_1(p_1), \cdots, k_4(p_4)] \simeq E_1\left[E_1\left[k_1\left(\frac{p_1}{p_1+p_2}\right), k_2\left(\frac{p_2}{p_1+p_2}\right)\right](p_1+p_2), E_1\left[k_3\left(\frac{p_3}{p_3+p_4}\right), k_4\left(\frac{p_4}{p_3+p_4}\right)\right](p_3+p_4)\right]. \tag{4.22}$$

This same analysis applies directly to the factorial moments. In particular

$$F_{2,1}[k_1(p_1), \cdots, k_4(p_4)] \simeq F_{2,1}\left[E_1\left[E_1\left[k_1\left(\frac{p_1}{p_1+p_2}\right), k_2\left(\frac{p_2}{p_1+p_2}\right)\right]\left(\frac{p_1+p_2}{p_1+p_2+p_3}\right), k_3\left(\frac{p_3}{p_1+p_2+p_3}\right)\right](p_1+p_2+p_3), k_4(p_4)\right]. \tag{4.23}$$

From (4.20) and (4.23) an approximate value for $\sigma_1^2[k_1(p_1), k_2(p_2), k_3(p_3), k_4(p_4)]$ can be obtained. This procedure is also perfectly general and so an estimate of $\sigma_1^2[\ ]$ can be obtained for any number of boxes.

This same method can be immediately applied to the approximation of $E_n[k_1(p_1), \cdots, k_n(p_n)]$. One simply considers the boxes two at a time, computing $E_2[\ ]$ at each stage instead of $E_1[\ ]$.

4.5. *Solution for $E_s[k^n]$ and $E_s[k_1^{n-1}, k_2]$.* When s is different from 1 or n, the complexities of the problem force one into the consideration of only the quantities given in the title of this subsection. The corresponding problem for three boxes, namely $E_2[k_1(p_1), k_2(p_2), k_3(p_3)]$, has been treated for general k_i and p_i by McCarthy [5]. However, the resulting expression is so complicated that it will not be given here.

The process to be used consists of reducing the subscript s by a series of steps

until the subscript 2 is reached. This expression can then be evaluated by the use of the curves or by simple computation. For the sake of convenience, the case $E_3[k^4]$ will be considered in detail. It will then be possible to write down the expression for general s and n.

As a starting point, look upon the first three boxes as a single unit. Then there is a definite probability π_i that one of these boxes will have k balls in it for the first time on the x_ith throw into these three boxes and that the other two boxes of the unit will each have less than k balls. Then if one of the other of the three boxes has u balls ($u < k$) the third box will have $(x_i - k - u)$ balls, $(x_i - k - u < k)$. Meanwhile the fourth box will also have been receiving balls, and the number in it at this time will be denoted by j, $(j = 0, 1, 2, \cdots, \infty)$. For each x_i there is a probability associated with u, namely $P(u \mid x_i)$, and another probability associated with j, $P(j \mid x_i)$. For the moment, consider that box 1 has received k balls, box 2 the $(x_i - k - u)$ balls, box 3 the u balls and box 4 the j balls. This numbering is of course immaterial since the situation is symmetric with respect to the first three boxes.

Now if $j \geq k$, either $(2k + u - x_i)$ balls will be required in the second box or $(k - u)$ balls in the third box in order to obtain three properly occupied boxes. On the other hand, if $j < k$, the specified number will be required in any two of boxes two, three and four. Consequently, with this conditional description of the situation, the required number of balls necessary to obtain three out of the four boxes occupied in the proper manner is

$$x_i + j + E_2[(2k + u - x_i), (k - u), (k - j)], \tag{4.24}$$

where $(k - j)$ will be taken as zero if j is greater than or equal to k. From this description, it is evident that the desired mean value may be obtained by summing (4.24) over all possible values of x_i, j and u. Therefore

$$E_3[k^4] = \sum_i \pi_i \Big\{ x_i + \sum_{j=0}^{\infty} P(j \mid x_i) \cdot \Big(j + \sum_u P(u \mid x_i) E_2[(2k + u - x_i), (k - u), k - j)] \Big) \Big\}. \tag{4.25}$$

It is to be noticed that the probabilities inside the $E_2[\ ]$ in (4.24) and (4.25) do not add to one but only to 3/4. This can be easily remedied by the application of a formula similar to (4.6) and the result is obtained that

$$E_3[k^4] = \sum_i \pi_i \Big\{ x_i + \sum_{j=0}^{\infty} P(j \mid x_i) \cdot \Big(j + 4/3 \sum_u P(u \mid x_i) E_2[(2k + u - x_i), (k - u), (k - j)] \Big) \Big\}, \tag{4.26}$$

where each probability inside $E_2[\ ]$ is now 1/3.

By simple considerations

$$P(u \mid x_i) = \frac{\dfrac{(x_i - k)!}{u!(x_i - k - u)!}\left(\tfrac{1}{2}\right)^u\left(\tfrac{1}{2}\right)^{x_i-k-u}}{\sum_u \dfrac{(x_i - k)!}{u!(x_i - k - u)!}\left(\tfrac{1}{2}\right)^u\left(\tfrac{1}{2}\right)^{x_i-k-u}}, \tag{4.27}$$

where u and $(x_i - k - u)$ are both less than k, and

$$P(j \mid x_i) = \frac{(x_i + j - 1)!}{(x_i - 1)!j!}\left(\tfrac{3}{4}\right)^{x_i}\left(\tfrac{1}{4}\right)^j. \tag{4.28}$$

From (4.27) and (4.28)

$$\sum_j jP(j \mid x_i) = x_i/3, \tag{4.29}$$

and

$$\sum_u uP(u \mid x_i) = \frac{x_i - k}{2}. \tag{4.30}$$

(4.25) can be written as

$$\begin{aligned} E_3[k^4] &= \sum_i \pi_i x_i + \sum_i \pi_i \sum_j jP(j \mid x_i) \\ &\quad + \frac{4}{3}\sum_i \pi_i \sum_j P(j \mid x_i) \sum_u P(u \mid x_i) E_2[(2k + u - x_i), (k - u), (k - j)]. \end{aligned} \tag{4.31}$$

Finally, making use of (4.29), (4.30), the definition of x_i and π_i and the procedure of replacing random variables inside an $E_2[\ ]$ by their mean values,

$$E_3[k^4] \simeq \frac{4}{3}\left\{E_1[k^3] + E_2\left[\left(\tfrac{3}{2}k - \frac{E_1[k^3]}{2}\right), \left(\tfrac{3}{2}k - \frac{E_1[k^3]}{2}\right), \left(k - \frac{E_1[k^3]}{3}\right)\right]\right\}, \tag{4.32}$$

and this in turn can be written as

$$E_3[k^4] \simeq \frac{4}{3}\left\{E_1[k^3] + E_2\left[\left(\tfrac{3}{2}k - \frac{E_1[k^3]}{2}\right)^2, \left(k - \frac{E_1[k^3]}{3}\right)\right]\right\}. \tag{4.33}$$

This method of analysis which has just been applied to $E_3[k^4]$ can be used equally well for $E_s[k^n]$. Here one simply considers the first $(n - 1)$ boxes and proceeds as above. The final result is immediately apparent, namely that

$$\begin{aligned} E_s[k^n] \simeq \frac{n}{n-1}\Big\{&E_1[k^{n-1}] + \\ &E_{s-1}\left[\left(\frac{n-1}{n-2}k - \frac{E_1[k^{n-1}]}{n-2}\right)^{n-2}, \left(k - \frac{E_1[k^{n-1}]}{n-1}\right)\right]\Big\}. \end{aligned} \tag{4.34}$$

It will be noticed that in reducing (4.34) further it will be necessary to consider expressions of the form $E_s[k_1^{n-1}, k_2]$. However, it will be seen from the foregoing

analysis that no use was made of the fact that the integers attached to the first $(n-1)$ boxes were the same. Accordingly,

$$E_s[k_1^{n-1}, k_2] \simeq \frac{n}{n-1}\left\{E_1[k_1^{n-1}] + E_{s-1}\left[\left(\frac{n-1}{n-2}k_1 - \frac{E_1[k_1^{n-1}]}{n-2}\right)^{n-2}, \left(k_2 - \frac{E_1[k_1^{n-1}]}{n-1}\right)\right]\right\}. \tag{4.35}$$

Now, by the use of (4.34) and (4.35), it is possible to reduce s as much as may be desired.

5. Some considerations concerning the error of the approximations.

5.1. *Preliminary remarks.* This discussion of the errors of the approximations given in the preceding sections has been left until now so that a broad perspective might be gained, and the errors seen in relationship to one another. Such an arrangement is advantageous in this instance since both the analytical and computational results bearing on the subject are scanty, and consequently, any intelligent leads which their inter-relationships can give are most helpful.

The difficulty involved in obtaining exact values for the various quantities considered in this paper has been pointed out quite frequently, and the approximations have been devised to overcome this very difficulty. The same complexity which prevents the computation of many exact values also prevents any effective analytic approach to the problem of evaluating the errors. For these reasons the author has been unable to carry through any general analytic treatment of the errors of the approximations. However, because the intelligent use of approximations requires some knowledge of their accuracy, certain isolated cases have been investigated by a combination of computational, graphical and analytic methods. These investigations are detailed in the remainder of this section, and conjectures concerning the general behavior of the errors are made whenever possible. As has been stated earlier, no consideration will be given to the approximation formulae for the variance.

5.2. *Errors of the approximations for* $E_1[k_1(p_1), \cdots, k_n(p_n)]$ *and*

$$E_n[k_1(p_1), \cdots, k_n(p_n)].$$

Taking n equal to 3, we have from (4.11) that

$$|E_1[k_1(p_1), k_2(p_2), k_3(p_3)] - E_1[a(p_1+p_2), k_3(p_3)]| \leq \tfrac{1}{2}\sigma_1^2\left[k_1\left(\frac{p_1}{p_1+p_2}\right), k_2\left(\frac{p_2}{p_1+p_2}\right)\right] \text{Max}\,|E_1^2[X(p_1+p_2), k_3(p_3)]|, \tag{5.1}$$

where $\text{Max}\,|E_1^2[X(p_1+p_2), k_3(p_3)]|$ is the maximum absolute value of the second derivative of $E_1[X(p_1+p_2), k_3(p_3)]$ with respect to X, and a is equal to $E_1[k_1(p_1/(p_1+p_2)), k_2(p_2/(p_1+p_2))]$. Now an examination of the curves

given in section 6 indicates that, for fixed p_3 and k_3, the maximum curvature of $E_1[X(p_1 + p_2), k_3(p_3)]$, considered as a function of X, is a monotone decreasing function of k_3. Since this curvature is negative, this geometric observation is equivalent to

$$\text{Max} \mid E_1^2[X(p_1 + p_2), (k_3 + 1)(p_3)] \mid \leq \text{Max} \mid E_1^2[X(p_1 + p_2), k_3(p_3)] \mid, \tag{5.2}$$

although it is not necessarily true that

$$\mid E_1^2[x_1(p_1 + p_2), (k_3 + 1)(p_3)] \mid \leq \mid E_1^2[x_1(p_1 + p_2), k_3(p_3)] \mid .$$

Moreover,

$$E_1[k_1(p_1), k_2(p_2), k_3(p_3)] \leq E_1[k_1(p_1), k_2(p_2), (k_3 + 1)(p_3)]. \tag{5.3}$$

From (5.1), (5.2) and (5.3) one readily obtains that the absolute value of the percentage error of the approximation to $E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$ is bounded by a function, say $U_1[k_1(p_1), k_2(p_2), k_3(p_3)]$, which is a monotone decreasing function of k_3 as k_3 increases. It should be noticed that the results of 4.2 have already shown not only that this upper bound for the percentage error approaches zero as k_3 becomes infinite, but also that the absolute difference between the true and approximate values approach zero as k_3 becomes infinite.

Computation of $U_1[k_1(p_1), k_2(p_2), k_3(p_3)]$ is very time consuming because of the difficulty in obtaining $\text{Max} \mid E_1^2[X(p_1 + p_2), k_3(p_3)] \mid$, and because the direct computation of $E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$ is laborious when any of k_1, k_2 and k_3 are much larger than 2 or 3. In order to surmount these difficulties and still give some indication of the behavior of $U_1[k_1(p_1), k_2(p_2), k_3(p_3)]$, the following expedients were adopted:

1. The values of k_1, k_2 and k_3 were each fixed at 5,

2. $\text{Max} \mid E_1^2[X(p_1 + p_2), k_3(p_3)] \mid$ was obtained by graphical means, namely drawing the slopes of the appropriate curve in section 6, graphing these slopes and then taking off the maximum slopes of these curves.

3. $E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$ was replaced by its approximation,

$$E_1[a(p_1 + p_2), k_3(p_3)],$$

in the computation of the percentage error. This new bound will be denoted by $U_1^*[k_1(p_1), k_2(p_2), k_3(p_3)]$.

4. Carefully chosen values of $U_1^*[k_1(p_1), k_2(p_2), k_3(p_3)]$ were plotted on triangular coordinates, and contour lines interpolated and extrapolated to cover in large part the range of p_1, p_2 and p_3.

The use of the third of the above listed assumptions is no detriment to the usefulness of the results since

$$\frac{E_{1a}[\,] - E_1[\,]}{E_1[\,]} = \frac{\dfrac{E_{1a}[\,] - E_1[\,]}{E_{1a}[\,]}}{1 - \dfrac{E_{1a}[\,] - E_1[\,]}{E_{1a}[\,]}} \leq \frac{U_1^*[k_1(p_1), k_2(p_2), k_3(p_3)]}{100 - U_1^*[k_1(p_1), k_2(p_2), k_3(p_3)]},$$

where $E_{1a}[\ ] = E_1[a(p_1 + p_2), k_3(p_3)]$ and $E_1[\ ] = E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$. Since $U_1^*[\ ]$ is a monotone decrease function of k_3, this new bound on the percentage error is also monotone decreasing for increasing k_3. Absolute values were not required in this derivation since $E_{1a}[\ ]$ is always greater than or equal to $E_1[\ ]$, as is apparent from (5.1) and an examination of the curves of section 6. The contours of $U_1^*[5(p_1), 5(p_2), 5(p_3)]$ are shown in Fig. 1. The interpretation of this figure is very straightforward. For example, for $p_3 \leqslant .5$, the value

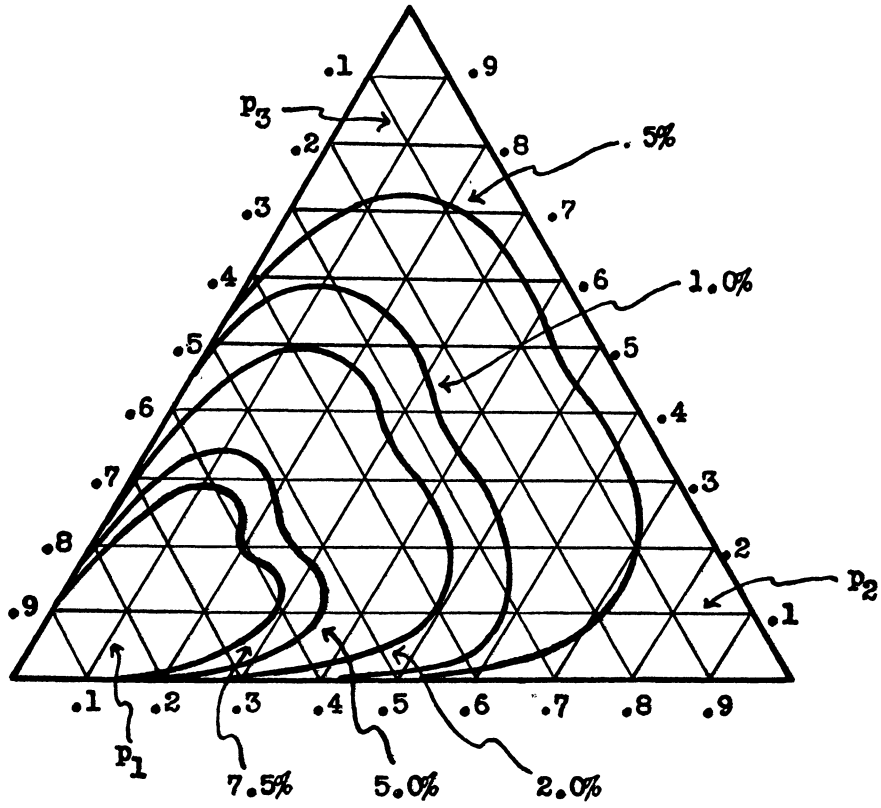


FIG. 1. CONTOURS OF $U_1^*[5(p_1), 5(p_2), 5(p_3)]$ CONSIDERED AS A FUNCTION OF p_1, p_2 AND p_3

of $U_1^*[5(p_1), 5(p_2), 5(p_3)]$ is less than 5.0%. Making use of the definition of $U_1^*[\ ]$, and especially its monotone characteristic, one can then say: the approximation for $E_1[5(p_1), 5(p_2), k_3(p_3)]$, where $k_3 \geqslant 5$, $p_3 \leqslant .50$ is in error by not more than 5.3%. Moreover, as has been already observed $E_1[a(p_1 + p_2), k_3(p_3)]$ is always greater than or equal to $E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$.

It will be noticed from Fig. 1 that $U_1^*[\ ]$ is increasing steadily as p_3 approaches 1. It has been demonstrated by McCarthy [5] that this behavior of the upper bound does not mean that the percentage error itself becomes larger as p_3 ap-

proaches 1. As a matter of fact, for fixed k_1, k_2 and k_3, the percentage error approaches zero as p_3 approaches 1. However, this demonstration does not furnish any reasonable bounds with which to fill in the lower left hand corner of Fig. 1. This fact is not as serious as it may at first seem because there is nothing to prevent one from reordering the boxes. For example, consider $E_1[5(.2), 5(.2), 5(.6)]$. From Fig. 1, the error of the approximation for this quantity, namely $E_1[E_1[5(.5), 5(.5)](.4), 5(.6)]$, is not more than approximately

$$7.5/(100 - 7.5) = 8.1\%.$$

On the other hand this same figure shows that $E_1[E_1[5(.25), 5(.75)](.80), 5(.20)]$, which is also an approximation to $E_1[5(.2), 5(.2), 5(.6)]$, is in error by not more than approximately .8%. Consequently one would choose the second ordering.

The procedure which has been used to obtain an upper bound on the percentage error of the approximation to $E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$, k_1 and k_2 fixed and k_3 greater than or equal to that integer at which the bound is evaluated, can also be applied to $E_3[k_1(p_1), k_2(p_2), k_3(p_3)]$. All the assumptions remain the same and in this case the bounds corresponding to $U_1[\ ]$ and $U_1^*[\ ]$ are denoted by $U_3[\ ]$ and $U_3^*[\ ]$. As in the case of $U_1[\ ]$, we have

$$\frac{E_3[\,] - E_{3b}[\,]}{E_3[\,]} = \frac{\dfrac{E_3[\,] - E_{3b}[\,]}{E_{3b}[\,]}}{1 + \dfrac{E_3[\,] - E_{3b}[\,]}{E_{3b}[\,]}} \leq \frac{U_3^*[k_1(p_1), k_2(p_2), k_3(p_3)]}{100}.$$

Here the approximation, $E_2[b(p_1 + p_2), k_3(p_3)]$, is always less than or equal to the exact value, $E_3[k_1(p_1), k_2(p_2), k_3(p_3)]$. The contours of $U_3^*[5(p_1), 5(p_2), 5(p_3)]$ are shown in Fig. 2. In using $U_3^*[5(p_1), 5(p_2), 5(p_3)]$ it is sometimes advantageous to reorder the boxes. For example, consider $E_3[5(.2), 5(.2), 5(.6)]$. Fig. 2 shows that, as an approximation, $E_2[E_2[5(.5), 5(.5)](.4), 5(.6)]$ is in error by not more than approximately 9%. However, $E_2[E_2[5(.25), 5(.75)](.80), 5(.20)]$, which is also an approximation for $E_3[5(.2), 5(.2), 5(.6)]$, is in error by not more than about 7%. There is a gain here, but it is not as great as the corresponding situation for $E_1[5(.2), 5(.2), 5(.6)]$.

As has already been stated, one may minimize the error by correctly choosing the two boxes which are to be combined first. Some discussion will be given here of a procedure for choosing these two boxes. Of course an experimental scheme may be used which makes use of the fact that the approximation to $E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$ is always an overestimate. In other words, *that grouping is used which gives rise to the smallest value of the approximation.* However, this can be replaced by a few preliminary computations.

As can be seen from (5.1), the error of the approximation depends upon two quantities, namely the variance of the two box situation obtained by combining two of the boxes, and the maximum value of the second derivative of the curve representing the function $E_1[X(p_1 + p_2), k_3(p_3)]$ over the proper range of X values. The error will be zero of $E_1[X(p_1 + p_2), k_3(p_3)]$ is either a constant or

linear in X over the range of X values in which one is interested, that is $k_1 \leq X \leq k_1 + k_2 - 1$, $k_1 \leqslant k_2$. If this is not possible, then one wishes to make it as near so as possible, subject to the restriction that

$$\sigma_1^2[k_1(p_1/(p_1 + p_2)), k_2(p_2/(p_1 + p_2))]$$

is not unnecessarily large.

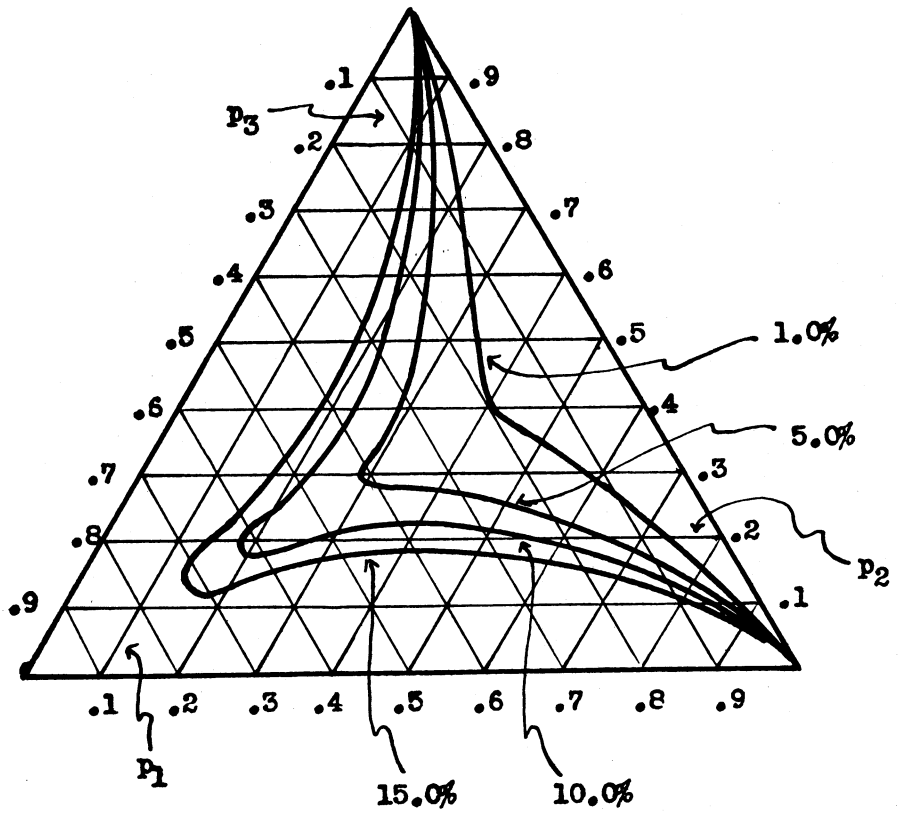


FIG. 2. CONTOURS OF $U_3^*[5(p_1), 5(p_2), 5(p_3)]$ CONSIDERED AS A FUNCTION OF p_1, p_2 AND p_3

An indication of the relationship between the boxes for both linearity and contribution to variance can be obtained from expressions (3.10) and (3.11). Thus for each box one computes k_i/p_i and $k_i(1 - p_i)/p_i^2$. Then in order to most nearly achieve linearity one orders the boxes in accordance with the increasing order of k_i/p_i and combines them in that order. If there is a tie between two or more boxes with respect to the k_i/p_i ordering, then one orders these "tied" boxes in accordance with increasing $k_i(1 - p_i)/p_i^2$.

Some computations have been carried out to illustrate these points and they are given in Table 1. The notation ((2, 4), 6) means that one first combines the boxes with integers 2 and 4, and then combines this result with the box with

associated integer 6. All values in this table were obtained by direct computation. No use of the curves was made.

In these three situations, one obtains the values given in Table 2.

Thus in the first case there is nothing to choose with respect to k_i/p_i, but $k_i(1 - p_i)/p_i^2$ indicates the ordering ((6, 4), 2). Actually the percentage error in this instance is 1.0 as compared with 1.7 and 2.4 for the other two orderings. In case two, k_i/p_i indicates the ordering ((2, 6), 4). Although this does not turn out to be the best ordering, Table 1 shows that the ordering in this instance makes little difference. In the last case, the indicated ordering is ((2, 4), 6) and the percentage error for this is zero, as opposed to 1.3 and 1.6. Since at any stage in the operation of combining boxes two at a time (4.13) holds, the

TABLE 1

Effect of Order of Combination on Error of Approximation

p_1 $1/6$ k_1	p_2 $1/3$ k_2	p_3 $1/2$ k_3	$E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$	% Error of Approximation		
				Order of Combination		
				((2, 4), 6)	((2, 6), 4)	((4, 6), 2)
2	4	6	6.96	+1.7	+2.4	+1.0
4	6	2	3.92	+0.3	+0.5	+0.5
6	4	2	3.77	+0.0	+1.3	+1.6

TABLE 2

p_i	1/6	1/3	1/2	1/6	1/3	1/2	1/6	1/3	1/2
k_i	2	4	6	4	6	2	6	4	2
k_i/p_i	12	12	12	24	18	4	36	12	4
$k_i(1 - p_i)/p_i^2$	60	24	12	120	36	4	180	24	4

above procedure will also give the minimum error for the approximation to $E_3[k_1(p_1), k_2(p_2), k_3(p_3)]$. *Moreover, the approximation for this quantity is always an underestimate of the true value, and therefore that ordering should be taken which gives the greatest value for the approximation.*

When the error of the approximation to $E_1[k_1(p_1), \cdots, k_n(p_n)]$ and

$$E_n[k_1(p_1), \cdots, k_n(p_n)],$$

for n greater than three, is considered, it is immediately obvious that the general considerations already given in this section still apply. In addition to these considerations, there is the difficulty that errors may cumulate. However, the results already quoted for three boxes, in conjunction with those which are to be given in 5.3, indicate that this cumulation is not serious. There are two factors which eventually prevent (i.e. as more and more boxes are considered) this percentage error from becoming unduly large, and, in fact, make it approach zero. These are:

1. The value of p_3 will, in most instances, be decreasing as more and more boxes are considered (see Fig. 1), and

2. The true value is usually becoming larger and larger as more and more boxes are considered.

In order to minimize the error, the following precautions should be taken:

1. At each stage in the computation, try to avoid, as much as possible, making readings where $E_1[X(p_1' + p_2'), k_3(p_3')]$ is curving sharply. If all readings are made where the curves are nearly linear, the percentage error will be very close to zero. On the other hand, if many readings must be made where the slopes of the curves are changing most sharply, larger errors must be expected.

2. Use that ordering of the boxes which provides the minimum value for the approximation to $E_1[\ ]$ or the maximum value for the approximation to $E_n[\ ]$.

3. In order to approximate the ordering which (2) would give, compute k_i/p_i and $k_i(1 - p_i)/p_i^2$ at each stage at which two boxes are to be combined and use the rules of procedure already given for three boxes.

5.3. *Error of the approximation for* $E_s[k^n]$. Repeated applications of the reduction formulae (4.34) and (4.35) allow one to evaluate $E_s[k^n]$ by means of the solution for the two box case, or more explicitly, by means of the curves given in section 6. Here the error of this approximation will be discussed primarily from a computational point of view.

$E_s[1^n]$ can be treated in detail since it is possible to obtain exact values for this expression by means of (4.1). This has been done by McCarthy [5], but the details will not be repeated here because of lack of space. The results simply add more credence to the conjectures which will soon be made.

When k is taken to be larger than one, the difficulty arises that it is almost impossible to compute the exact value of $E_s[k^n]$ in a large number of cases. Consequently it was necessary to devise an experimental model to estimate these exact values so that the amount of error would be known within bounds. A set of 10,000 punched cards[1] was obtained on which were recorded 100,000 random numbers drawn from a rectangular distribution. Thus if the cards are ordered on a particular set of columns, and one reads off the digits 0–9 on another specified column, one card at a time, it is equivalent to using a table of random numbers such as those prepared by Tippett [7]. By the use of these cards, it was possible to run off on an IBM Tabulator any desired number of experiments in order to obtain an experimental distribution from which to calculate an estimate of $E_s[k^n]$ and the variance of this estimate. For example, in determining an estimate of $E_1[2^5]$ one hundred experimental trials were made, as described above, with the following results:

Number of Trials Required	Frequency
2	23
3	32
4	31
5	11
6	3

[1] These punched cards were prepared at the Mayo Clinic, Rochester, Minn., under the direction of Doctor Joseph Berkson.

From this distribution the estimate of $E_1[2^5]$ is 3.39, with a variance computed from the distribution of .011. The 95% symmetric confidence limits for the mean, computed from the Student t-distribution, are 3.17 and 3.61. Such estimates will be used in the remainder of this section. It should be pointed out that in order to prevent a prohibitive amount of machine time, it was

TABLE 3

Percentage Errors for $E_s[k^n]$

s	k	n 3	4	5
1	1	−	−	−
	2	+ .7	+ 2.2	− .3 +13.6
	5	+ 1.1	− 3.1 +5.7	+ .6 +10.7
	10			− 2.9 + 5.1
2	1	− 5.6	− .4	+ 1.3
	2	− 4.6	− 4.4 +4.4	+ .6 +10.4
	5	−4.6 +1.7	+ 3.0 +9.3	+ 7.9 +14.8
	10	−3.7 +2.1	− .3 +5.5	+ 4.3 +10.7
	15		+ 1.0 +7.2	
	20	−2.5 +2.4		
3	1	−18.2	−12.7	− 3.1
	2	− 6.3	−16.5 −7.3	−2.9 + 6.0
	5	−9.7 −2.2	−10.7 −5.5	+ .8 + 5.8
	10			− 2.1 + 3.1
4	1		−12.0	−15.6
	2		−13.6 +6.1	−11.6 − 3.9
	5		−13.9 −7.2	− 9.9 − 4.0
	10		− 8.9 −2.6	− 6.4 − 1.2
5	1			−8.8
	2			−18.1 − 6.0
	5			−12.5 − 5.6
	10			− 8.9 − 2.9

necessary to use many of the same runs to determine values of $E_s[k^n]$ for different values of s, k and n. This means that the errors are correlated to some slight extent, but it would be extremely difficult to determine how much.

A summary of the computed percentage errors for various values of s, k and n is given in Table 3. In the instances where there are two entries, they are calculated on the basis of the 95% confidence limits for the experimental mean. These confidence limits are symmetric and were determined by using the Student t-distribution. For k equal to 2 and 5 the distribution contained 100 trials,

while for k greater than 5, the distribution were made up of approximately 50 trials.

The computations given in this table show for various values of s, k and n, the percentage error of the approximation for $E_s[k^n]$. In addition to showing the values of these percentage errors, the computations lead one to conjecture that

1. For fixed s and k, there exists an n_0 such that for $n > n_0$ the absolute value of the percentage error of the approximation for $E_s[k^n]$ is a monotone decreasing function for increasing n. It was shown by McCarthy [5] that this absolute value approaches zero as n approaches infinity for $E_s[1^n]$, and in fact, that the difference between the true and approximate values approaches zero.

2. For fixed s and n, there exists a k_0 such that for $k > k_0$, the absolute value of the percentage error of the approximation for $E_s[k^n]$ is a monotone decreasing function for increasing k.

6. Computation.

6.1. *Curves to aid in the computation of* $E_1[k_1(p_1), k_2(p_2)]$. In 3.1 it was shown that $E_1[k_1(p_1), k_2(p_2)]$ is equal to

$$\frac{k_1}{p_1} I_{p_1}(k_1 + 1, k_2) + \frac{k_2}{p_2} I_{p_2}(k_2 + 1, k_1),$$

where $I_x(p, q)$ is the Incomplete Beta-Function as tabled by Karl Pearson [6]. There are three principal difficulties connected with the use of these tables as they apply to the approximations of this paper. These are:

1. The tables must be available,

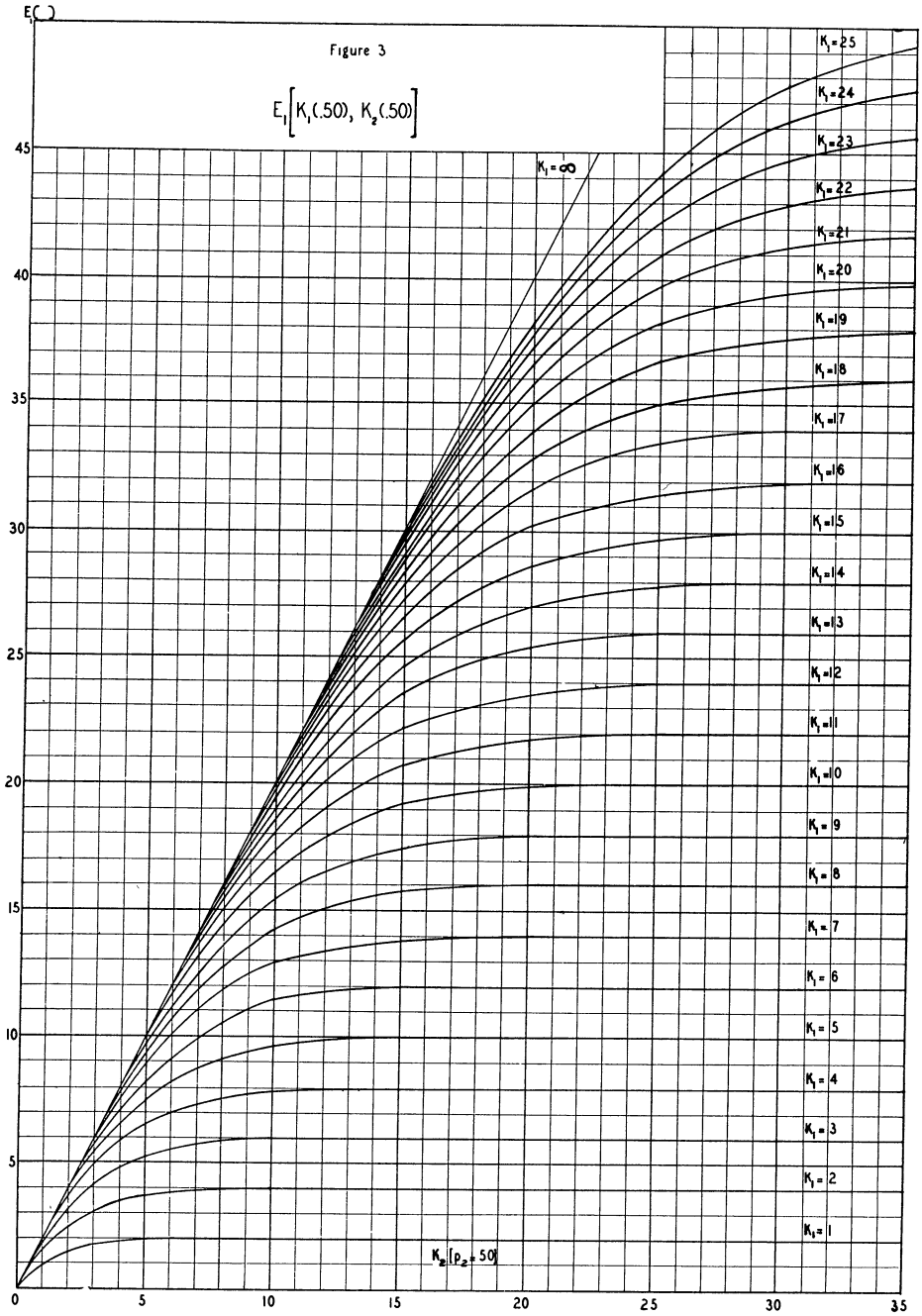
2. The tables give directly only values for integer or half-integer values of k_1 and k_2, and

3. Since many different values of $E_1[k_1(p_1), k_2(p_2)]$ are often required to obtain a single approximation, the computational burden would be very heavy.

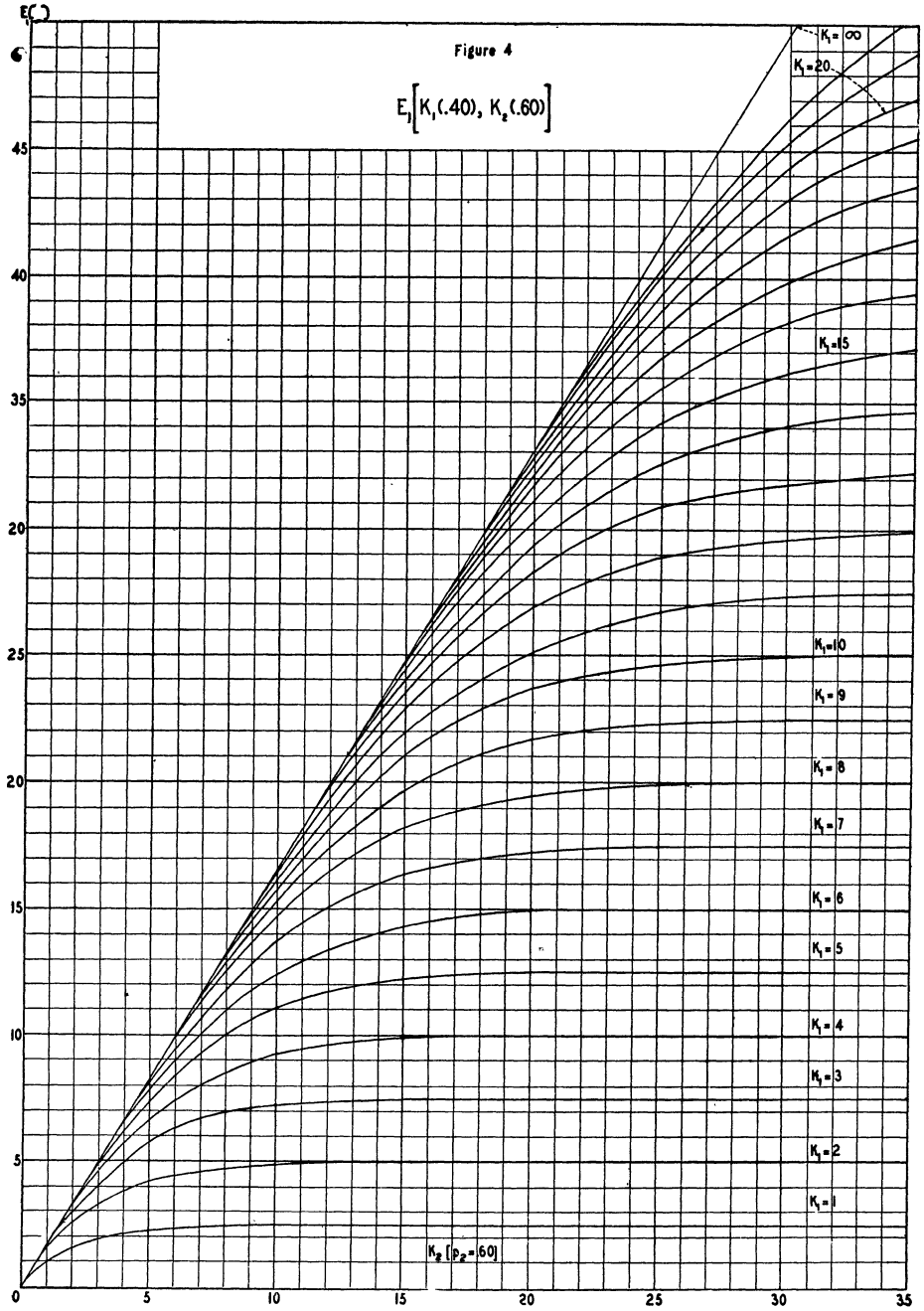
In order to surmount these difficulties, it seemed advisable to prepare curves giving the values of $E_1[k_1(p_1), k_2(p_2)]$ for various values of k_1, k_2, p_1 and p_2. These curves would give values of $E_1[\ ]$ with sufficient accuracy for most problems not only for integer values of k_1 and k_2, but for all values over the range considered.

Such curves have been prepared by computing $E_1[k_1(p_1), k_2(p_2)]$ for integral values of k_1 and k_2 (for fixed p_1 and p_2) and then joining these points with a smooth curve. A summary of the graphs prepared is as follows:

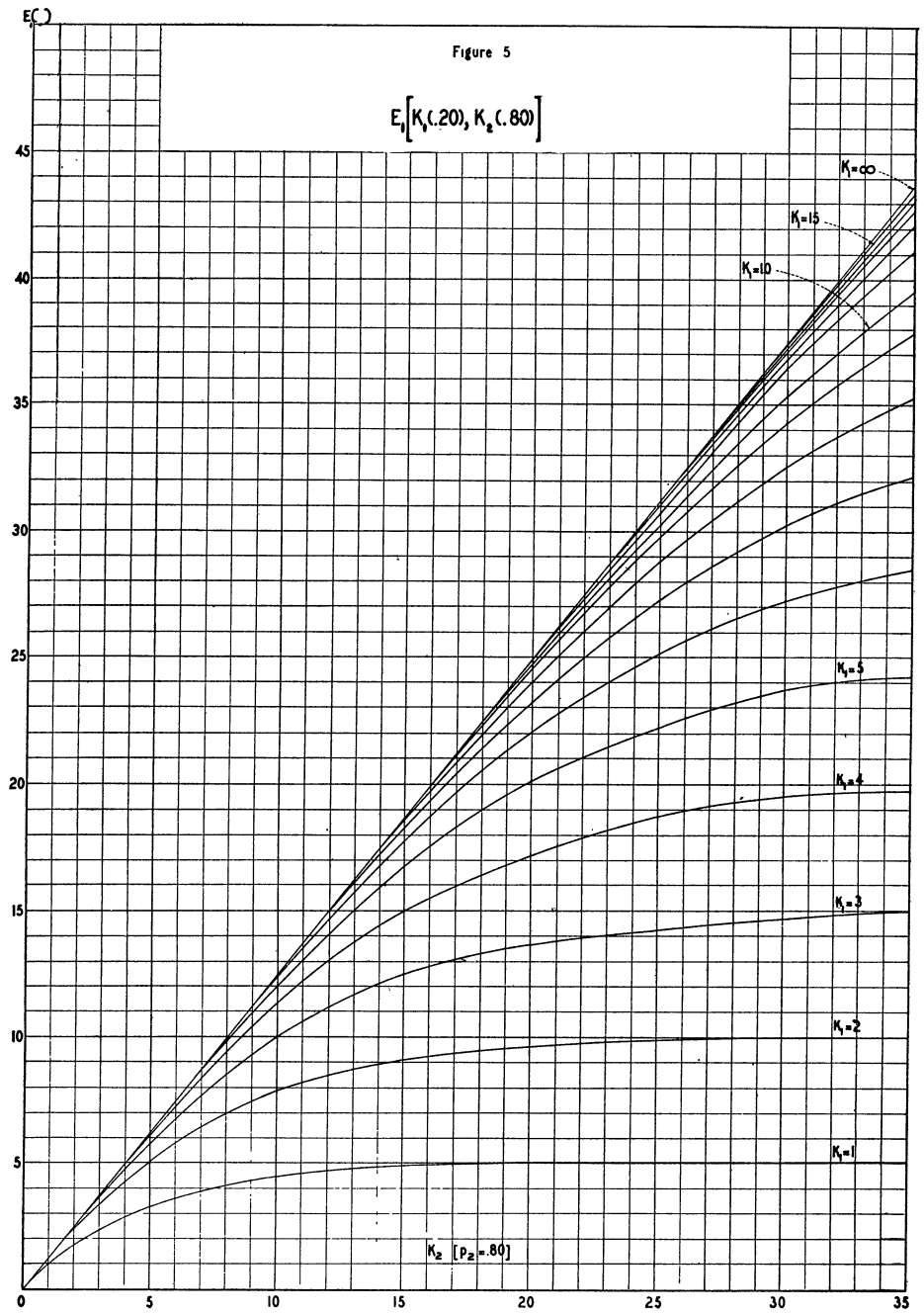
	k_1	k_2	p_1	p_2
Fig. 3	1, 2, $\cdots$, 25, ∞	1, 2, $\cdots$, 35	.50	.50
Fig. 4	1, 2, $\cdots$, 20, ∞	1, 2, $\cdots$, 35	.40	.60
Fig. 5	1, 2, $\cdots$, 15, ∞	1, 2, $\cdots$, 35	.20	.80
Fig. 6	1, 2, $\cdots$, 10, ∞	1, 2, $\cdots$, 15	.80	.20
Fig. 7	1, 2, $\cdots$, 7, ∞	1, 2, $\cdots$, 15	.60	.40
Fig. 8	1, 2, $\cdots$, 8, ∞	1, 2, $\cdots$, 15	.50	.50
Fig. 9	1, 2, $\cdots$, 6, ∞	1, 2, $\cdots$, 15	.40	.60
Fig. 10	1, 2, $\cdots$, 5, ∞	1, 2, $\cdots$, 15	.20	.80



Figures 8, 9, and 10 are simply portions of figures 3, 4 and 5 drawn on an expanded scale in order to permit greater accuracy in reading the curves. Also figures 6 and 10 and figures 7 and 9 form pairs in that a member of one pair can

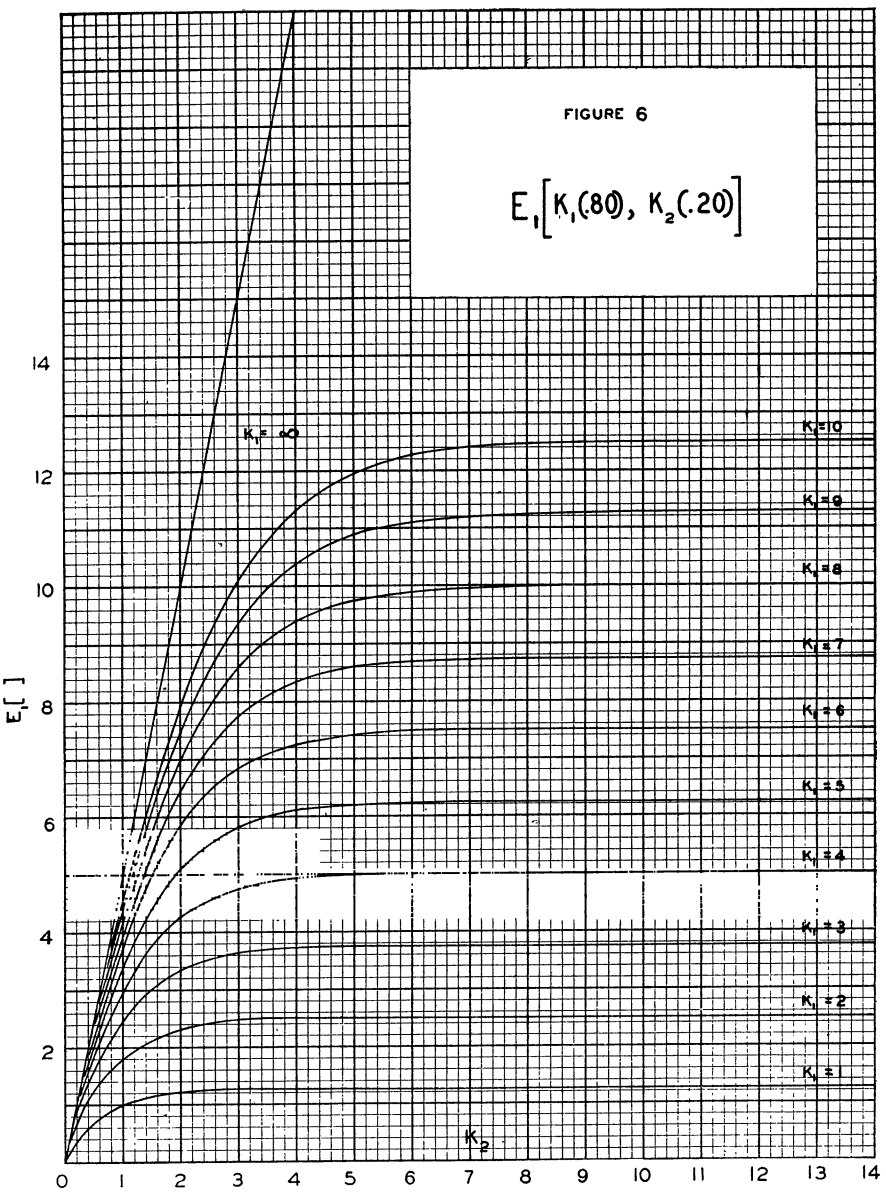


be obtained from the other member of the pair. Both members of the pair are given on the expanded scale in order to facilitate interpolation. Values of the mean for combinations of k_1 and k_2 not given directly can usually be obtained

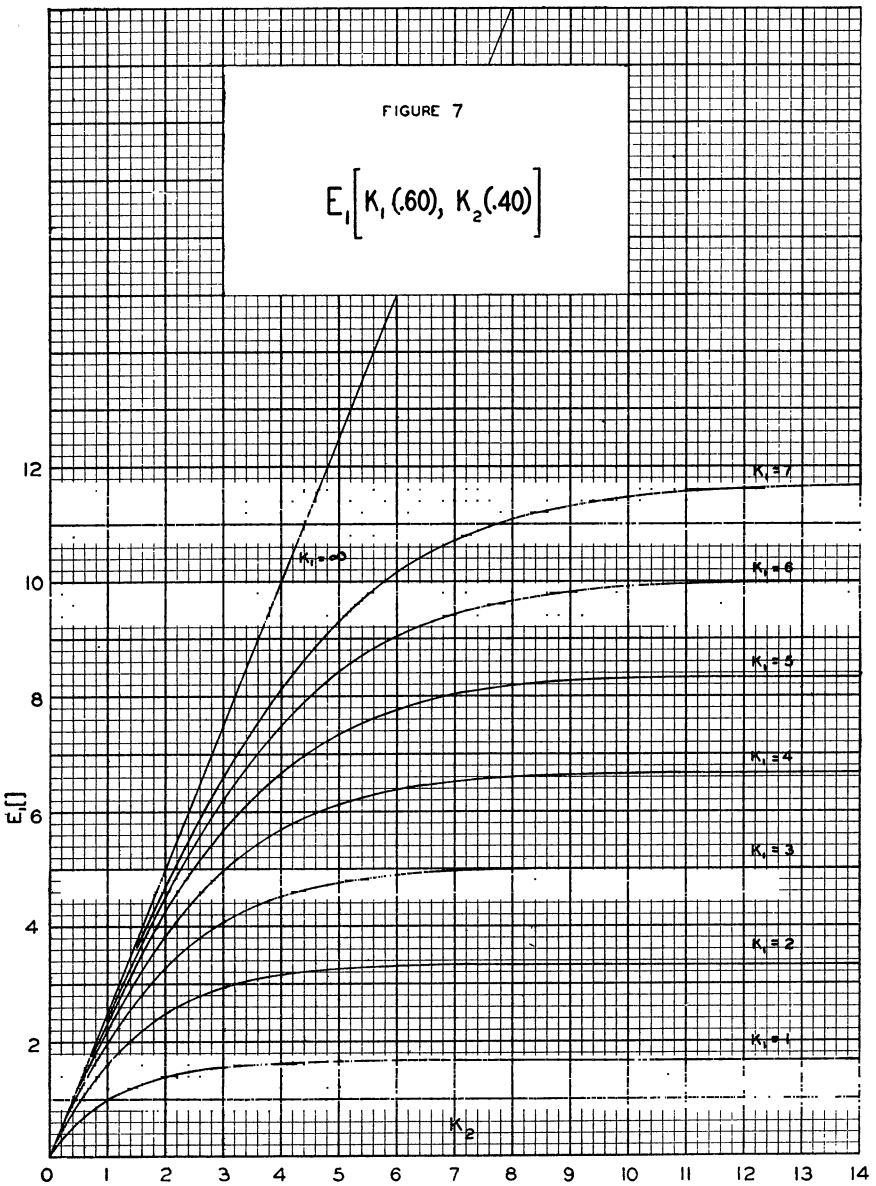


with sufficient accuracy with linear interpolation. Interpolation for p_1 and p_2 should be done graphically since in some instances linear interpolation would be extremely poor.

As an example, suppose one has two boxes with $k_1 = 2$, $k_2 = 5$, $p_1 = .40$ and $p_2 = .60$. Consulting Fig. 9, one goes along the horizontal axis to $k_2 = 5$.



Following up the vertical line through this point to the curve $k_1 = 2$, $E_1[2(.40), 5(.60)]$ is read as 4.25. The actually computed value to four decimals is 4.2224.

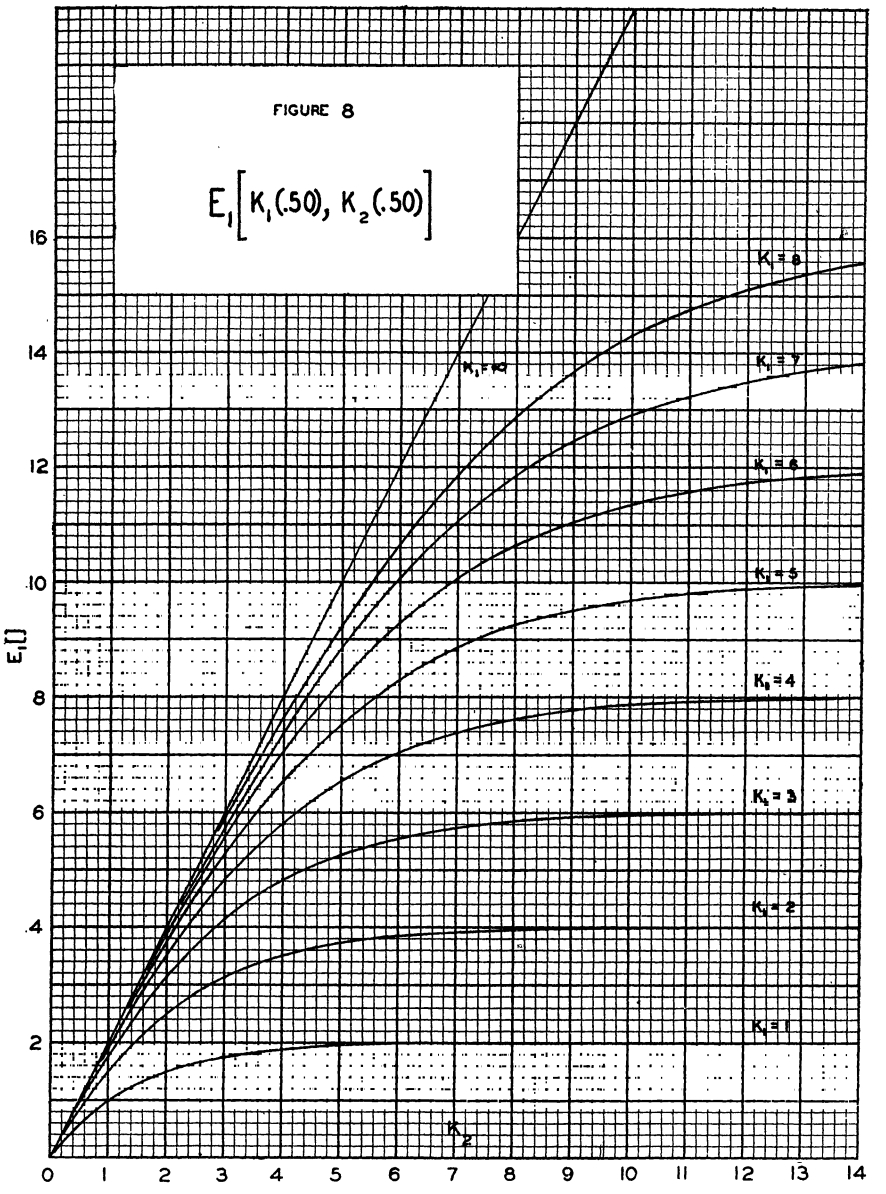


It is immediately evident that $E_2[k_1(p_1), k_2(p_2)]$ can also be obtained from the curves since

$$E_2[k_1(p_1), k_2(p_2)] = (k_1/p_1) + (k_2/p_2) - E_1[k_1(p_1), k_2(p_2)].$$

6.2. *Use of the curves to obtain exact values (i.e. subject only to the error of reading the curves) for* $E_1[k_1(p_1), k_2(p_2), k_3(p_3)]$. Referring back to (4.8), one obtains that

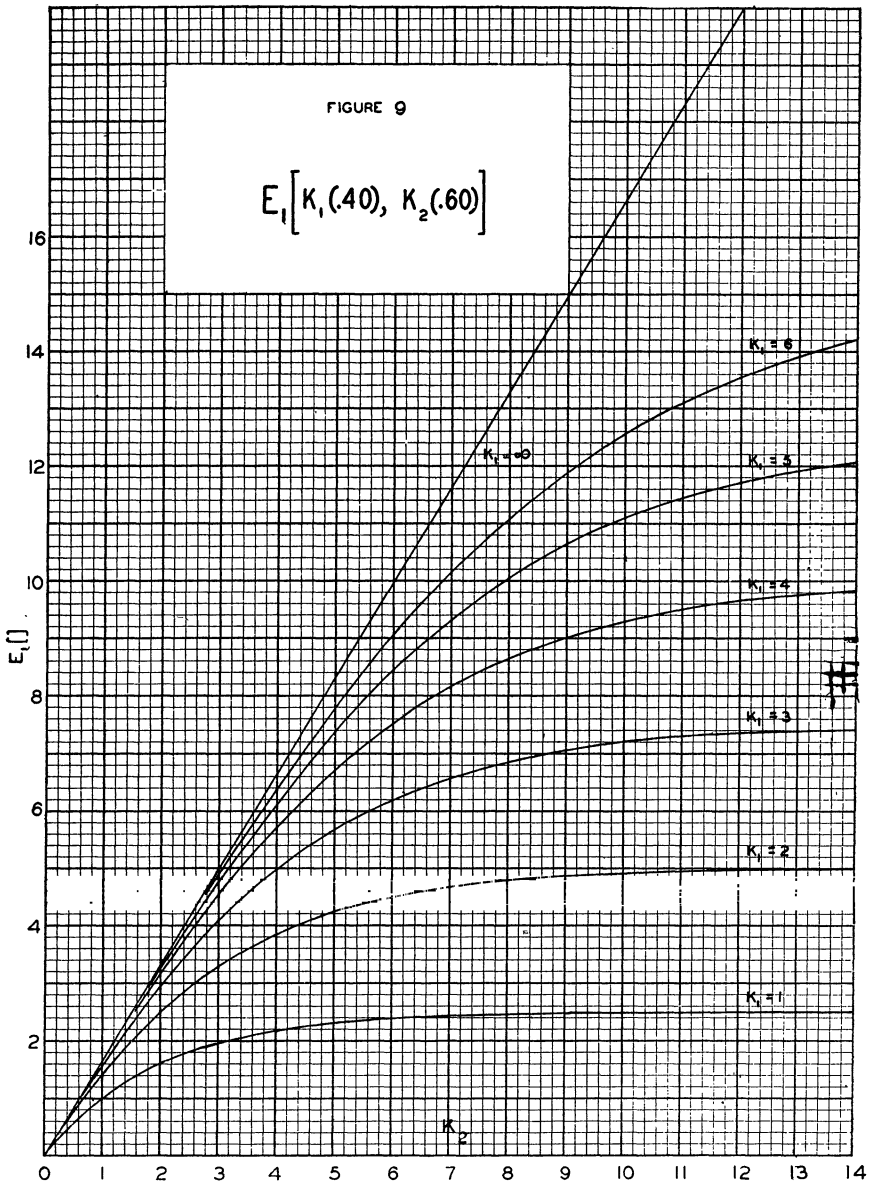
$$E_1[k_1(p_1), k_2(p_2), k_3(p_3)] = \sum_i \pi_i E_1[x_i(p_1 + p_2), k_3(p_3)], \tag{6.1}$$



where π_i is the probability that either k_1 balls are obtained in the first box or k_2 balls are obtained in the second box on the x_i *th* throw for the first time, assuming balls can go only in boxes one and two. x_i takes on values

$$k_1 , k_1 + 1, \cdots , k_1 + k_2 - 1$$

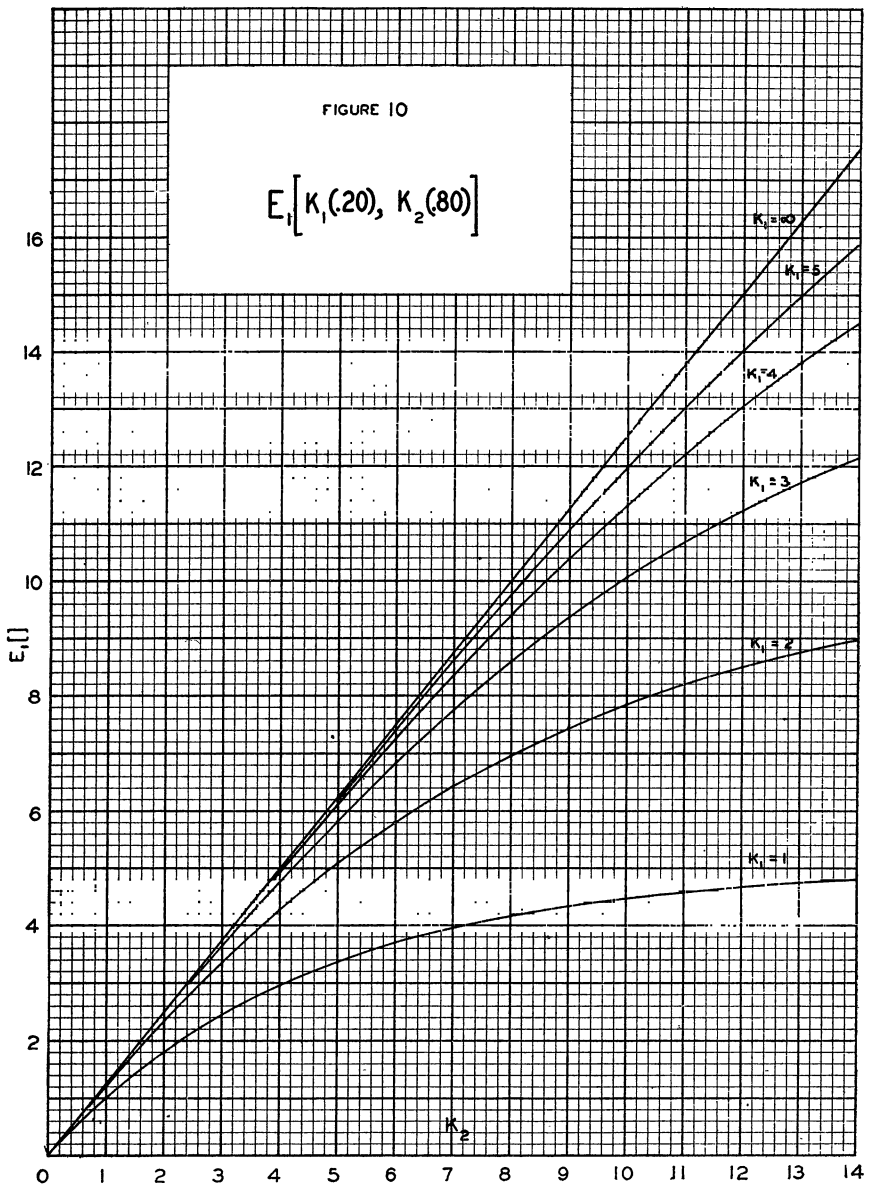
when $k_1 \leq k_2$. Now π_i can be easily computed and $E_1[x_i(p_1 + p_2), k_3(p_3)]$ can be obtained from the curves. The only difficulty in using this procedure



arises when the range of x_i is large. Then a large amount of computation is involved.

In order to illustrate this computation, consider $E_1[2(.1), 3(.1), 5(.8)]$. Here x_i takes on the values 2, 3 and 4. We have $x_1 = 2$, $\pi_1 = 2/8$; $x_2 = 3$, $\pi_2 = 3/8$; and $x_3 = 4$, $\pi_3 = 3/8$. From Fig. 6

$$E_1[2(.2), 5(.8)] = 5.09$$
$$E_1[3(.2), 5(.8)] = 5.88$$
$$E_1[4(.2), 5(.8)] = 6.11.$$



Consequently, $E_1[2(.1), 3(.1), 5(.8)]$ is equal to

$$(5.09)(2/8) + (5.88)(3/8) + (6.11)(3/8) = 5.77.$$

Using computed values for $E_1[x_i(.2), 5(.8)]$, $E_1[2(.1), 3(.1), 5(.8)]$ is equal to 5.75. Thus the use of the curves has only led to an error of .3%.

6.3. *Use of the curves in approximating* $E_1[k_1(p_1), \cdots, k_n(p_n)]$,

$$E_n[k_1(p_1), \cdots, k_n(p_n)]$$

and $E_s[k^n]$. In illustrating the application of the curves and the reduction formulae (4.34) and (4.35), one example will be worked through in detail. This example will provide illustrations of all the details involved in such problems. Consider $E_4[5^5]$. Applying formula (4.34)

$$E_4[5^5] \simeq 5/4\left\{E_1[5^4] + E_3\left[\left(\frac{4}{3}\cdot 5 - \frac{E_1[5^4]}{3}\right)^3, \quad \left(5 - \frac{E_1[5^4]}{4}\right)\right]\right\}. \tag{6.2}$$

Consequently, the first step must be to compute $E_1[5^4]$. Using the principles of 4.4

$$E_1[5^4] \simeq E_1[E_1[5^2](.50), 5(.25), 5(.25)]. \tag{6.3}$$

From Fig. 8, $E_1[5^2] = 7.55$. Therefore $E_1[5^4]$ is approximately equal to

$$E_1[7.55(.50), 5(.25), 5(.25)].$$

Now applying the same principle again,

$$E_1[5^4] \simeq E_1[E_1[7.55(\tfrac{2}{3}), 5(\tfrac{1}{3})](.75), 5(.25)]. \tag{6.4}$$

By the use of figures 7, 8, 9 and 10, graphical interpolation may be applied to find that $E_1[7.55(\frac{2}{3}), 5(\frac{1}{3})]$ is equal to 9.84. The approximation procedure now says that

$$E_1[5^4] \simeq E_1[9.84(.75), 5(.25)]. \tag{6.5}$$

Again applying the curves and using graphical interpolation for p_1 and p_2, $E_1[5^4] \simeq 11.88$.

Substituting this value in (6.2),

$$E_4[5^5] \simeq \tfrac{5}{4}\{11.88 + E_3[2.71, 2.71, 2.71, 2.03]\}. \tag{6.6}$$

Now formula (4.35) must be applied to $E_3[2.71, 2.71, 2.71, 2.03]$, i.e.

$$E_3[2.71, 2.71, 2.71, 2.03] \simeq$$

$$\tfrac{4}{3}\left\{E_1[(2.71)^3] + E_2\left[\left(\tfrac{3}{2}\,2.71 - \frac{E_1[(2.71)^3]}{2}\right)^2, \quad \left(2.03 - \frac{E_1[(2.71)^3]}{3}\right)\right]\right\}. \tag{6.7}$$

$E_1[(2.71)^3]$ can be evaluated by the same method used for $E_1[5^4]$. This leads to the result

$$E_3[2.71.\ 2.71, 2.71, 2.03] \simeq \tfrac{4}{3}\{4.40 + E_2[1.86, 1.86, .56]\}. \tag{6.8}$$

Once more applying (4.35)

$$E_2[1.86, 1.86, .56] \simeq$$

$$\tfrac{3}{2}\left\{E_1[(1.86)^2] + E_1[(2\cdot 1.86 - E_1[(1.86)^2]), \quad \left(.56 - \frac{E_1[(1.86)^2]}{2}\right)\right]\right\}. \tag{6.9}$$

$E_1[1.86, 1.86]$ is equal, by the curves, to 2.25. Therefore

$$E_2[1.86, 1.86, .56] \simeq \tfrac{3}{2}\{2.25 + E_1[1.47, -.56]\}. \tag{6.10}$$

However, since the convention is observed that a negative quantity is replaced by zero,

$$E_1[1.47, -.56] = E_1[1.47, 0] = 0. \tag{6.11}$$

Now working back through these various expressions,

$$E_4[5^5] \simeq \tfrac{5}{4}[11.88 + \tfrac{4}{3}[4.40 + \tfrac{3}{2}[2.25 + 0]]] = 27.81. \tag{6.12}$$

From Table 2 it can be seen that the percentage errors for this approximation to $E_4[5^5]$, corresponding to the 95% confidence limits for this quantity, are -4.0% and -9.9%.

This example has illustrated most of the situations which will arise in the use of the approximations of this paper.

6.4. *Miscellaneous approximation formulae useful for computation.* There exists a relatively simple approximation to $E_1[k_1(p_1), k_2(p_2)]$, $p_1 + p_2 = 1$, when p_2 is near one. Using (3.8) and making some obvious simplifications, one obtains

$$E_1[k_1(p_1), k_2(p_2)] = \frac{k_2}{p_2} + \frac{1}{p_2}\cdot\frac{(k_1 + k_2)!}{(k_1 - 1)!(k_2 - 1)!}\frac{1}{p_1}\int_0^{p_1} t^{k_1-1}(1 - t)^{k_2-1}(t - p_1)\, dt.$$

Since p_1 is near zero, $(1 - t)$ can be replaced by one, and the result is obtained that

$$E_1[k_1(p_1), k_2(p_2)] \simeq \frac{k_2}{p_2} - \frac{1}{p_2}p_1^{k_1}\frac{(k_1 + k_2)!}{(k_1 + 1)!(k_2 - 1)!}.$$

An approximation to the Incomplete Beta-Function, given by Tukey and Scheffé [8], may also prove useful at times. The expression, changed slightly by those authors since publication, is

$$I_b(n - r + 1, r) \simeq 1 - \frac{1}{2\Gamma(r)}\int_0^{\chi_\alpha^2}\left(\frac{\chi^2}{2}\right)^{r-1} e^{-(\chi^2/2)}\, d\chi^2,$$

where

$$\chi_\alpha^2 = 2r\left[\frac{(1 - b)\frac{n + \frac{1}{2}}{r} - 1}{\sqrt{b}}\right] + 2r.$$

The right hand side of the first expression will be recognized as the χ^2 distribution with $2r$ degrees of freedom. In the event that the tables of χ^2 are not adequate for the application of these expressions, the approximation of Wilson and Hilferty [10] should be used. This approximation states that $(\chi^2/\nu)^{\frac{1}{3}}$ where ν is the number of degrees of freedom, is approximately normally distributed with mean $1 - 2/(9\nu)$ and variance $2/(9\nu)$, for large ν.

7. Acknowledgements. The author wishes to express his grateful appreciation for the many helpful comments and suggestions received from Professors W. G. Cochran, A. M. Mood, J. W. Tukey and S. S. Wilks.

REFERENCES

[1] R. A. FISHER AND F. YATES, *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, London, 1943, Table XXII.

[2] M. A. GIRSHICK, FREDERICK MOSTELLER, AND L. J. SAVAGE, "Unbiased estimates for certain binomial sampling problems with applications", *Annals of Math. Stat.*, Vol. 17 (1946), pp. 13–23.

[3] J. B. S. HALDANE, "On a method of estimating frequencies", *Biometrika*, Vol. 33 (1945), pp. 222–225.

[4] P. S. LAPLACE, *Théorie Analytique des Probabilités*, Mme. Ve Courcier, Paris, 1820, pp. 194–219.

[5] P. J. MCCARTHY, *Approximate Solutions for Means and Variances in a Certain Class of Box Problems*, unpublished thesis, Library, Princeton University, 1946.

[6] KARL PEARSON, *Tables of the Incomplete Beta-Function*, The "Biometrika" Office, London, 1934.

[7] L. H. C. TIPPETT, *Random Sampling Numbers*, Cambridge Univ. Press, 1927.

[8] J. W. TUKEY, AND H. SCHEFFÉ, "A formula for sample sizes for population tolerance limits", *Annals of Math. Stat.*, Vol. 15 (1944), p. 217.

[9] J. V. USPENSKY, *Introduction to Mathematical Probability*, McGraw-Hill, 1937, p. 181.

[10] E. B. WILSON AND M. M. HILFERTY, "The distribution of chi-square", *Proc. Nat. Acad. Sci.*, Vol. 17 (1931), pp. 684–688.