Tables of the Hypergeometric Probability Distribution

GERALD J. LIEBERMAN
AND
DONALD B. OWEN

TABLES OF THE HYPERGEOMETRIC PROBABILITY DISTRIBUTION

GERALD J. LIEBERMAN *Stanford University*
DONALD B. OWEN *Sandia Corporation*

STANFORD UNIVERSITY PRESS
Stanford, California
1961

PREFACE

The hypergeometric probability distribution is not an easy function to compute, since reasonably accurate computation involves extensive factorial expansion with laborious calculations. For this reason, the function has usually been approximated by means of binomial, Poisson, and normal distributions.

In connection with another project, it was decided to undertake the programming and calculation of the hypergeometric distribution function for various sample and lot sizes. The sample and lot values were chosen to provide exact (six-decimal-place) point and cumulative probability values in the ranges where most sampling is done and where the usual approximations are poor.

The material presented here should be useful in many different disciplines. Research workers in the physical sciences and engineering, in industrial management, and in the social sciences, in particular, should find many applications of the distribution-free statistics whose probabilities can now be evaluated from these tables. Several of these statistics, including those based upon 2 × 2 tables and the number of exceedances, are discussed in the Introduction.

The mathematician and the statistician will find the equivalence of various sums of combinatorials of special interest. The results are similar to the well-known equivalence of the cumulative distribution of the negative binomial to the cumulative distribution of an ordinary binomial distribution. Shown for the first time, we believe, is the equivalence of the work on the hypergeometric distribution, per se, and the work on exceedance theory.

One author (G.J.L.) was partially supported in this study by the Office of Naval Research under Contract No. Nonr 225(53). The other author (D.B.O.) is at Sandia Corporation, a prime contractor to the United States Atomic Energy Commission; many of his contributions to this book first appeared in the form of technical memoranda issued by Sandia Corporation. These memoranda are listed in the Bibliography.

Drs. G. P. Steck and E. J. Gilbert contributed significantly to the Sandia Corporation memoranda as co-authors on two of them. Mr. C. M. Williams and Miss M. K. Weston, both of Sandia Corporation, programmed the computations of the logarithms of factorials and of the hypergeometric probability

distribution, respectively. Mrs. Marjorie E. Endres of Sandia Corporation made spot checks of the tables for accuracy and completeness. Miss Anna Glinski of Stanford University made some calculations on the approximations to the hypergeometric distribution. The authors are indebted to all of these persons for their help, although full responsibility for the accuracy of the results presented here rests with the authors.

G. J. Lieberman
D. B. Owen

Stanford, California
Albuquerque, New Mexico
December 1960

CONTENTS

I THE HYPERGEOMETRIC PROBABILITY DISTRIBUTION

1. Introduction 3
2. Applications 7
3. Approximations to the Hypergeometric Probability Distribution 16
4. Interpolation in the Hypergeometric Tables 23
5. Summary of Some Useful Formulas on Sums of Combinatorials 24

II TABLES OF THE HYPERGEOMETRIC PROBABILITY DISTRIBUTION

Table for $N = 2$, $n = 1$ through $N = 100$, $n = 50$ 33

Table for $N = 1000$, $n = 500$ 628

Table for $k = n - 1$, n; $n = N/2$; $N = 100$, $n = 50$ through $N = 2000$, $n = 1000$ 706

Appendix: Logarithms of Factorials 717

PART I

THE HYPERGEOMETRIC PROBABILITY DISTRIBUTION

THE HYPERGEOMETRIC PROBABILITY DISTRIBUTION

1. INTRODUCTION

1.1 The Hypergeometric Probability Distribution

Tabulations of the hypergeometric probability distribution have many potentially useful applications, some of which are not generally recognized as having any connection with the distribution. In Part I of this book we have attempted to present the theory and rationale of the hypergeometric probability distribution and to indicate as many of its specific applications as possible.

1.2 Definitions

The nomenclature of sampling inspection (one of the applications) is convenient to describe the parameters of the hypergeometric probability distribution and will be used here. The following symbols are defined:

N = number of items in a lot,
n = number of items in a sample taken from the lot,
k = number of defective items in the lot,
x = number of defective items observed in the sample.

Then the probability

$$\Pr\{\text{Exactly } x \text{ defectives are found in the sample}\} = p(x) = p(N, n, k, x) = \frac{k!\,n!}{(k-x)!\,(n-x)!\,x!}\,\frac{(N-k)!\,(N-n)!}{N!\,(N-k-n+x)!},$$

where x is an integer such that $\max[0, n+k-N] \leqq x \leqq \min[n, k]$, and

$$\Pr\{x \text{ defectives or fewer are found in the sample}\} = P(x) = P(N, n, k, x) = \sum_{i=\max[0,n+k-N]}^{x} \frac{k!\,n!}{(k-i)!\,(n-i)!\,i!}\,\frac{(N-k)!\,(N-n)!}{N!\,(N-k-n+i)!}$$

1.3 Calculation of the Tables

Included at the end of the hypergeometric tables is a table of $\log N!$ for $N = 1(1)2000$ taken from [50]. This table was put on magnetic tape (all 15 decimal places for each N) and a program was prepared for an IBM 704 computer to sum the proper logarithms to give the logarithms of the point probabilities to 15 decimal places. The point probabilities were obtained by taking antilogarithms correct to at least eight decimal places; usually nine decimal places were obtained. The cumulative probabilities were calculated by summing the point probabilities. The results were rounded off to six decimal places within the IBM 704 computer and printed. The tables given here were produced by photographic means from this output.

As a check on the accuracy of the tables we made calculations on desk calculators of randomly selected values from each set of 200. We found no discrepancies. The cumulative probabilities were checked to see that for each set of values N, n, k there was an entry equal to 1 for $x = k$ and $N \leq 100$. For $N > 100$, a check was made for

$$P(N, \tfrac{1}{2}N, k, \tfrac{1}{2}[k-1]) = \tfrac{1}{2},$$

where N is even and k is odd.

1.4 Symmetries and Check

Since n and k may be interchanged in either of the probabilities $P(x)$ and $p(x)$ without changing the values of the probabilities, it is necessary to tabulate only for $k \leq n$. *If $n < k$, it is necessary only to interchange n and k to read the probabilities directly from the tables.*

This volume tabulates the hypergeometric distribution for $N = 1(1)49$, $50(10)100$, and 1000. The values for $N = 1000$ are given only for $n = 500$. Some values for $N = 100(100)2000$ are also given.

All possible hypergeometric probabilitity distributions with $N \leq 25$ are tabulated below, and the only symmetry that may have to be used in entering the table is the one on n and k mentioned above.

For $N > 25$, three additional symmetries were taken into account in order to keep the table to a reasonable size. These symmetries are given by the following equations. For point probabilities,

$$\begin{aligned} p(N, n, k, x) &= p(N, n, N-k, n-x) \\ &= p(N, N-n, k, k-x) \\ &= p(N, N-n, N-k, N-n-k+x); \end{aligned}$$

and for cumulative probabilities,

$$\begin{aligned} P(N, n, k, x) &= P(N, N-n, N-k, N-n-k+x) \\ &= 1 - P(N, n, N-k, n-x-1) \\ &= 1 - P(N, N-n, k, k-x-1), \end{aligned}$$

where the value of $P(N, n, k, x)$ is 1 if either $n - x - 1$ or $k - x - 1$ becomes negative. These three symmetries are immediately obvious if one considers the effect of interchanging the roles of defective and nondefective, and if one keeps in mind also that n and k can always be interchanged. They may be proved formally by substituting in the defining equations for $p(N, n, k, x)$ and $P(N, n, k, x)$ and showing that the resulting factorials and

TABLE I

N	n	k	x	$P(x)$	$p(x)$
			Set I		
16	6	4	0	0.115385	0.115385
16	6	4	1	0.510989	0.395604
16	6	4	2	0.881868	0.370879
16	6	4	3	0.991758	0.109890
16	6	4	4	1.000000	0.008242
			Set II		
16	10	4	0	0.008242	0.008242
16	10	4	1	0.118132	0.109890
16	10	4	2	0.489011	0.370879
16	10	4	3	0.884615	0.395604
16	10	4	4	1.000000	0.115385
			Set III		
16	12	6	2	0.008242	0.008242
16	12	6	3	0.118132	0.109890
16	12	6	4	0.489011	0.370879
16	12	6	5	0.884615	0.395604
16	12	6	6	1.000000	0.115385
			Set IV		
16	12	10	6	0.115385	0.115385
16	12	10	7	0.510989	0.395604
16	12	10	8	0.881868	0.370879
16	12	10	9	0.991758	0.109890
16	12	10	10	1.000000	0.008242

sums of factorials are equal. The symmetry involving n and k may be written

$$p(N, n, k, x) = p(N, k, n, x) \quad \text{and} \quad P(N, n, k, x) = P(N, k, n, x) .$$

The usefulness of this symmetry cannot be overemphasized and should always be kept in mind when dealing with the hypergeometric probability distribution.

As an example of the usefulness of the three symmetries mentioned above, consider the four sets of equivalent entries given in Table I. Clearly, when one of the sets of values is given, it will be easy to obtain the other three. These symmetries were used in producing the table for $N > 25$ in order to reduce the number of entries. This involved restricting the parameters so that $x \leq k \leq n \leq \frac{1}{2}N$. These restrictions eliminated practically all of the duplications, the exceptions being values such that N is even and $n = \frac{1}{2}N$, where the distributions themselves are symmetric about $x = \frac{1}{2}k$. In this connection, for N even, $n = \frac{1}{2}N$, k odd, and $x = \frac{1}{2}(k-1)$, the value of $P(N, n, k, x)$ is $\frac{1}{2}$. This computation was found to be a useful and quick check on parts of the tables. Another useful and quick check on the tables is given by

$$\sum_{x=\max[0,n+k-N]}^{k} P(x) = 1 + k\left(1 - \frac{n}{N}\right), \quad \text{for } n \geqq k .$$

The tables give all possible hypergeometric probability distributions for $N \leqq 25$. For $25 < N \leqq 50$ all possible distributions are given except that one of the set of three symmetries may have to be used to find the proper table entry. For $N = 60(10)100$ all values also are given except that one of the set of three symmetries may have to be used. For $N = 1000$, $n = 500$, all distributions are given except that point probabilities equal to 0 to six decimal places (0.000000) are not explicitly given and one of the set of three symmetries may have to be used. A table of probabilities is also given for $N = 100(100)2000$, $n = \frac{1}{2}N$ and $k = n - 1$, n, and all values of x.

As an example of the use of the symmetries in looking up probabilities in the table, consider the problem of finding $P(50, 40, 30, 20)$. The three equivalent values of the parameters are

$$P(N, N-n, N-k, N-n-k+x) = P(50, 10, 20, 0) ,$$
$$1 - P(N, n, N-k, n-x-1) = 1 - P(50, 40, 20, 19) ,$$
$$1 - P(N, N-n, k, k-x-1) = 1 - P(50, 10, 30, 9) .$$

The value that can be found in the table is the one with $n \leqq \frac{1}{2}N$ and $k \leqq n$. Therefore the appropriate value given in the table is $P(50, 20, 10, 0) = 0.002925$. Note the interchange of n and k. Hence we find $P(50, 40, 30, 20) = 0.002925$.

As a second example, suppose the value of $P(1000, 565, 500, 287)$ is needed. Note that this can be obtained from the tables, since n and k can be interchanged and the tables include all values of $N = 1000$, $n = 500$ which are greater than 0 to six decimal places and less than 1. The equivalent values are $P(1000, 435, 500, 222)$, $1 - P(1000, 565, 500, 277)$, and $1 - P(1000, 435, 500, 212)$. Now the value sought may be obtained from $1 - P(1000, 500, 435, 212) = 1 - 0.261791$. Hence we have $P(1000, 565, 500, 287) = 0.738209$.

1.5 Number of Entries in the Tables

If only the symmetry given by the interchangeability of n and k is taken into account, then the number of entries in a table of the hypergeometric distribution beginning with $N = 2$ is given by

$$S = \frac{N^4 + 12N^3 + 2N^2 - 12N - \begin{cases} 0 \text{ if } N \text{ is even} \\ 3 \text{ if } N \text{ is odd} \end{cases}}{48} .$$

If all entries such that $0 \leqq x \leqq k \leqq n \leqq \frac{1}{2}N$ are tabulated beginning with $N = 2$ up to $N = N$, the number of entries for even values of N is

$$S^* = \frac{1}{192}[N(N+2)(N^2 + 14N + 16)] ,$$

and for odd values of N,

$$S^* = \frac{1}{192}[(N-1)(N+1)(N+3)(N+13)]\,.$$

Thus for the tables given here, the number of entries for $N \leqq 25$ is $S = 12{,}064$; the number of entries from $26 \leqq N \leqq 50$ is $S^* = 43{,}550 - 3{,}458 = 40{,}092$. Similarly, the number of entries for $N = 60, 70, 80, 90$, and 100 can be computed, and for these values of N the number of entries is 66,750. Through $N = 100$, the total number of entries in the tables given here is then 118,906 entries. Since zero entries were eliminated for $N = 1000$, $n = 500$, and for $N = 100(100)2000$, $n = \frac{1}{2}N$ and $k = n - 1$, n, it was not possible to predict the number of entries for these cases. There were, however, 15,433 entries for $N = 1000$, $n = 500$ and 1535 entries for the cases covered with $N = 100(100)2000$, giving a grand total of 135,874 entries in the hypergeometric tables given here.

2. APPLICATIONS

2.1 Applications to a Sequential Procedure

Given a lot of N items containing k defectives, a question that frequently arises is "How many items must be sampled from the lot to produce n nondefectives?" The solution to this problem may be obtained as follows:

Pr $\{x + n$ trials or fewer will be required to produce n nondefectives$\}$

$$= \frac{(N-k)!\,(N-n)!}{(N-k-n)!\,N!}\left[1 + n\frac{k}{N-n} + \frac{n(n+1)}{2}\,\frac{k(k-1)}{(N-n)(N-n-1)} + \cdots + \frac{n(n+1)\cdots(n+x-1)}{x!}\,\frac{k(k-1)\cdots(k-x+1)}{(N-n)(N-n-1)\cdots(N-n-x+1)}\right]$$

$$= \left[\binom{N-k}{n}\Big/\binom{N}{n}\right]\sum_{i=0}^{x}\binom{n+i-1}{i}\binom{k}{i}\Big/\binom{N-n}{i}$$

$$= \left[1\Big/\binom{N}{k}\right]\sum_{i=0}^{x}\binom{n+i-1}{n-1}\binom{N-n-i}{N-n-k},$$

where $0 \leqq x \leqq k$ and $N \geqq k + n$.

It can be shown that this probability reduces to

$$1 - P(N, x+n, N-k, n-1) = P(N, n+x, k, x)$$

by the following argument. Let N_n be the number of trials until n nondefectives first appear. Let y_h be the number of nondefectives in h trials. Suppose there are fewer than n nondefectives in the first h trials. Then more than h trials will be required to obtain n nondefectives. Also, if the number of trials needed to obtain n nondefectives is larger than h, then in the first h trials there will be fewer than n nondefectives. Since $N_n > h$ for exactly the same sequences for which $y_h < n$, it follows that $N_n \leqq h$ for exactly the same sequences for which $y_h \geqq n$.

Thus, we obtain

$$\begin{aligned}\Pr\{N_n \leqq h\} &= \Pr\{y_h \geqq n\} \\ &= 1 - P(N, x+n, N-k, n-1) \\ &= P(N, n+x, k, x)\,.\end{aligned}$$

This proves the identity

$$\sum_{j=0}^{N-n-k} \binom{k+j}{x}\binom{N-k-1-j}{n-1} \equiv \sum_{j=0}^{x} \binom{N-k}{n+x-j}\binom{k}{j},$$

which is a generalization of the identity (12.16) in Feller [16].

For example, suppose a lot of 50 items contains 10 defectives and it is necessary to obtain 20 nondefective items from the lot. The sampling will stop when the 20 nondefectives are obtained. What is the probability that the 20 nondefective items can be obtained with a sample of 25 or fewer? The answer is $P(50, 25, 10, 5) = 0.637399$.

The entire probability distribution for the possible sample sizes may be obtained from the hypergeometric table also. The distribution is given in Table II. Hence, for example, about 66 times out of 1000 it will be necessary to take a sample of 28 or more to obtain 20 nondefectives from the lot.

TABLE II

Sample size	Look up
20	$P(50, 20, 10, 0) = 0.002925$
21	$P(50, 21, 10, 1) = 0.022424$
22	$P(50, 22, 10, 2) = 0.085964$
23	$P(50, 23, 10, 3) = 0.219096$
24	$P(50, 24, 10, 4) = 0.417561$
25	$P(50, 25, 10, 5) = 0.637399$
26	$P(50, 26, 10, 6) = 0.820598$
27	$P(50, 27, 10, 7) = 0.934006$
28	$P(50, 28, 10, 8) = 0.983930$
29	$P(50, 29, 10, 9) = 0.998051$
30	$P(50, 30, 10, 10) = 1.000000$

As a second example, suppose a lot of 35 items is at hand, and it is necessary to obtain 20 nondefectives from this lot before sampling can cease. When sampling is stopped, it is necessary to make a statement about the number of defectives in the original lot, e.g., that the number of defectives in the lot is no more than k with 90 per cent assurance.

This problem may be solved by solving the inequality $P(35, x+20, k, x) \leqq 0.10$ for k. The results are shown in Table III. Now, for example, it can be said that if sampling stopped with a sample of 25 items to produce 20 nondefective items, one is at least 90 per cent sure that there were no more than 10 defective items in the lot before sampling; or equivalently, one is 90 per cent sure that there are no more than 5 defective items in the remaining 10 items.

TABLE III

Sample size taken	x	k	Actual probability
20	0	3	0.070
21	1	5	0.071
22	2	7	0.050
23	3	8	0.070
24	4	9	0.084
25	5	10	0.089
26	6	11	0.084
27	7	12	0.070
28	8	13	0.050
29	9	14	0.028
30	10	14	0.070
31	11	15	0.026
32	12	15	0.070

It might be instructive to compare this case with the case of an infinite lot. Suppose a sample of 25 is taken from a continuous production process to produce 20 nondefective items. With 90 per cent assurance, what is the upper bound on the proportion of defective items coming from this process? The answer is obtained by computing the following upper confidence limit on a proportion where sampling is from a negative binomial distribution:

$$1 - \frac{n+1}{n+1+xF_{\gamma,2x,2n+2}},$$

where F_γ is an upper γ percentage point of the F distribution based on $2x$ degrees of freedom for the numerator and $2n+2$ degrees of freedom for the denominator. Here $n = 20$ and $x = 5$; hence $F_{0.90} = 1.75$ and the upper bound on the proportion of defectives produced by the process is $1 - 0.706 = 0.294$ with 90 per cent assurance. For the finite lot with $N = 35$, as pointed out above, the proportion of defectives in the original lot is less than $10/35 = 0.286$ with at least 90 per cent assurance, or the proportion of defectives in the remaining part of the lot is less than $5/10 = 0.500$ with at least 90 per cent assurance.

2.2 Applications to Tests of the Equality of Two Proportions (2 x 2 Tables)

A 2×2 contingency table is represented below.

Characteristic II	Characteristic I: Has	Characteristic I: Does not have	Totals
Has	x	$k - x$	k
Does not have	$n - x$	$N - n - k + x$	$N - k$
	n	$N - n$	N

The analysis considered here of this type of table is due to Fisher [19], p. 85; various probability tables have been prepared for easy application of the method [17], [18], [33], [37], and [62].

An example will make clear the usefulness of the hypergeometric table in testing a 2×2 table. We wish to evaluate two methods of encapsulating small batteries in plastic. The data are presented as follows:

Treatment	Performance		Totals
	Failure	Success	
Encapsulation Method I	9	6	15
Encapsulation Method II	3	11	14
	12	17	29

Hence, take $n = 12$, $k = 15$, $x = 9$, and $N = 29$. But in order to read this from the hypergeometric table, it is necessary to interchange k and n and to make use of the symmetry $P(N, n, k, x) = 1 - P(N, N - n, k, k - x - 1)$. Entries are now taken from the main table as given in Table IV.

TABLE IV

N	n	k	x	$P(x)$	$p(x)$
29	15	12	0	0.000002	0.000002
29	15	12	1	0.000107	0.000105
29	15	12	2	0.002132	0.002025
29	15	12	3	0.019685	0.017553
29	15	12	4	0.098672	0.078987
29	15	12	5	0.297267	0.198595
29	15	12	6	0.586885	0.289618
29	15	12	7	0.835130	0.248244
29	15	12	8	0.959252	0.124122
29	15	12	9	0.994357	0.035105
29	15	12	10	0.999623	0.005266
29	15	12	11	0.999991	0.000368
29	15	12	12	1.000000	0.000009

The question to be answered is "Does Method II have a better effect on performance than Method I?" This is a one-sided test since the outcome is interesting only if Method II is better than Method I. The relative proportion of successes using Method II is $11/14 = 0.79$ and for Method I is $6/15 = 0.40$. Hence, in this sample Method II shows a better performance than Method I. The question now becomes "Is this due to chance or is Method II really better than Method I?" Next, the statistical test will be performed to determine if Method II is indeed better than Method I. If Method II were worse than Method I in the sample, then no further statistical test would be performed because the hypothesis that Method II is no better than Method I is automatically accepted. Note that the procedure outlined above enables one to make a one-tailed test. For a two-tailed test no preliminary look at the proportions is necessary.

The probability of observing exactly nine failures is then $p(x) = 0.035105$. But it is necessary to find the probability of nine or more failures (a deviation as extreme as, or more extreme than, the one observed), and this can be obtained from the table by taking

$$1 - \Pr\{x \leqq 8\} = 1 - 0.959252 = 0.040748 .$$

Since this probability is less than 0.05, there is a significant difference between Method I and Method II at the 95 per cent level of significance.

If in the problem solved above there were no prior understanding that the outcome would be interesting only if Method II were better than Method I, then a two-sided test should be run. Looking at Table IV, one should say there is a significant difference at the 95 per cent level of significance only if $x \leqq 3$ or if $x \geqq 10$, since

$$\begin{aligned}\Pr\{x \leqq 3\} + \Pr\{x \geqq 10\} &= 0.019685 + 1 - 0.994357\\ &= 0.025328 < 0.05\,.\end{aligned}$$

Table IV shows that x cannot be raised to 4, since then $P(x) = 0.098672$. But x could be lowered to 9 at the upper end if x were lowered to 2 at the lower end. That is, another two-sided rule for determining significance at the 95 per cent significance level is to conclude that Method I differs from Method II if $x \leqq 2$ or if $x \geqq 9$, since

$$\begin{aligned}\Pr\{x \leqq 2\} + \Pr\{x \geqq 9\} &= 0.002132 + 1 - 0.959252\\ &= 0.042881 < 0.05\,.\end{aligned}$$

Both rules, $x \leqq 3$ or $x \geqq 10$ and $x \leqq 2$ or $x \geqq 9$, are equally good. The second might be preferred to the first on the ground that the actual test probability is closer to 0.05. The first might be preferred to the second on the ground that both tails of the distribution are below 0.025. The difficulty, of course, arises because of the discreteness of x. For further discussion of this point as applied to the above problem see [62].

2.3 Applications to the Distribution of the Number of Exceedances

Consider a random sample of size n taken from a continuous distribution. Let another random sample of size m, independent of the first sample, be drawn from the same population. The probability that x observations among the observations of the second sample will exceed the rth-largest observation in the first sample is given by

$$\begin{aligned}&\Pr\{x \text{ among } m \text{ future trials will exceed the } r\text{th-largest observation in a sample of } n\}\\ &\quad = \binom{m+n-r-x}{n-r}\binom{x+r-1}{r-1}\Big/\binom{m+n}{n}\\ &\quad = \frac{n}{m+n}\,p(m+n-1, m, x+r-1, x)\\ &\quad = \frac{r}{x+r}\,p(m+n, m, x+r, x)\,,\end{aligned}$$

where $p(m+n, m, x+r, x)$ is the quantity defined in Section 1.2 and is equal to $p(x)$ in the hypergeometric table.

The probability that the largest among n past observations will be exceeded at most x times in m future trials is given by

$$\Pr\{x \text{ or fewer among } m \text{ future trials will exceed the largest among } n \text{ observations}\} = \sum_{y=0}^{x} \binom{m+n-y-1}{n-1} \Big/ \binom{m+n}{n} = 1 - \binom{m+n-1-x}{n} \Big/ \binom{m+n}{n}.$$

The summation of the binomial coefficients was accomplished by means of Equation (12.6) of [16]. Hence,

$$\Pr\{x \text{ or fewer among } m \text{ future trials will exceed the largest among } n \text{ observations}\} = 1 - P(m+n, n, x+1, 0), \qquad \text{for } 0 \leqq x \leqq m,$$

and

$$\Pr\{x \text{ or more among } m \text{ future trials will exceed the largest among } n \text{ observations}\} = P(m+n, n, x, 0), \qquad \text{for } 0 \leqq x \leqq m.$$

Also,

$$\Pr\{x \text{ or fewer among } m \text{ future trials will exceed the smallest observation in a sample of } n\} = \sum_{y=0}^{x} \binom{y+n-1}{n-1} \Big/ \binom{m+n}{n} = \binom{n+x}{n} \Big/ \binom{m+n}{n}.$$

The summation of the binomial coefficients was accomplished by means of Equation (12.8) of [16]. Then,

$$\Pr\{x \text{ or fewer among } m \text{ future trials will exceed the smallest observation in a sample of } n\} = P(m+n, n, m-x, 0), \qquad \text{for } 0 \leqq x \leqq m.$$

In general,

$$\Pr\{x \text{ or more among } m \text{ future trials will exceed the } r\text{th-largest among } n \text{ observations}\} = \sum_{y=x}^{m} \binom{m+n-r-y}{n-r} \binom{y+r-1}{r-1} \Big/ \binom{m+n}{n} = P(m+n, n, x+r-1, r-1), \qquad \text{for } 1 \leqq r \leqq n.$$

(See Section 2.1 for a proof of the equivalence of these sums.)

If $r = 1$, the formula for the probability that x or more among m future trials will exceed the largest among n observations is obtained.

Also, if $r = n$,

$$\Pr\{x \text{ or more among } m \text{ future trials will exceed the smallest among } n \text{ observations}\} = P(m+n, n, x+n-1, n-1);$$

or