The Representation of Hypergeometric Random Variables using Independent Bernoulli Random Variables

Stefen Hui* and C. J. Park†
Mathematics & Statistics
San Diego State University
San Diego, CA 92182

June 2, 2011

Running Head: The Hypergeometric and Bernoulli Random variables

Abstract

In this paper we show that a hypergeometric random variable can be represented as a sum of independent Bernoulli random variables that are, except in degenerate cases, not identically distributed. In the proof we use the factorial moment generating function. An asymptotic result on the probabilities of the Bernoulli random variables in the sum is also presented. Numerical examples are used illustrated the results.

*Corresponding author, hui@sciences.sdsu.edu, Department of Mathematics, San Deiego State University, San Diego, CA, 92182, USA
†chongjin.park@mail.sdsu.edu

1 Introduction

It is well known that a binomial random variable is the sum of independent identically distributed Bernoulli random variables, and conversely. It is also well known that a hypergeometric random variable is the sum of *dependent* identically distributed Bernoulli random variables (see for example, Feller [2], Ross [4], or Wilks [5].) In this paper we show that a hypergeometric random variable is also the sum of *independent* Bernoulli random variables; see Harris and Park [3] for a similar result for a different probability model.

To motivate our approach, we use the following standard interpretation of the hypergeometric distribution: Consider a list containing N binary digits, with b ones and r zeros. Select n numbers from the list at random one at a time without replacement, Let S_n be the sum of the numbers selected. Then the random variable S_n follows a hypergeometric distribution. We show that the random variables S_n has the same distribution as a sum of independent Bernoulli random variables that are not necessarily identically distributed.

Theorem 1 *Let S_n be random variable that follows a hypergeometric distribution with parameters N, b, n. Then there exist* $\min(b, n)$ *independent Bernoulli random variables such that their sum has the same probability distribution as S_n.*

Theorem 1 follows from the following theorem. In the remainder of this paper, we let $\widetilde{n} = \min(b, n)$.

Theorem 2 *The factorial moment generating function* $g_n(t) = E(1+t)^{S_n}$ *of* S_n *can be factored into*

$$g_n(t) = \prod_{j=1}^{\widetilde{n}} (1 + p_j t), \tag{1}$$

where $0 < p_j \leq 1$ *for* $j = 1, \ldots, \widetilde{n}$.

The proof the theorems will explicitly give $P(Y = 1)$ for each Bernoulli random variable Y in terms of the zeros of polynomials, which can be evaluated numerically.

A preliminary technical result required for the proofs of the main theorems is given in Section 2 and the proofs of Theorems 1 and 2 are given in Section 3. Examples, along with an asymptotic result on the p_j's, are given in Section 4. Section 5 contains the conclusions.

2 A Technical Result

In this section, we prove a technical result on the zeros of certain polynomials required in the proof of the main result.

Let N be a positive integer and let p be a polynomial with degree B, $0 < B \leq N$, and $p(0) = 1$. Let $f_N = p$ and, for $n = 1, \ldots, N$, iteratively define

$$f_{n-1}(t) = f_n(t) - \frac{t}{n} f_n'(t),$$

where f_n' denotes the derivate of f_n. The following facts follow immediately from the definitions of f_n:

1. $f_0(0) = \cdots = f_N(0) = 1$

2. If $\deg f_n = n$, then $\deg f_{n-1} = n - 1$.

3.

$$f_{n-1}(t) = f_n(t) - \frac{t}{n} f'_n(t) = -\frac{t^{n+1}}{n} \frac{d}{dt} \left[\frac{f_n(t)}{t^n} \right]$$

We next show that all zeros of f_n are all real and in the interval $(-\infty, -1]$ if all the zeros of f_N are real and in the interval $(-\infty, -1]$.

Theorem 3 *If p has B real zeros in $(-\infty, -1]$, then*

$$\deg f_n = \min\{n, B\}$$

and the zeros of each f_n are real and in $(-\infty, -1]$.

Proof By assumption, f_N satisfies the stated claims. Suppose $B < N$. The the function $q_N(t) = f_N(t)/t^N$ is rational with a pole at $t = 0$ and has B zeros in $(-\infty, -1]$ and a zero at $-\infty$. It follows that q_N has B critical points in $(-\infty, -1]$. Since f_{N-1} is a polynomial with degree at most B and has B zeros in $(-\infty, -1]$, it is a polynomial with degree B with all real zeros (and in $(-\infty, -1]$.) This argument is repeated for $n = B + 1, \ldots, N - 1$. For $n = 0, \ldots, B$, we have by Lemma 2 that $\deg f_n = n = \min\{n, B\}$. Also, the function $q_n(t) = f_n(t)/t^n$ has n zeros in $(-\infty, -1]$ but does not have a zero at $-\infty$. Therefore q_n only has $n - 1$ critical points in $(-\infty, -1]$ and so f_{n-1} has all its zeros in $(-\infty, -1]$.

If $B = N$, then by Lemma 2,

$$\deg f_n = n = \min\{n, N\},$$

for $n = 0, \ldots, N$, and the proof as above for the case of no zeros at $-\infty$ shows that all the zeros of f_n are real and in $(-\infty, -1]$. The proof of the theorem is complete.

Remark 1 *Note the statement concerning the critical points does require the zeros of the f_n's be distinct. It holds even if the zeros are repeated as in $p(t) = (1+t)^B$. One can explicitly compute f_n for $B = N$ and for $B = 1$ but things get tedious rapidly for the other B's.*

3 Proofs of Theorems 1 & 2

Let $g_n(t) = E(1+t)^{S_n}$ be the factorial moment generating function of S_n. We obtain the coefficients of g_n by equating the known form of the factorial moments (see, for example, Wilks [5], p. 135) and the derivatives of g_n at $t = 0$ to give

$$g_n(t) = \sum_{k=0}^{\tilde{n}} \frac{\begin{pmatrix} b \\ k \end{pmatrix}\begin{pmatrix} n \\ k \end{pmatrix}}{\begin{pmatrix} N \\ k \end{pmatrix}} t^k, \tag{2}$$

where $\widetilde{n} = \min(b, n)$, $N = b + r$, and $n \leq N$. A straightforward calculation using Equation 2 shows that for $n = 1, \ldots, N$,

$$g_{n-1}(t) = g_n(t) - \frac{t}{n} g_n'(t).$$

Also, since $\widetilde{N} = \min(b, N) = b$,

$$g_N(t) = \sum_{k=0}^{b} \left(\begin{array}{c} b \\ k \end{array} \right) t^k = (1+t)^b. \tag{3}$$

We can now apply Theorem 3 to conclude that the zeros of g_n are real and are in the interval $(-\infty, -1]$. Therefore g_n can be written as

$$g_n(t) = \prod_{j=1}^{\widetilde{n}} (1 + p_j t), \tag{4}$$

where $0 < p_j \leq 1$. This completes the proof of Theorem 2.

For Theorem 1, let Y_j, $j = 1, \ldots, \widetilde{n}$, be a sequence of independent Bernoulli random variables with $P(Y_j = 1) = p_j$. Then

$$\begin{aligned} E[(1+t)^{Y_1 + \cdots + Y_{\widetilde{n}}}] &= E(1+t)^{Y_1} \times \cdots \times E(1+t)^{Y_{\widetilde{n}}} \\ &= \prod_{j=1}^{\widetilde{n}} (1 + p_j t) \\ &= g_n(t). \end{aligned}$$

Thus S_n and $Y_1 + \cdots + Y_n$ have the same factorial moment generating function and are therefore equal in distribution. This completes the proof of Theorem 1

4 Examples and Further Results

In this section, we illustrate our results with numerical examples and and present a result related to the distribution of the probabilities p_j as the population size N gets large.

The first example uses small parameters and the calculation can be carried out by hand. The second example uses larger parameters and Matlab is used to compute the derivative and find the zeros of g_n. A small simulation is also given to compare the actual values of the hypergeometric distribution to the values obtained by summing the Bernoulli random variables.

Example 1 *Consider an urn that contains 5 marbles, 3 black and 2 red marbles. Select two marbles one at a time without replacement and let S_2 denote the number of black marbles selected in two draws. The factorial moment generating function of S_2 is*

$$\begin{aligned} g_2(t) &= \frac{1}{10} + \frac{6}{10}(1+t) + \frac{3}{10}(1+t)^2 \\ &= 1 + \frac{12}{10}t + \frac{3}{10}t^2. \end{aligned}$$

The zeroes of g_2 are

$$r_1 = -2 + \frac{\sqrt{6}}{3} \quad \textit{and} \quad r_2 = -2 - \frac{\sqrt{6}}{3},$$

and thus

$$p_1 = -\frac{1}{r_1} = \frac{3}{5} + \frac{\sqrt{6}}{10} \quad \textit{and} \quad p_2 = -\frac{1}{r_2} = \frac{3}{5} - \frac{\sqrt{6}}{10}.$$

Simple calculations show that

$$p_1 p_2 = \frac{3}{10},\ p_1(1-p_2) + p_2(1-p_1) = \frac{6}{10},\ (1-p_1)(1-p_2) = \frac{1}{10},$$

as required. Thus $S_2 = Y_1 + Y_2$*, where* Y_1, Y_2 *are Bernoulli random variables with probabilities* p_1, p_2 *respectively.*

The frequency domain remains the same, as in an infinite population case which relate to Hardy-Weinberg Theorem. Hence Example 1 demonstrates that Hardy-Weinberg Theorem holds for a finite population with adjusted frequencies.

Example 2 *In this example, we use* $N = 50$*,* $b = 20$ *and* $n = 10$*. We use Matlab to find the zeros of* g_{10}*, which has degree* 10*, as given by Equation 2 and then the* p_j*'s, which are, to* 4 *decimal places,*

$$\{0.0947, 0.1511, 0.2123, 0.2784, 0.3491, 0.4234, 0.5006, 0.5802, 0.6620, 0.7483\}.$$

We generated 200,000 *sums of* 10 *independent Bernoulli random variables using these probabilities and calculated the empirical distribution of the sum for* S_{10}*.*

The actual and the distributions are summarized in the following table:

S_n	*Actual*	*Empirical*
0	0.0029	0.0028
1	0.0279	0.0286
2	0.1083	0.1074
3	0.2259	0.2251
4	0.2801	0.2811
5	0.2151	0.2148
6	0.1034	0.1037
7	0.0306	0.0306
8	0.0053	0.0053
9	0.0005	0.0004
10	0.0000	0.0000

We see that the values of the empirically calculated distribution closely match the actual values.

As illustrated by Example 2, it is easy to calculate the p_j's for moderate values of N, b, and n. It is interesting to note that the p_j's satisfy the following conditions:

$$\sum_{j=1}^{\tilde{n}} p_j = \tilde{n}\frac{b}{N}, \ \sum_{j=1}^{\tilde{n}} p_j(1-p_j) = n\frac{b}{N}\frac{(N-b)}{N}\frac{(N-n)}{(N-1)}.$$

Example 2 and other numerical experiments show that p_j's are fairly uniformly distributed and are not clustered in a tight neighborhood of b/N for moderate values of b and N. However, the p_j's do cluster around b/N for large b and N. More precisely, we have:

Theorem 4 *Let n be fixed positive integer and let p be a positive real number. Suppose $b/N = p$. Let $g_n^{(N)}$ denote the factorial moment generating function for S_n with population size N. Then*

1. *$g_n^{(N)}$ converges uniformly to $(1+pt)^n$ on compact subsets of the complex plane $\mathbb{C}$ and*

2. *for every $\epsilon > 0$, there is a positive integer N_0 such that for $N \geq N_0$, the coefficient p_j for each factor $1 + p_j t$ of $g_n^{(N)}$ satisfies $|p_j - p| < \epsilon$.*

Proof For N large, we have from Equation 2 that

$$g_n^{(N)}(t) = \sum_{k=0}^{n} \frac{\binom{b}{k}\binom{n}{k}}{\binom{N}{k}} t^k \tag{5}$$

$$= \sum_{k=0}^{n} \frac{(1-\frac{1}{b})\cdots(1-\frac{k-1}{b})}{(1-\frac{1}{N})\cdots(1-\frac{k-1}{N})} \binom{n}{k} (pt)^k \tag{6}$$

As $N \to \infty$, we have $b = pN \to \infty$ and

$$g_n^{(N)}(t) \to \sum_{k=0}^{n} \binom{n}{k} (pt)^k = (1+pt)^n,$$

pointwise and uniformly on compact subsets of $\mathbb{C}$. Let $\epsilon > 0$. Since $(1+pt)^n$ has a zero of order n at $-1/p$, Hurwitz's Theorem (see Duncan [1], p. 225) implies that there is a positive integer N_0 such that n of the zeros of $g_n^{(N)}$ are in the disk $\{z \in \mathbb{C} : |z + 1/p| < \epsilon\}$. As a polynomial of degree n, $g_n^{(N)}$ has exactly n zeros (which are all real by Theorem 1.) Therefore for all p_j for which $1 + p_j t$ is a factor of $g_n^{(N)}$, we have

$$\left| -\frac{1}{p_j} + \frac{1}{p} \right| < \epsilon$$

and thus

$$|p_j - p| < p_j p \epsilon \leq \epsilon,$$

since $0 < p, p_j \leq 1$. The proof is complete.

Theorem 4 is another confirmation that for $N \gg n$, sampling without replacement is almost the same as sampling with replacement. The difference is that with replacement, S_n is a sum of independent identically distributed Bernoulli random variables, and for sampling without replacement, S_n is a sum of independent distributed Bernoulli random variables that are almost identically distributed in the sense that the probabilities $p_j = P(Y_j = 1)$ are almost equal.

5 Conclusions

We have shown that the hypergeometric random variable can be represented as a sum of independent Bernoulli random variables, not identically distributed, whereas the Binomial random variable can be represented a sum of independent identically distributed Bernoulli random variables. We illustrated our result with numerical examples.

References

[1] Duncan, J., (1968), The Elements of Complex Analysis, John Wiley and Sons, New York.

[2] Feller, W., (1968), An Introduction to Probability Theory and Its Applications, Volume I, Third Edition, John Wiley and Sons, New York

[3] Harris, B. and Park, C. J., (1971), "A Note on the Asymptotic Normality of the Distribution of the Number of Empty Cells in Occupancy Problems," Ann. of the Institute of Statistical Mathematics, 23, pp. 507-513.

[4] Ross, Sheldon, (2002), Introduction to Probability Model, Academic Press.

[5] Wilks, S.S., (1962), Mathematical Statistics, John Wiley and Sons, New York.