# Representation of a Hypergeometric Random Variable as a Sum of Independent Bernoulli Variables

Jerry Alan Veeh

December 19, 2016

## Contents



## 1 Introduction

The objective of this note is to show that any hypergeometric random variable can be written as a sum of independent Bernoulli random variables. Writing a hypergeometric random variable as a sum of identically distributed, but dependent, Bernoulli variables is easy: the $i$th Bernoulli variable takes the value 1 exactly when the $i$th draw is a success. The representation here will be as a sum of independent but not identically distributed Bernoulli variables.

This work expands the details of the proof of this result given in M. P. Quine *Probability Approximations for Discrete Divisible Distributions* (Australian Journal of Statistics 36(3), 1994, 339–349), and similar results given by similar methods in Vatutin and Mikhailov *Limit Theorems for the Number of Empty Cells in*

*an Equiprobable Scheme for Group Allocation of Particles* (Theory of Probability and Its Applications 27(4), 1982, 734–743) and in L. H. Harper *Stirling Behavior is Asymptotically Normal* (Annals of Mathematical Statistics 38, 1967, 410–414).

# 2 A General Bernoulli Representation Theorem

Suppose $X$ is a bounded non-negative integer valued random variable with probability generating function $G(z) = E[z^X]$. Note that $G$ is a polynomial with non-negative coefficients and that $G(1) = 1$. Denote by $d$ the degree of the polynomial $G$. Since at least one coefficient of $G$ is strictly positive, $G(z) > 0$ for $z > 0$. Thus all real roots of $G$ are non-positive.

Suppose that $G$ has $d$ real roots, $r_1 \leq \ldots \leq r_d \leq 0$. Let $B_1, \ldots, B_d$ be independent Bernoulli random variables defined on some probability space with the success probability for $B_i$ being $0 < 1/(1-r_i) \leq 1$. Denote by $B$ the sum of these Bernoulli variables. Then the generating function of $B$ is

$$E\left[z^B\right] = \prod_{i=1}^{d} E\left[z^{B_i}\right] = \prod_{i=1}^{d} \left(\frac{z - r_i}{1 - r_i}\right)$$

by independence of the $B$'s and direct computation. Now the generating functions of $X$ and $B$ are two polynomials of degree $d$ with the same roots which also take the value 1 at $z = 1$ since they are generating functions. Thus the generating functions of $X$ and $B$ are equal for all $z$, which means that $X$ and $B$ have the same distribution.

# 3 Hypergeometric Random Variables

Recall the commonly used model for a hypergeometric random variable. An urn contains $s$ balls labelled *success* and $f$ balls labelled *failure*, and $n$ balls are drawn from the urn at random and without replacement. The number, $H_n$, of success balls in the sample is a hypergeometric random variable. Here $0 \leq n \leq s + f$, and the extreme values of $n$ yield a degenerate random variable.

Notice too that $H_n$ takes each of the integer values from $\max\{0, n - f\}$ to $\min\{n, s\}$ with positive probability. Thus the generating function $G_n(z)$ of $H_n$ is a polynomial of degree $\min\{n, s\}$ which has a root at $z = 0$ of multiplicity $\max\{0, n - f\}$.

The first objective is to study the connection between the generating function of $H_{n+1}$ and $H_n$. To this end, observe that conditionally given $H_n$, $H_{n+1}$ takes the value $1+H_n$ with probability $(s-H_n)/(s+f-n)$ and the value $H_n$ with probability $1-(s-H_n)/(s+f-n) = (H_n+f-n)/(s+f-n)$. With this in mind, for $0 \le n < s+f$

$$\begin{aligned} E[z^{H_{n+1}}|H_n] &= z^{1+H_n}\left(\frac{s-H_n}{s+f-n}\right) + z^{H_n}\left(\frac{H_n+f-n}{s+f-n}\right) \\ &= z^{H_n}\left(\frac{(1-z)H_n}{s+f-n}\right) + z^{H_n}\left(\frac{zs+f-n}{s+f-n}\right). \end{aligned}$$

Now $E[H_n z^{H_n}] = z\frac{d}{dz}E[z^{H_n}]$. Using this and taking expectations yields

$$G_{n+1}(z) = G_n'(z)\left(\frac{z(1-z)}{s+f-n}\right) + G_n(z)\left(\frac{zs+f-n}{s+f-n}\right).$$

Since $H_n$ has a degenerate distribution if either $s=0$ or $f=0$, the next part of the discussion assumes that $s \ge 1$ and $f \ge 1$. The objective is to prove under this assumption that for $1 \le n < s+f$ the generating function $G_n$ has $R(n) = \min\{n,s\} - \max\{0, n-f\}$ simple negative real roots and a root at $z=0$ of multiplicity $\max\{0,n-f\}$. Since the degree of $G_n$ is $\min\{n,s\}$, this will imply that all of the roots of $G_n$ are real and non-positive. Note that $R(n) \ge 1$ under these assumptions. The proof will proceed by induction on $n$.

Since $G_0(z) = 1$, the above recursion gives $G_1(z) = (zs+f)/(s+f)$. Thus $G_1$ has a simple negative real root, as claimed. Suppose now that $G_n$ has real roots as claimed. The assertion about the roots at $z=0$ follows directly from the definition of generating function and the fact that $H_{n+1}$ takes its minimum value $\max\{0, n+1-f\}$ with positive probability. The result will be proved if the assertion about strictly negative roots is established.

First note a general fact. Suppose $r$ is a negative root of $G_n$. The recursion shows that $G_{n+1}(r)$ and $G_n'(r)$ have opposite signs. Since the negative roots of $G_n$ are simple and $G_n$ is a polynomial, $G_n'$ alternates signs at these roots. So $G_{n+1}$ alternates signs at these roots too. Thus $G_{n+1}$ has $R(n)-1$ roots entwined between the negative roots of $G_n$.

If $n < \min\{s,f\}$, then $n+1 \le \min\{s,f\}$ and $R(n) = n$ and $R(n+1) = n+1$. The assertion then is that $G_{n+1}$ has one more simple negative root than $G_n$. For these values of $n$ and $n+1$, $G_n(0) > 0$ and $G_{n+1}(0) > 0$. So at the largest negative root of $G_n$, $G_n'$ must be positive. Thus $G_{n+1}$ is negative at this root, which implies

that $G_{n+1}$ has a root between the largest negative root of $G_n$ and 0. Also for these values of $n$ and $n+1$, $G_n$ has degree $n$ while $G_{n+1}$ has degree $n+1$. Thus at the smallest negative root of $G_n$, $G_{n+1}$ will have the opposite sign of its limit as $z \to -\infty$. So $G_{n+1}$ has a root smaller than the smallest root of $G_n$. This shows that $G_{n+1}$ has the desired number of simple roots in this case.

Suppose now that $\min\{s,f\} \le n < \max\{s,f\}$. Then $R(n) = R(n+1)$, and the assertion is that $G_n$ and $G_{n+1}$ have the same number of negative simple roots. As noted above, $G_{n+1}$ has $R(n)-1$ roots entwined with the negative roots of $G_n$. If $s < f$, then $G_n(0) > 0$ and $G_{n+1}(0) > 0$ and the same argument as above shows that $G_{n+1}$ has a root between the largest negative root of $G_n$ and 0. If $s \ge f$, then $G_{n+1}$ is of degree $n+1$ while $G_n$ has degree $n$. So as before, $G_{n+1}$ has a root smaller than the smallest root of $G_n$. Thus in either case, $G_{n+1}$ and $G_n$ have the same number of simple negative roots.

Finally, suppose $n \ge \max\{s,f\}$. Then $R(n) = s+f-n = 1+R(n+1)$. As before, $G_{n+1}$ has $R(n)-1$ roots entwined between the roots of $G_n$, as desired.

# 4 The Hypergeometric Representation

Suppose $s \ge 0$, $f \ge 0$ and $0 < n \le s+f$. The assertion is that $H_n$ can be written as the sum of $n$ independent Bernoulli random variables. As shown above, the generating function $G_n$ of $H_n$ has $\min\{n,s\} - \max\{0, n-f\}$ simple negative real roots and a root at $z = 0$ of multiplicity $\max\{0, n-f\}$. As seen in the general discussion, each strictly negative real root gives rise to a Bernoulli random variable with success probability between 0 and 1 and these random variables are independent. The root at $z = 0$ gives rise to $\max\{0, n-f\}$ Bernoulli random variables with success probability one, which, being constant, are independent of each other and any other random variables. The remaining $n - \min\{n,s\}$ Bernoulli variables have success probability zero, and, being constant, are independent of each other and any other random variables. The required number of independent Bernoulli summands has been found, completing the proof of the representation.