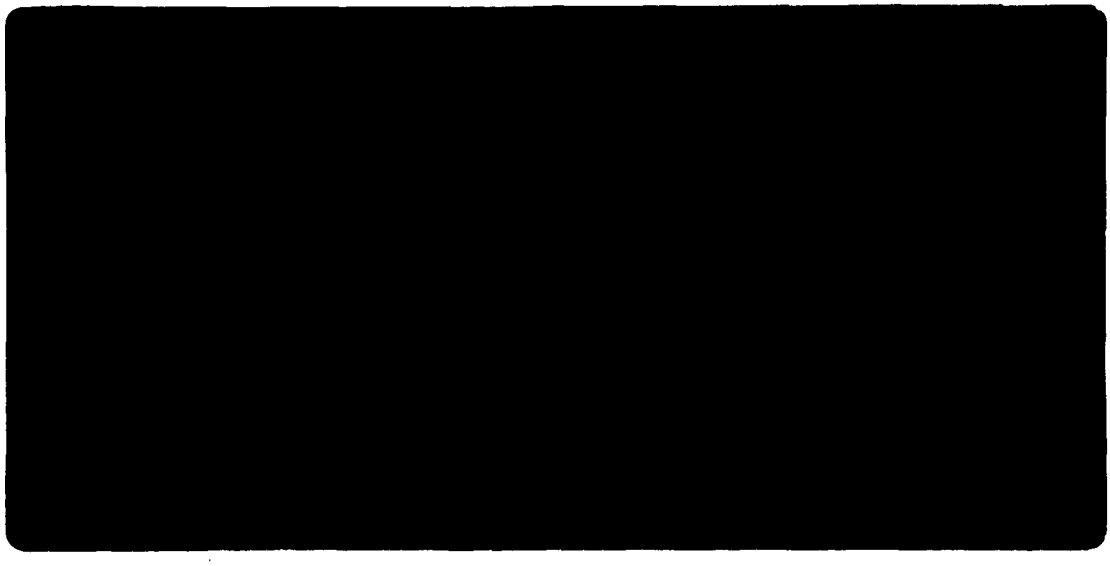
Institute of Paper Science and Technology
Atlanta, Georgia

IPST TECHNICAL PAPER SERIES



NUMBER 418

A NEW SOLUTION FOR THE PROBABILITY OF COMPLETING SETS IN RANDOM SAMPLING: DEFINITION OF THE "TWO-DIMENSIONAL FACTORIAL"

J.D. LINDSAY

JANUARY 1992

A New Solution for the Probability of Completing Sets in Random Sampling: Definition of the "Two-Dimensional Factorial"

J.D. Lindsay

To appear in
The Mathematical Scientist

INSTITUTE OF PAPER SCIENCE & TECHNOLOGY MEMORANDUM

FROM Jeff Lindsay DATE Jan. 23, 1992

TO IPST Members

SUBJECT Revision of a previous paper

The attached Technical Paper Series, "A New Solution for the Probability of Completing Sets in Random Sampling: Definition of the 'Two-Dimensional Factorial'," is the revised version of a similar paper you received several months ago (Series #400, Sept. 1991). The revisions are based upon an extremely helpful review of a still-unknown statistician who suggested a variety of improvements and performed some detailed computations. This article will appear in *The Mathematical Scientist*, a multidisciplinary journal published by the Applied Probability Group at the University of Sheffield.

The previous version of this paper included a section on nonuniform distributions. The reviewer showed that my conditional-probability solution, while correct, was excessively difficult to use and offered no clear advantage over an unconditional probability approach – both are sensitive to numerical resolution on a computer. A more elegant and computationally robust approach using the two-dimensional factorial may still be possible but has not yet been found. For now, I am simply deleting the discussion of nonuniform distributions. If you are interested in more information on this topic, please let me know.

The findings presented in this paper have application to the computation of probability in random sampling to complete a set or subset of uniformly distributed distinct items. Extensions using this work can be made to set completion for nonuniform distributions and to estimates of the number of missing species based on partial sampling. These extensions may be treated in future work.

I will contact the reviewer who spent so much time on this paper and pursue the possibility of collaboration in future work.

A NEW SOLUTION FOR THE PROBABILITY OF COMPLETING SETS IN RANDOM SAMPLING: DEFINITION OF THE "TWO-DIMENSIONAL FACTORIAL"

Jeffrey D. Lindsay
Institute of Paper Science and Technology
Atlanta, GA 30318

ABSTRACT

A new solution to a classical sampling problem is found. The problem concerns the probability of completing a subset of items when randomly sampling with replacement from a finite population (or equivalently of completing a subset of classes when sampling from an infinite population whose members are uniformly distributed among a finite collection of classes). In deriving the solution, an interesting recursive function is obtained which can be described as a "two-dimensional factorial." This function is partially tabulated, and several of its properties are investigated, including limits for large numbers. Use of this function offers significant computational advantages over the previous classical solution to the probability problem considered here. The function is not known to have been noted in previous work.

1. INTRODUCTION

In probability theory, a classical sampling problem concerns the likelihood of collecting a set of items by randomly sampling a population[1]. A simple example can be found in the collection of sets of promotional items offered inside cereal boxes. The items are presumably randomly and uniformly distributed and remain unidentified until the package has been opened. For instance, one cereal manufacturer offered miniature license plates from all 50 states with one plate per box. If somebody desires to collect all 50, how many boxes should one plan to purchase to be 95% confident that the set will be completed? A less ambitious consumer may simply want to know the probability that at least 10 different plates will be obtained by purchasing 12 boxes. Related problems may be found in sampling problems in scientific studies.

We will begin by considering the simple problem when the different classes in the population each compose an equal fraction of the population. In general terms, our problem statement becomes:

> *If U distinct classes of items are randomly and uniformly distributed among an infinite population, what is the probability that a specified number, U-M, of the classes will be acquired in N trials? (M is the number of missing classes in the sample.)*

We will introduce the notation P(N,U,M) to denote this probability. Feller[2] shows that this probability is

$$P(N,U,M) = \binom{U}{M} \sum_{k=0}^{U-M} (-1)^k \binom{U-M}{k} \left[1 - \frac{M+k}{U}\right]^N. \qquad (1)$$

By taking an independent approach in the solution, we will show a new form for the solution to be

$$P(N,U,M) = \frac{U!}{M!\ U^N} F(D, U-M), \tag{2}$$

where D is the number of duplicate items among the N samples, or D = N-(U-M), and F is a recursive function defined by

$$F(D,U-M) \equiv \sum_{j=1}^{U-M} j\ F(D-1,j), \tag{3}$$

$$F(0,j) = 1 \text{ for all } j = 1,2,3,....$$

After derivation of the probability formulas, we will discuss properties and limiting values of the recursive function F, which can be described as a two-dimensional factorial function.

2. DERIVATION

The two-dimensional factorial function was found by noting obvious patterns while determining the permutations for obtaining U-M distinct items in N trials. That number, divided by the total number of possible permutations, U^N, gives the desired probability. For example, consider the problem of collecting all three items of a set in six tries. Here U = 3, M = 0, and N = 6, and the number of duplicates, D, is 3. The permutations are treated in the following table. There are 10 cases to consider, one for each feasible combination of positions occupied by the duplicates. Duplicates are shown in bold, italic text. For example, in case 6, duplicates occur at trials 2, 5, and 6. For trial 1, any of the three distinct items can be chosen. If a duplicate occurs in trial 2, there is only one possibility, the same item that was selected in the first trial. The remaining two items appear in trials 3 and

4, so the number of possibilities becomes 2 and 1, respectively. For trials 5 and 6, any selection will be a duplicate, so the number of possibilities becomes 3 and 3.

Case	Trials						Permutations
1.	3	***1***	***1***	***1***	2	1	=3! * (1*1*1)
2.	3	***1***	***1***	2	***2***	1	=3! * (1*1*2)
3.	3	***1***	***1***	2	1	**3**	=3! * (1*1*3)
4.	3	***1***	2	***2***	***2***	1	=3! * (1*2*2)
5.	3	***1***	2	***2***	1	***3***	=3! * (1*2*3)
6.	3	***1***	2	1	***3***	***3***	=3! * (1*3*3)
7.	3	2	***2***	***2***	***2***	1	=3! * (2*2*2)
8.	3	2	***2***	***2***	1	***3***	=3! * (2*2*3)
9.	3	2	***2***	1	***3***	***3***	=3! * (2*3*3)
10.	3	2	1	***3***	***3***	***3***	=3! * (3*3*3)
						Total:	=3!* 90 =3!*F(3,3)

Table 1. Permutations for the 10 possible cases when U=3, N = 6, and M = 0.

The total number of permutations is the product of 3! and the total permutations for duplicates, which is the sum of the products in parentheses in the rightmost column of Table 1. The sum of numbers in parentheses can be written as either

$$F(3,3) = \sum_{a_3=1}^{3} \prod a_1\, a_2\, a_3 \text{ , with } a_3 \geq a_2 \geq a_1 \text{ , } a_i \in \{1,2,3\} \qquad (4)$$

or as

$$3*(1*1 + 1*2 + 2*2 + 1*3 + 2*3 + 3*3) + 2*(1*1+1*2+2*2) + 1*(1*1)$$

$$= F(3,3) = \sum_{j\,=\,1}^{3} j\, F(2,j) \text{ ,} \qquad (5)$$

$$\text{where } F(2,j) = \sum_{a_2=1}^{j} \prod a_1\, a_2 \text{ , with } a_2 \geq a_1 \text{ , } a_i \in \{1,2 \ldots j\}. \qquad (6)$$

The number of cases is given by the number of feasible combinations of positions for the D duplicates among N=U-M+D sample positions, with duplicates able to occur only after at least one element of U has been selected. The number of cases is thus (N-1)!/(D! [N-D-1]!). The number of choices available for a duplicate equals the number of unique items previously selected in that case.

In general, when all U distinct items are selected in an arbitrary N≥U trials, there are U! permutations for the first selections of these items. For the D = N-U duplicates, the kth duplicate can be any one of a_k items, where a_k denotes the number of distinct items already selected, $1 \leq a_1 \leq a_2 \leq \ldots. \leq a_D \leq U$. The number of permutations of duplicates is then $\prod a_1\, a_2 \ldots. a_D$. Summing over all possible values of a_D, the number of permutations for obtaining all U distinct items in N trials, resulting in D = N-U duplicates, is therefore

$$U! \sum_{a_D=1}^{U} \prod a_1 a_2 \ldots a_D \text{ , with } a_D \geq a_{D-1} \geq a_{D-2} \geq \ldots \geq a_1,\ a_i \in \{1,2 \ldots U\} \qquad (7)$$

which can also be written as U! F(D,U), where

$$F(D,U) \equiv \sum_{j=1}^{U} j\, F(D-1,j), \text{ with } F(0,j) = 1 \text{ for all } j = 1,2,3,\ldots. \qquad (8)$$

When M of the U unique items are missing in the sampled subset, the number of duplicates becomes D = N-(U-M). By considering the permutations of duplicates and first occurrences, as was done in deriving Equation (7), it is easily shown that the total number of permutations becomes

$$\frac{U!}{M!} F(D,U-M) \qquad (9)$$

with the function F defined in Equation (8). In general, then, the probability of obtaining U-M unique items from a possible U items, distributed uniformly throughout an infinite population, in N trials is

$$P(N,U,M) = \frac{U!}{M!\ U^N} F(D, U-M), \tag{10}$$

where D is the number of duplicate items among the N samples, D = N-(U-M), and F is a recursive function defined by

$$F(D,U-M) \equiv \sum_{j=1}^{U-M} j\, F(D-1,j), \quad \text{with } F(0,j) = 1 \text{ for all } j = 1,2,3,.... \tag{11}$$

Equating the r.h.s. of Equations (1) and (10) and simplifying yields

$$F(D,U-M) = \sum_{j=1}^{U-M} j\, F(D-1,j) = \sum_{j=0}^{U-M} \frac{(-1)^j\,(U-M-j)^N}{j!\ (U-M-j)!} \tag{12}$$

The identity in Equation (12) is by no means obvious and is an interesting result of itself.

The probability P(N,U,M) can be computed using either Equation (10) or Equation (1) from Feller. Likewise, F(D,U-M) can be determined using the recursive approach of Equation (11) or the alternating-sign series in Equation (12). Use of the recursive function offers a significant computational advantage for it is a summation of positive terms only, whereas the alternating-sign series involves small differences of large numbers. Limited numerical resolution on a computer thus greatly restricts the usefulness of Equation (1). For example, to compute F(D= 4, U = 43, M = 0) = 8.04E+11 with the alternating-sign series, differences between numbers 16 orders or magnitude greater are required. From j=6 to 11, the terms of the series are 1.45E+27, -2.04E+27, 2.35E+27, -2.25E+27, 1.80E+27, and -1.22E+27. Summing the series on a computer with 15 digits of resolution (the

Wingz™ spreadsheet by Informix was used on a Macintosh II) yielded a negative result, whereas accuracy was maintained with the recursive approach until sums exceeded the largest allowed number, 1.7E+308.

3. FURTHER PROPERTIES OF THE TWO-DIMENSIONAL FACTORIAL

The two-dimensional factorial appears to be an interesting function meriting further study. Table 2 shows values of F(D,U-M) for $1 \le D \le 25$ and $1 \le U\text{-}M \le 7$. Several interesting features are apparent in the columns of numbers shown here. Note that F(D,1) = 1 and $F(D,2) = 2^{D+1} - 1$ for all D. A logarithmic contour plot in Figure 1 for the range $1 \le D \le 30$ and $1 \le U\text{-}M \le 29$ shows how the numbers increase with U and D.

D \ U-M	1	2	3	4	5	6	7
1	1	3	6	10	15	21	28
2	1	7	25	65	140	266	462
3	1	15	90	350	1050	2646	5880
4	1	31	301	1701	6951	22827	63987
5	1	63	966	7770	42525	179487	627396
6	1	127	3025	34105	246730	1323652	5715424
7	1	255	9330	145750	1379400	9321312	49329280
8	1	511	28501	611501	7508501	63436373	408741333
9	1	1023	86526	2532530	40075035	420693273	3281882604
10	1	2047	261625	10391745	210766920	2734926558	25708104786
11	1	4095	788970	42355950	1096190550	17505749898	1.97E+11
12	1	8191	2375101	171798901	5652751651	110687251039	1.49E+12
13	1	16383	7141686	694337290	28958095545	6.93E+11	1.11E+13
14	1	32767	21457825	2798806985	147589284710	4.31E+12	8.23E+13
15	1	65535	64439010	11259666950	7.49E+11	2.66E+13	6.03E+14
16	1	131071	193448101	45232115901	3.79E+12	1.63E+14	4.38E+15
17	1	262143	580606446	1.82E+11	1.91E+13	9.99E+14	3.17E+16
18	1	524287	1742343625	7.28E+11	9.64E+13	6.09E+15	2.28E+17
19	1	1048575	5228079450	2.92E+12	4.85E+14	3.70E+16	1.63E+18
20	1	2097151	15686335501	1.17E+13	2.44E+15	2.25E+17	1.16E+19
21	1	4194303	47063200806	4.68E+13	1.22E+16	1.36E+18	8.29E+19
22	1	8388607	1.41E+11	1.87E+14	6.13E+16	8.22E+18	5.88E+20
23	1	16777215	4.24E+11	7.49E+14	3.07E+17	4.96E+19	4.17E+21
24	1	33554431	1.27E+12	3.00E+15	1.54E+18	2.99E+20	2.95E+22
25	1	67108863	3.81E+12	1.20E+16	7.71E+18	1.80E+21	2.08E+23

Table 2. F(D,U-M) for $1 \le D \le 25$ and $1 \le U\text{-}M \le 7$.

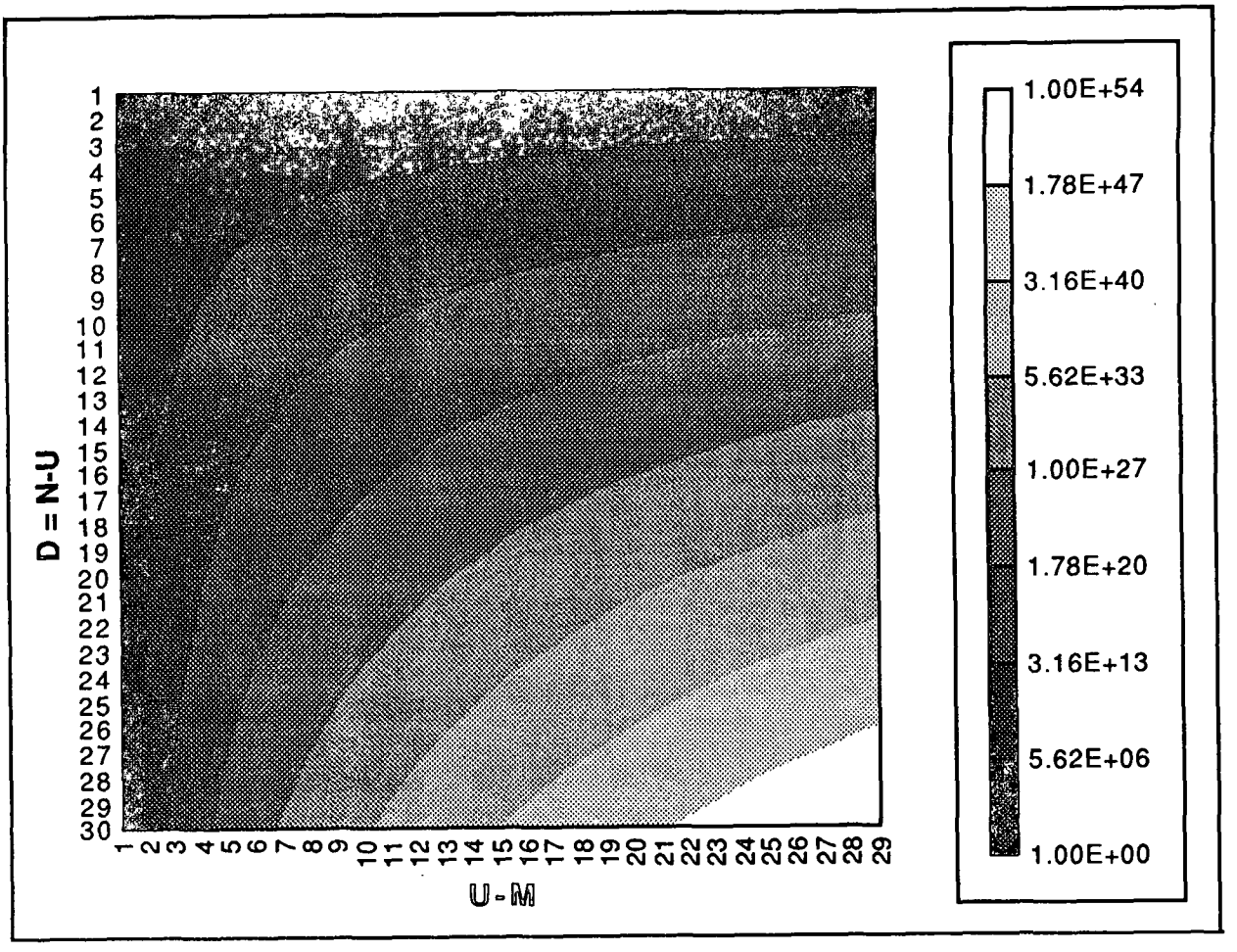


Figure 1. Logarithmic contour plot of F(D,U-M) for $1 \le D \le 30$ and $1 \le U-M \le 29$.

3.1 Limits for Large Numbers

As D, and hence N, become very large for a given U, P(N,U,0) approaches unity (it becomes nearly certain that all U items will be collected if enough samples are obtained). Thus,

$$\lim_{D \to \infty} F(D,U) = \frac{U^N}{U!}. \tag{13}$$

Therefore, the ratio F(D,U)/F(D-1,U) approaches U for large D. Likewise, for large D, the ratio of adjacent values in any row is

$$\lim_{D\to\infty} \frac{F(D,U)}{F(D,U-1)} = \left(\frac{U}{U-1}\right)^{N-1} \tag{14}$$

A more exact expression than Equation (13) is possible using a theorem for the limit of Equation (1) proved by Feller[3] and attributed (with a different proof) to von Mises[4]:

If U and N increase so that $\lambda = Ue^{-N/U}$ *remains bounded, then for fixed M:*

$$P(N,U,M) \to \frac{\lambda^M}{M!} e^{-\lambda}, \tag{15}$$

which is the Poisson distribution. The two-dimensional factorial for large N = U+D is then

$$F(D,U-M) = F(N-U+M,U-M) \to \frac{U^{(N+M)}}{U!} \exp - \left[\frac{NM}{U} + U \exp\left(\frac{-N}{U}\right)\right]. \tag{16}$$

For M = 0, this can be rewritten as

$$F(N-U,U) \to \frac{\left(U^U\right)^{N/U} \left(e^{-U}\right)^{e^{-N/U}}}{U!}, \tag{17}$$

or for finite M, we can re-express Equation (16) as

$$F(D,U-M) = F(N-U+M,U-M) \to \frac{\left(M^M\right)^{\ln(M/U)} \left(e^{-M}\right)^{N/U} \left(U^U\right)^{N/U} \left(e^{-U}\right)^{e^{-N/U}}}{U!}, \tag{18}$$

where the terms in the numerator bear some resemblance to Stirling's formula for large factorials,

$$n! \to \sqrt{2\pi n}\, n^n\, e^{-n}. \tag{19}$$

While the resemblance to the regular factorial function is somewhat superficial, the two-dimensional factorial is still suggested as an appropriate name for the

recursive F function introduced here. The main similarity to the factorial is through the recursive expression given in Equation (11).

Comparisons of the approximate form in Equation (15) with the exact probability form in Equation (10) suggest that the approximate form must be used with caution for M>0. For example, for a given U and M, the approximation may be close for a certain range of N, but will become increasingly incorrect as N increases.

3.2 Number Analysis

One feature of the numbers produced by the two-dimensional factorial is that a large proportion of them seem to have seven and eleven as factors. In Table 3 (a subset of Table 2), numbers divisible by seven are in italics, and numbers divisible by eleven are in boldface. About 20% of the numbers examined are divisible by both seven and eleven. I have no explanation for this feature.

D \ U-M	1	2	3	4	5	6	7
1	1	3	6	10	15	*21*	*28*
2	1	*7*	25	65	*140*	*266*	***462***
3	1	15	90	*350*	*1050*	*2646*	*5880*
4	1	31	*301*	*1701*	*6951*	*22827*	***63987***
5	1	*63*	*966*	*7770*	*42525*	***179487***	***627396***
6	1	127	**3025**	34105	**246730**	**1323652**	**5715424**
7	1	255	9330	**145750**	**1379400**	***9321312***	***49329280***
8	1	*511*	**28501**	**611501**	***7508501***	***63436373***	***408741333***
9	1	**1023**	**86526**	***2532530***	***40075035***	***420693273***	***3281882604***

Table 3. Subset of Table 2 showing F(D,U-M) values divisible by seven in italic and values divisible by eleven in boldface.

Examination of the last digits of the numbers in columns 2 through 5 shows interesting repeating patterns if we consider that the initial, undisplayed row for D = 0 consists of ones. The repeating final digits are:

Column 2: 1-3-7-5

Column 3: 1-6-5-0

Column 4: 1-0-5-0

Column 5: 1-5-0-0

Column 6 shows an interesting pattern in the final digits. The sequence is 1-1 – 6-6 – 7-7 – 2-2 – 3-3 – 8-8 – 9-9 – 4-4 – 5-5 – 0-0, which apparently repeats (I am not sure because of limited numerical resolution). These pairs of digits change according to a specific pattern: add 5, add 1, subtract 5, add 1, and repeat.

4. APPLICATIONS AND EXAMPLES

4.1 Probability of Collecting *at Least* U-M Sets

P(N,U,M) in Equation (10) gives the probability of obtaining exactly U-M distinct items in a random sample of size N from a uniform, infinite population. The collector, however, is usually more interested in the probability of collecting at least a specified number of distinct items. For varying M with constant N and U, each P(N,U,M) is independent. Therefore, the probability that no more than M_{max} items are missing in a random sample of size N is given by

$$P(N,U,M\le M_{max}) = \frac{U!}{U^N}\sum_{M=0}^{M_{max}} \frac{F(N-U+M,U-M)}{M!}. \qquad (20)$$

Since the probability of having at least one distinct item is unity,

$$\frac{U!}{U^N}\sum_{M=0}^{N-U} \frac{F(N-U+M,U-M)}{M!} = 1. \qquad (21)$$

4.2 Expected Number of Trials to Complete a Set

For a series of U distinct items sampled with replacement, Feller[5] shows that the expected number of samples to obtain U-M distinct items is

$$E(N_{U-M}) = U\left\{\frac{1}{U} + \frac{1}{U-1} + \frac{1}{U-2} + \ldots + \frac{1}{M+1}\right\}, \tag{22}$$

which, for large U, can be approximated by

$$E(N_{U-M}) \approx U \ln\left(\frac{U}{M+1}\right). \tag{23}$$

In the limit of large U, E(N) = U ln(U) when M = 0. Applying the Poisson approximation to P(N,U-0) for large N, we see that the probability of collecting all U sets in E(N) trials approaches $e^{-1} \approx 0.3679$ as U becomes large.

4.3 Sample Probability Results

Table 4 shows the smallest number of samples required to complete a set of U distinct items (M=0) with minimum p-values from 0.5 to 0.99. Equation (10) was used for all values of U; for comparison, several results from the Poisson approximation in Equation (15) for U = 50 are also shown in the last four rows. Based on Table 4, the would-be collector of items hidden in packages should plan on buying three to five times as many packages as there are items to be collected to be fairly sure (ca. 90% confident) of collecting a complete set with less than 20 items. For larger sets (say > 25 items), it may be necessary to buy six or more times as many packages as there are items to be collected.

	Smallest N required for a p-value of at least				
U	0.5	0.75	0.9	0.95	0.99
2	2	3	5	6	8
3	5	7	9	11	15
4	7	10	13	16	21
5	10	14	18	21	28
6	13	18	23	27	36
7	17	22	28	33	43
8	20	26	33	38	51
9	23	30	38	44	58
10	28	35	44	51	66
11	31	39	50	57	74
12	35	44	55	63	82
...	.....	.....	.....	.....	.....
25	90	111	135	152	192
...	.....	.....	.....	.....	.....
50	214	257	306	341	422
	Estimates using Equation (15):				
5	*10*	*15*	*20*	*23*	*32*
12	*35*	*45*	*57*	*66*	*86*
25	*90*	*112*	*137*	*155*	*196*
50	*214*	*258*	*308*	*345*	*426*

Table 4. Samples required to complete sets of distinct items at several p-values.

For the case of M = 0, the Poisson approximation corresponds well to the exact results of Equation (10). We noted above that the accuracy decreases substantially when M > 0. For example, the probability of getting exactly 48 out of 50 distinct items (M = 2) from 159 samples is 0.297 from Equation (10), but the Poisson approximation gives 0.270. However, cases with M > 0 are often of less interest than completed sets.

In computing results for U = 25 and 50 in Table 4, rescaling of F values was necessary to avoid numeric overflow. To rescale on a spreadsheet with columns of U and rows of D, divide a row of F values at constant D by a large number such as 10^{250} (the choice depends on the numerical limits of the computer and software and the value of U being considered). F values in subsequent rows (higher D) are then computed recursively with Equation (8). To obtain P(N,U,M) from Equation (10) using the reduced F values, replace U^N in the denominator with the appropriately scaled number, e.g., $10^{(N \log_{10}(U) - 250)}$, which may prevent the numeric overflow that can occur in evaluating U^N. In rescaling, numeric underflow could occur in columns of small U, but these have a negligible effect on F at the larger U values where rescaling is needed.

For the originally considered case of U = 50 license plates, Table 5 shows the probabilities of obtaining partially completed sets with various M values if one buys N = 100 boxes. The likelihood of collecting plates from all 50 states is 0.00017 and the chance that no more than three states will be missing is only 5.18%. The most likely outcome is that six states will be missing, although there is a 52% probability that even more than six will be missing. With N = 180, the probability of completing the set is still only 24.5% (25.5% according to the approximation of Equation [15]). To be 90% confident of getting all 50 states, 306 boxes must be purchased, as shown in Table 4. (Consumers may do well to simply contact the manufacturer of the collectable items and buy a complete set directly.)

M	P(100,50,M)	Cumulative
0	0.00017	0.00017
1	0.00202	0.00219
2	0.01129	0.01348
3	0.03835	0.05183
4	0.08910	0.14093
5	0.15071	0.29164
6	0.19294	0.48458
7	0.19183	0.67641
8	0.15082	0.82723
9	0.09501	0.92224
10	0.04841	0.97065
11	0.02009	0.99074
12	0.00682	0.99756
13	0.00190	0.99947
14	0.00044	0.99990
15	0.00008	0.99999

Table 5. Probabilities for the case of U = 50 and N =100.

5. CLOSURE

A new form of the solution to a classical probability problem has yielded an interesting function which may be termed a two-dimensional factorial. The function allows computation of set collection probabilities with improved accuracy compared to the classical alternating-sign series solution in Equation (1) for uniformly distributed populations.

ACKNOWLEDGMENT

The author is grateful for the valuable comments and suggestions offered by Dr. Bruce Collings of the Brigham Young University Statistics Department and by Kendra L. Lindsay.

REFERENCES

1. Feller, W. (1950), *An Introduction to Probability Theory and Its Applications*, Vol. 1, New York: John Wiley and Sons, pp. 51-66.

2. Feller, p. 69, see also p. 64.

3. Feller, pp. 72-75.

4. von Mises, R. (1939), "Über Aufteilungs- und Besetzungswarhscheinlichkeiten," *Revue de la Faculté des Sciences de l'Université d'Istanbul*, N.S., Vol. 4 , pp. 1-19, as cited by Feller, p. 72.

5. Feller, pp. 174-175.