Spring 4-28-2015

Using Independent Bernoulli Random Variables to Model Gender Hiring Practices

Kimberly D. Hildebrand

USING INDEPENDENT BERNOULLI RANDOM VARIABLES TO MODEL GENDER HIRING PRACTICES

An honors paper submitted to the Department of Mathematics
of the University of Mary Washington
in partial fulfillment of the requirements for Departmental Honors

Kimberly D Hildebrand
April 2015

By signing your name below, you affirm that this work is the complete and final version of your paper submitted in partial fulfillment of a degree from the University of Mary Washington. You affirm the University of Mary Washington honor pledge: "I hereby declare upon my word of honor that I have neither given nor received unauthorized help on this work."

Kimberly D. Hildebrand
(digital signature)

04/28/15

Using Independent Bernoulli Random Variables to Model Gender Hiring Practices

Kimberly D. Hildebrand

Submitted in partial fulfillment of the requirements for Honors in Mathematics at the University of Mary Washington

December 2014

This thesis by **Kimberly D. Hildebrand** is accepted in its present form as satisfying the thesis requirement for Honors in Mathematics.

DATE	APPROVED
__________	________________________________
	Debra L. Hydorn, Ph.D. thesis advisor
__________	________________________________
	Melody B. Denhere, Ph.D. committee member
__________	________________________________
	Julius N. Esunge, Ph.D. committee member

Contents

Abstract

Gender bias is a problem in the workforce at large. In order for society to progress it is important that hiring practices do not use gender as a competitive factor. Hiring practices based on gender can be represented statistically using Bernoulli Random Variables and the Beta and Binomial distributions. Using the moment generating function (MGF) of the Bernoulli and Binomial Distributions, it is possible to calculate the expected value (mean) and variance for the number of women hires for n positions. The probability generating function (PGF) of a sample size n can be used to find the probability of hiring a specific number of women (X). The PGF when solved for $P(X = 0)$ reveals the probability of no women hired for n positions, while $P(X \leq 1)$ gives the probability that one or no women were hired. A computer program was used to run trials to simulate different male/female distributions using recent data on the proportion of women earning a PhD in a variety of disciplines. The simulations were used to represent hiring results for seven faculty positions. Situations where the female proportion is centered at 0.3, 0.5, and 0.7 were studied to account for a range of situations explained by the data researched. Trials that included random proportions of women for each position were run as well. The Chi-Squared Goodness-of-Fit Test will compare the Binomial cumulative distribution function to the sum of Bernoulli cumulative distribution function in order to find a critical value at which it is acceptable to approximate the Bernoulli distribution with a Binomial distribution for various simulations. Simulations will be run to find the average difference between the probabilities that one or more women are hired. Results reveal that it is actually unusual for employers to hire one or no women for seven positions, which could provide evidence of gender bias and that the Binomial distribution approximates each situation fairly well for varying measures of central tendency.

1 Introduction

One of the major problems that many college graduates face is acquiring a job in the field in which they earned their degree. There are several aspects that employers look for when hiring. For example, employers consider each applicant's qualifications, experience, and educational attainment. However, employers must also contemplate hiring decisions based on personality, attitude, and other character traits that are important to the position, such as communication skills and motivation. For the purposes of this study, the focal qualification for hiring is educational attainment within a discipline. This study concentrates on the chances that, of seven available tenure track faculty positions, either one women or no women are hired.

The Council of Graduate Schools released data that indicates that the proportion of doctoral degrees earned by women increased to approximately 50.4 percent within the last decade. According to the data, the categories or disciplines with the highest proportion of doctorates earned by women in 2008 and 2009 were health sciences, education, public administration and services, and social and behavioral sciences. However, the proportion of degrees in sciences and mathematics earned by women is substantially lower. The results from the study are shown in Table 1 below.

Discipline	Percentage of Female Graduates
Health Sciences	70%
Education	67%
Public Administration and Services	61%
Social and Behavioral Sciences	60%
Arts and Humanities	53%
Biological and Agricultural Sciences	51%
Business	39%
Physical and Earth Sciences	33%
Math and Computer Sciences	27%
Engineering	22%

Table 1: Percentage of Women Doctoral Recipients by Discipline, 2008-2009 (Jaschik)

Based on the data provided by the Council of Graduate Schools, a conclusion could be that the percentage of women PhD graduates should be directly correlated to the number of women hired for faculty positions in each discipline. That is, it is expected that among seven positions in the health sciences for example, about five would be filled by women candidates on average. Thus, only one or no positions filled by women is an unlikely outcome. Similarly, for seven positions open in engineering, it is expected that only one would be filled by a woman on average. In this scenario, one or no positions filled by women is a likely outcome. This assumes that the proportion of women who apply for positions is the same as the proportion that earned a doctorate and that there is no gender bias in hiring.

Suppose that a college has seven positions to fill in a mix of disciplines. For seven positions in a mix of different disciplines, it is more difficult to predict the average number of positions filled by women. It is also difficult to guess how likely one or none positions filled by women would be. If a person is hired only based on their educational attainment, how likely is it that one or no women will be hired out of the seven available positions?

2 Methods

Since each position has a different proportion of females earning a PhD and therefore a different proportion of female applicants, and an applicant either will be hired or not, the Bernoulli distribution is used to represent whether a person hired is male or female. Let X_i be the gender of the person hired for the ith faculty position and let $X_i = 1$ if the person is female and $X_i = 0$ if the person is male. The total number of women hired for n faculty positions is the sum of n independent Bernoulli random variables, $X = X_1 + X_2 + X_3 + ... + X_n$, and X takes the values 0 through n. As the person hired is determined separately for each position, this suggests that the gender of those hired are independent Bernoulli random variables. In the situation to be studied, there are seven faculty positions that need to be filled, leading one to think that the Binomial distribution would be a more appropriate fit for modeling the number of women hired. This would be valid if and only if each position had the exact same proportion of female applicants. The Binomial distribution will however be useful for making comparisons with using the sum of Bernoulli Random Variables when solving for probabilities.

Along with the Bernoulli and Binomial distributions, this study will require one more type of distribution, the Beta distribution, which will be used to randomly generate values for the proportion of female applicants for each position for a simulation study. The Beta distribution models random variables that can have a value between 0 and 1 and the shape of the distribution is determined by two parameters, α and β. For each of the n positions a Beta random variate P_i will be generated to represent the proportion of female applicants for that position. By varying the parameters of the Beta distribution, the average number of women hired and the probability that one or no women are hired for n positions will be studied for hiring situations with different proportions of female applicants.

2.1 Formulas

To study the likelihood of one or no women hired for seven positions, three different equations for sums of Bernoulli Random variables are necessary: the expected value, the variance, and the probability generating function (PGF). The expected value and variance equations can be found using the moment generating function (MGF). Let $M_X(t)$ be the MGF of the sum of n independent Bernoulli random variables and let $M_{X_i}(t)$ be the MGF of the ith Bernoulli Random Variable. Then, using Property D from Section 4.5 we have that $M_X(t) = \prod_{i=1}^{n} M_{X_i}(t))$ (Rice pg 155). The expected value $E(X)$ is calculated by solving for the first derivative of the MGF when $t = 0$. Similarly, the variance $Var(X)$ can be found by first solving for the second derivative of the MGF at $t = 0$ (Rice p. 155) to give $E(X^2)$ and then using the result to find

$Var(X) = E(X^2) - E(X)^2$.

The expected value for the sum of $n = 7$ Bernoulli random variables can be predicted as it is known that the expected value of a Bernoulli random variable is p and the expected value of a Binomial Random Variable is np when all p's are equivalent. One can predict that the expected value of $n = 7$ independent Bernoulli Random variables is $E(X) = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7$. To test this theory, instead of directly solving for the expected value when $n = 7$, I will verify the expected value of X by finding $E(X) = M'_X(0)$ for $n = 1, 2$ and 3 and use inductive reasoning to establish the result.

The same procedure can be followed for the variance of n independent Bernoulli's by combining knowledge of the variances of the Bernoulli and Binomial random variables. It is known that the variance, σ^2 of a Bernoulli random variable is $\sigma^2 = p(1-p)$ and the variance of a Binomial random variable is $\sigma^2 = np(1-p)$. A likely outcome for the variance in this case is that $\sigma^2 = \sum_{i=1}^{7} p_i(1 - p_i)$. As the expected value will be shown by following a pattern for smaller n, the variance can be shown the same way.

The next formula that is necessary to derive is the probability generating function, PGF. The PGF is defined as $G(s) = \sum_{k=0}^{\infty} p_k s^k = E(s^X)$ (Gribakin). The PGF for the Bernoulli random variable with probability p is given as $G(s) = (1 - p) + ps$ if and only if $p_k \neq 0$ or 1, and for the Binomial it is given as the product of the PGFs of n independent Bernoulli random variables, with the same proportion p, and is written as $G_X(s) = ((1 - p) + ps)^n$ (Gribakin). From this, one can conclude that the PGF of independent Bernoulli random variables with different proportions P_i can be found by the general form $\prod_{i=1}^{n}((1 - p_i) + p_i s)$. Using the PGF, the probability of k successes in n trials is the k^{th} derivative of $G_X(s)$ with respect to s evaluated at $s = 0$. Similar to the MGF, finding the pattern is the most efficient way to establish a formula for the PGF. Once the PGF formula is reached, by replacing s with zero, the function will give the probability that zero women were hired. By evaluating the first derivative of the formula of the PGF and substituting s with zero, this gives the probability that one women would be hired.

2.2 Simulations

Once the formulas for $E(X)$, $Var(X)$, and $P(X \leq 1)$ are determined, the next step is to write a computer program to generate $n = 7$ Bernoulli random Variables to represent the gender of the applicants who are hired and calculate the average number of women hired and the probability that one or no women are hired for repeated trials. In this project, the software R is used. The program (Program 1) written is provided in the Appendix. The program works so that a counter, *count*, is set to equal 0 and seven places are established to represent the proportion of female applicants for the seven job openings. The program runs through one loop one thousand times. In each loop, the seven slots are filled with either a zero (a male was hired) or a one (a female was hired). The variable *count* then keeps track of the total number of times a one is observed. At the end of the one thousand trials, the software prints out the five number summary (upper and lower quartiles, median, minimum, and maximum value) of the number of women hired in each trial and the mean for the number of women hired. The standard deviation and a histogram will also be provided.

The seventh line in the program has the formula $p[j] = rbeta(1, a, b, ncp = 0)$ to generate the proportion of women applicants. The Beta distribution has $E(X) = \frac{\alpha}{\alpha+\beta}$ and $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Three main types of Beta distributions used are one skewed right over 0.3, one skewed left over 0.7, and one symmetric around 0.5 to represent multiple hiring situations. The Beta parameters α and β were determined to give these expected values with values ranging mostly between 0.3 and 0.7 following the values of P_i in Table 1. Simulations will also be run using a Uniform$(0.3, 0.7)$ distribution to generate values for P_i by changing the seventh line of the program (in the appendix) to $p[j] = runif(1, min = 0.3, max = 0.7)$ for comparison. With the code written and knowing how to find the necessary equations, trials can now be run.

2.3 Approximating the Distribution of the Sum of Independent Bernoulli Trials

After running the above trials, the Chi-Squared Goodness-of-Fit test, or the Chi-Squared test, will be implemented to find a critical value at which it is acceptable to approximate the distribution of the sum of independent Bernoulli random variables with a Binomial distribution. The Chi-Squared test is a test that evaluates the goodness of fit by comparing the observed values and the expected values. The Chi-Squared test statistic is given by $\chi^2 = \sum_{i=1}^{n} \frac{(O_i-E_i)^2}{E_i}$, where the O_i's are the observed or predicted values and E_i's are the expected values following Table 1 and the results obtained from Program 1. Using the degrees of freedom, a p-value can be determined. Since the p-value is the smallest value for α in which we would reject the null hypothesis, we reject the null hypothesis when the p-value is less than the level of significance, α. The null hypothesis (H_0) and the alternative hypothesis (H_A) for the Chi-Squared test are :

H_0 :The Binomial distribution is a good approximate of the sum of Bernoullis distribution results.

H_A : The Binomial distribution is not a good approximate of the sum of Bernoullis distribution results.

I wrote a second program (Program 2), purpose of this program is to determine for which ranges of values the Binomial distribution will approximate the results of the distribution of the sum of Bernoulli random variables from each of the previous situations. To use the Binomial distribution, we need a value for P_i to represent a common proportion of women applicants among the n positions. Four different approaches to determine p are used. Let $p_1, p_2, ..., p_n$ be the proportion of women applicants for each of the n positions. The four estimates for p used are the mean, the median, the ten percent trimmed mean, and the twenty percent trimmed mean of the n P_i values. This program runs similar to the first program and produces an integer for the number of women hired in each situation based upon the P_i Beta random variate generated to represent the proportion of female applicants for each position given by the parameters of $Beta(\alpha, \beta)$. The Binomial distribution is then created for the number of women hired by substituting the four central tendency values from the data generated by each Beta distribution for the parameter p. The program keeps an overall tally of the frequency of the number of women hired using both the sum of independent Bernoulli random variables and the estimated Binomial distribution. With these values, the Chi-Squared Goodness of Fit test is performed and returns a p-value which will determine if the Binomial distribution is a good approximation for the independent Bernoulli data. The program runs under the premise that there are one thousand colleges or universities that each have seven open tenure-track faculty positions so that the p-value represents the number of times, out of one thousand, where the Binomial distribution was not a good fit for the sum of n independent Bernoulli random variables.

3 Results

3.1 Formulas

Given specific values for the proportions of women applicants for seven faculty positions, one can find the probability of one or no women being hired directly through the following formulas for expected value, variance, and the probability generating function, PGF. The first equation that needs to be derived is the MGF when $n = 7$ as shown below.

Given Bernoulli Moment Generating Function: $M(t) = ((1-p) + pe^t)$

For $n = 7$ Bernoulli random variables the MGF of the sum is:

$$
\begin{aligned}
M_7(t) &= ((1-p_1)+p_1e^t)((1-p_2)+p_2e^t)((1-p_3)+p_3e^t)((1-p_4)+p_4e^t)((1-p_5)+p_5e^t)((1-p_6)+p_6e^t)((1-p_7)+p_7e^t) \\
&= \prod_{i=1}^{7}(1-p_i) + e^t \sum_{j=1}^{7} p_j \prod_{i\neq j}^{7}(1-p_i) + e^{2t}\sum_{j<k}^{7} p_jp_k \prod_{i\neq j,k}^{7}(1-p_i) + e^{3t}\sum_{j<k<l}^{7} p_jp_kp_l \prod_{i\neq j,k,l}^{7}(1-p_i) \\
&+ e^{4t}\sum_{j<k<l<m}^{7} p_jp_kp_lp_m \prod_{i\neq j,k,l,m}^{7}(1-p_i) + e^{5t}\sum_{j<k<l<m<o}^{7} p_jp_kp_lp_mp_o \prod_{i\neq j,k,l,m,o}^{7}(1-p_i) \\
&+ e^{6t}\sum_{j<k<l<m<o<q}^{7} p_jp_kp_lp_mp_op_q \prod_{i\neq j,k,l,m,o,q}^{7}(1-p_i) + e^{7t}\prod_{i=1}^{7} p_i
\end{aligned}
$$

Since the expected value of the sum of n Bernoulli random variables is the sum of their expected values, when $n = 7$ is $E(X) = M_0(0) = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7$. To verify this result using $E(X) = M'_X(0)$ we need to take the derivative of $M_7(t)$ with respect to t and then set $t = 0$. Instead, $E(X) = M'_X(0)$ will be found for $n = 1, 2$ and 3 and a pattern will be established that verifies that $E(X) = p_1 + p_2 + ... + p_7$.

$$
\begin{aligned}
n = 1 : M_1(t) &= (1-p_1) + p_1e^t \\
M'_1(t) &= p_1e^t \\
M'_1(0) &= p_1
\end{aligned}
$$

$$
\begin{aligned}
n = 2 : M_2(t) &= ((1-p_1)+p_1e^t)((1-p_2)+p_2e^t) = (1-p_1)(1-p_2) + p_1e^t(1-p_2) + p_2e^t(1-p_1) + p_1p_2e^{2t} \\
&= (1-p_1)(1-p_2) + e^t(p_1(1-p_2) + p_2(1-p_1)) + p_1p_2e^t \\
M'_2(t) &= e^t(p_1(1-p_2) + p_2(1-p_1)) + p_1p_2e^t \\
M'_2(0) &= p_1 - p_1p_2 + p_2 - p_1p_2 + 2p_1p_2 = p_1 + p_2
\end{aligned}
$$

$$
\begin{aligned}
n = 3 : M_3(t) &= ((1-p_1)+p_1e^t)((1-p_2)+p_2e^t)((1-p_3)+p_3e^t) \\
M'_3(t) &= e^t[(1-p_1)(1-p_2)p_3 + (1-p_1)(1-p_3)p_2 + (1-p_2)(1-p_3)p_1] + 2e^{2t}[(1-p_1)p_2p_3 \\
&+ (1-p_2)p_1p_3 + (1-p_3)p_1p_2] + 3e^{3t}p_1p_2p_3 \\
M'_3(0) &= (1-p_1)(1-p_2)p_3 + (1-p_1)(1-p_3)p_2 + (1-p_2)(1-p_3)p_1 + (1-p_1)p_2p_3 + (1-p_2)p_1p_3 \\
&+ (1-p_3)p_1p_2 + 3p_1p_2p_3 = p_1 + p_2 + p_3
\end{aligned}
$$

Using inductive reasoning we have that for $n = k$, $M_k'(0) = \sum_{i=1}^{k} p_i$ so that when $n = 7$:
$M_7'(0) = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7$, which agrees with the result that the mean of the sum of random variables is the sum of their means.

Similarly, the variance of n independent Bernoulli random variables can be found to be the sum of their variances, $\sigma^2 = \sum_{i=1}^{7} p_i(1 - p_i)$. This can be derived from $E(X^2)$ which is the second derivative of the MGF when $t = 0$, and $var(X) = E(X^2) - E(X)^2$. Instead of taking the second derivative of $M(t)$ when $n = 7$ with respect to t and setting it to zero to find $E(X^2)$, we will find $E(X^2) = M_X''(0)$ for $n = 1, 2$ and 3 and establish the pattern to verify the form for the variance.

$n = 1 : M_1''(t) = p_1 e^t \qquad E(M_1''(0))^2 = p_1 \qquad \sigma^2 = p_1 - p_1^2 = p(1 - p)$

$n = 2 : M_2''(t) = p_1(1 - p_2)e^t + p_2(1 - p_1)e^t + 4p_1 p_2 e^{2t}$

$M_2''(0) = p_1(1 - p_2) + p_2(1 - p_1) + 4p_1 p_2 = p_1 + p_2 + 2p_1 p_2 = E(X^2)$

$\sigma^2 = p_1 - p_1^2 p_2 - p_2^2 + 4p_1 p_2 - (p_1 + p_2)^2 = p - 1 - p_1^2 + p_2 - p_2^2 - (p_1^2 + 2p_1 p_2 + p_2^2 = p_1 - p_1^2 + p_2 - p_2^2$

$= p_1(1 - p_1) + p_2(1 - p_2)$

$n = 3 : M_3''(t) = (p_1(1 - p_1)e^t)(p_2(1 - p_2)e^t)(p_3(1 - p_3)e^t)$

$M_3''(t) = e^t[(1 - p_1)(1 - p_2)p_3 + (1 - p_1)(1 - p_3)p_2 + (1 - p_2)(1 - p_3)p_2] + 4e^{2t}[p_1 p_2(1 - p_3)$
$+ p_1 p_3(1 - p_2) + p_2 p_3(1 - p_1)] + 9e^{3t} p_1 p_2 p_3$

$\sigma^2 = (1 - p_1)(1 - p_2)p_3) + (1 - p_1)(1 - p_3)p_2 + (1 - p_2)(1 - p_3)p_1 + 4p_1 p_2(1 - p_3) + 4p_1 p_3(1 - p_2)$
$+ 4p_2 p_3(1 - p_1) + 9p_1 p_2 p_3 = p_3 - p_2 p_3 - p_1 p_3 + p_1 p_2 p_3 + p_2 - p_1 p_2 - p_2 p_3 + p_1 p_2 p_3 + p_1$
$- p_1 p_2 - p_2 p_3 + p_1 p_2 p_3 + 4p_1 p_2 - 4p_1 p_2 p_3 + 4p_2 p_3 + 4p_1 p_2 p_3 - 4p_1 p_3 4p_1 p_2 p_3 + 9p_1 p_2 p_3$
$= p_1(1 - p_1) + p_2(1 - p_2) + p_3(1 - p_3)$

Using inductive reasoning, we find that for $n = 7$:
$\sigma^2 = M_7''(0) = p_1(1 - p_1) + p_2(1 - p_2) + p_3(1 - p_3) + p_4(1 - p_4) + p_5(1 - p_5) + p_6(1 - p_6) + p_7(1 - p_7)$, which agrees with the result that the variances of the sum of independent random variables is the sum of their varainces.

To find the probability that one or no women are hired we could use the PGF of X, $G_7(s) = \prod_{i=1}^{7}((1 - p_i) + p_i s)$. Instead, we will find $P(X \leq 1)$ for $n = 1, 2$ and 3 to establish a pattern which can be used to find $P(X \leq 1)$ for $n = 7$.

$n = 1 : G_1(s) = ((1 - p_1) + p_1 s)$

$G_1(0) = 1 - p = P(X = 0)$

$$G_1'(s) = p_1$$

$$G_1'(0) = p_1 = P(X = 1)$$

$$P(X \leq 1) = P(X = 1) + P(X = 1) = (1 - p_1) + p_1 = 1$$

$$n = 2 : G_2(s) = ((1 - p_1) + p_1 s)((1 - p_2) + p_2 s) = (1 - p_1)(1 - p_2) + (1 - p_1)p_2 s + (1 - p_2)p_1 s + p_1 p_2 s^2$$

$$G_2(0) = (1 - p_1)(1 - p_2) = P(X = 0)$$

$$G_2'(s) = (1 - p_1)p_2 + (1 - p_2)p_1 + 2p_1 p_2 s$$

$$G_2'(0) = (1 - p_1)p_2 + (1 - p_2)p_1 = P(X = 1)$$

$$P(X \leq 1) = (1-p_1)(1-p_2)+(1-p_1)p_2+(1-p_2)p_1 = 1-p_1-p_2+p_1p_2+p_2-p_1p_2+p_1-p_1p_2 = 1-p_1p_2$$

$$n = 3 : G_3(s) = ((1 - p_1) + p_1 s)((1 - p_2) + p_2 s)((1 - p_3) + p_3 s)$$

$$= (1 - p_1)(1 - p_2)(1 - p_3) + (1 - p_1)(1 - p_2)p_3 s + (1 - p_1)(1 - p_3)p_2 s + (1 - p_2)(1 - p_3)p_1 s$$

$$+ (1 - p_1)p_2 p_3 s^2 + (1 - p_2)p_1 p_3 s^2 + (1 - p_3)p_1 p_2 s^2 + p_1 p_2 p_3 s^3$$

$$G_3(0) = (1 - p_1)(1 - p_2)(1 - p_3) = P(X = 0)$$

$$G_3'(s) = (1 - p_1)(1 - p_2)p_3 + (1 - p_1)(1 - p_3)p_2 + (1 - p_2)(1 - p_3)p_1 + 2(1 - p_1)p_2 p_3 s$$

$$+ 2(1 - p_2)p_1 p_3 s + 2(1 - p_3)p_1 p_2 s + 3p_1 p_2 p_3 s^2$$

$$G_3'(0) = p_1 + p_2 + p_3 - 2p_1p_2 - 2p_1p_3 - 2p_2p_3 + 3p_1p_2p_3 = P(X = 1)$$

$$P(X \leq 1) = 1 - p_1p_2 - p_1p_3 - p_2p_3 + 2p_1p_2p_3$$

By using inductive reasoning, we have that for $n = 7$:

$$G_7(0) = (1 - p_1)(1 - p_2)(1 - p_3)(1 - p_4)(1 - p_5)(1 - p_6)(1 - p_7) = P(X = 0)$$

$$G_7'(0) = \sum_{i=1}^{7} p_i - \sum_{i=1}^{7} 2 \prod_{i \neq j}^{7} p_i p_j + \sum_{i=1}^{7} 3 \prod i \neq j, k7 p_i p_j p_k - \sum_{i=1}^{7} 4 \prod_{i \neq j,k,l}^{7} p_i p_j p_k p_l + \sum_{i=1}^{7} 5 \prod_{i \neq j,k,l,m}^{7} p_i p_j p_k p_l p_m$$

$$- \sum_{i=1}^{7} 6 \prod_{i \neq j,k,l,m,o}^{7} p_i p_j p_k p_l p_m p_o + \sum_{i=1}^{7} 7 \prod_{i \neq j,k,l,m,o,q}^{7} p_i p_j p_k p_l p_m p_o p_q = P(X = 1)$$

$$P(X \leq 1) = 1 - \sum_{i=1}^{7} \prod_{i=1}^{7} p_i p_j + 2 \sum_{i \neq j}^{7} \prod_{i \neq j,k}^{7} p_i p_j p_k - 3 \sum_{i=1}^{7} \prod_{i \neq j,k,l}^{7} p_i p_j p_k p_l + 4 \sum_{i=1}^{7} \prod_{i \neq j,k,l,m}^{7} p_i p_j p_k p_l p_m$$

$$- 5 \sum_{i=1}^{7} \prod_{i \neq j,k,l,m,o}^{7} p_i p_j p_k p_l p_m p_o + 6 \sum_{i=1}^{7} \prod_{i \neq j,k,l,m,o,q}^{7} p_i p_j p_k p_l p_m p_o p_q$$

As shown above, the expected value and variance agree with what was predicted. Although the PGF was not predicted, it became clear what the result would be for $n = 7$ after using smaller values of n to reveal a pattern of alternating signs and increasing coefficients with products of the P_is of increasing order.

3.2 Simulations

If specific values for the proportion of female PhD recipients are known for $n = 7$ positions, the exact values for the expected value, variance, and probability of one or no women hired can be found using the established formulas. However, in many situations the exact proportions of female PhD recipients may not be known and the proportion of female applicants may differ from the proportion of PhD recipients. In these situations, simulations can be run using the distribution of the number of women hired for cases when the proportion of women PhD recipients varies using Table 1 as a guide. Values for the proportion of women applicants will be generated using the Beta distribution and also using the Uniform distribution. The first set of trials run used the Beta distribution with $\alpha = \beta = 12$ so that the distribution of the P_i is symmetric around 0.5 with most of the distribution between 0.3 and 0.7. The parameters of the Beta distribution were established through trial and error and the distribution is shown in Figure 1 below. The average value for this Beta distribution is $E(P_i) = \frac{12}{12+12} = 0.5$ and $Var(P_i) = \frac{12*12}{(12+12)^2(12+12+1)} = 0.01$.

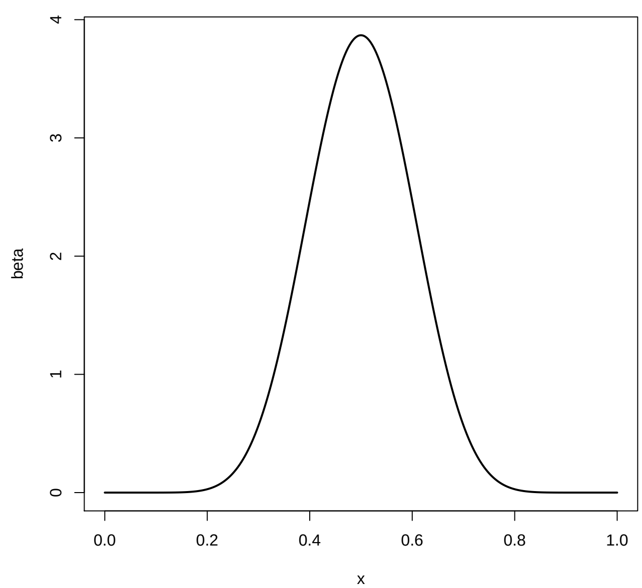


Figure 1: $Beta(12, 12)$

The simulation was run using this Beta distribution and three example histograms for the total number of women hired in $n = 7$ positions are shown below in Figure 2. These histograms show that the number of women hired is symmetric around 3.5 which is expected if the proportion of women applicants is 0.5 and there is no gender bias.

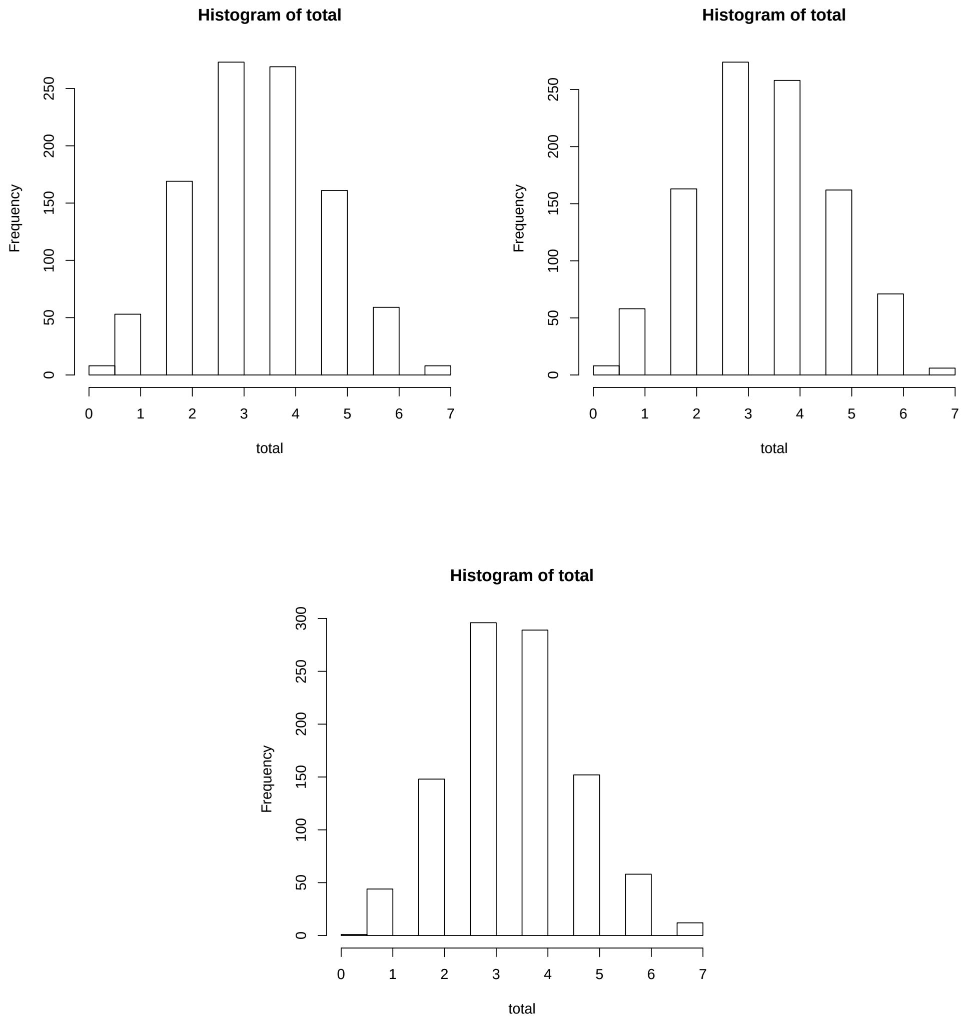


Figure 2: Histograms of Total Number of Women Hired with $Beta(12, 12)$

The next set of trials uses a right skewed Beta distribution over 0.3 with $\alpha = 35$ and $\beta = 50$. This distribution was established through trial and error based on Table 1 and shown in Figure 3 below. The average value for this Beta distribution is $E(P_i) = \frac{35}{35+50} = 0.41$ and $Var(P_i) = \frac{35*50}{(35+50)^2(35+50+1)} = 0.002816$. It represents the situation where most of the $n = 7$ positions are for disciplines that have a lower proportion of women earning a PhD such as in mathematics and computer science.

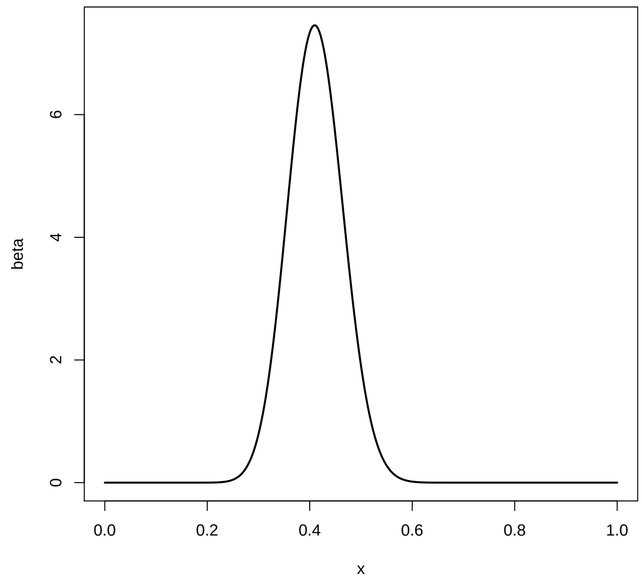


Figure 3: Beta(35,50)

The simulation was run using this Beta distribution and three example histograms for the total number of women hired in $n = 7$ positions are shown above in Figure 4. These histograms show that the number of women hired is skewed to the right with a higher probability that one or no women are hired. In this situation, one or no women hired for n=7 academic positions would not provide strong evidence for a gender bias in hiring since a low number of women hired is a likely outcome.

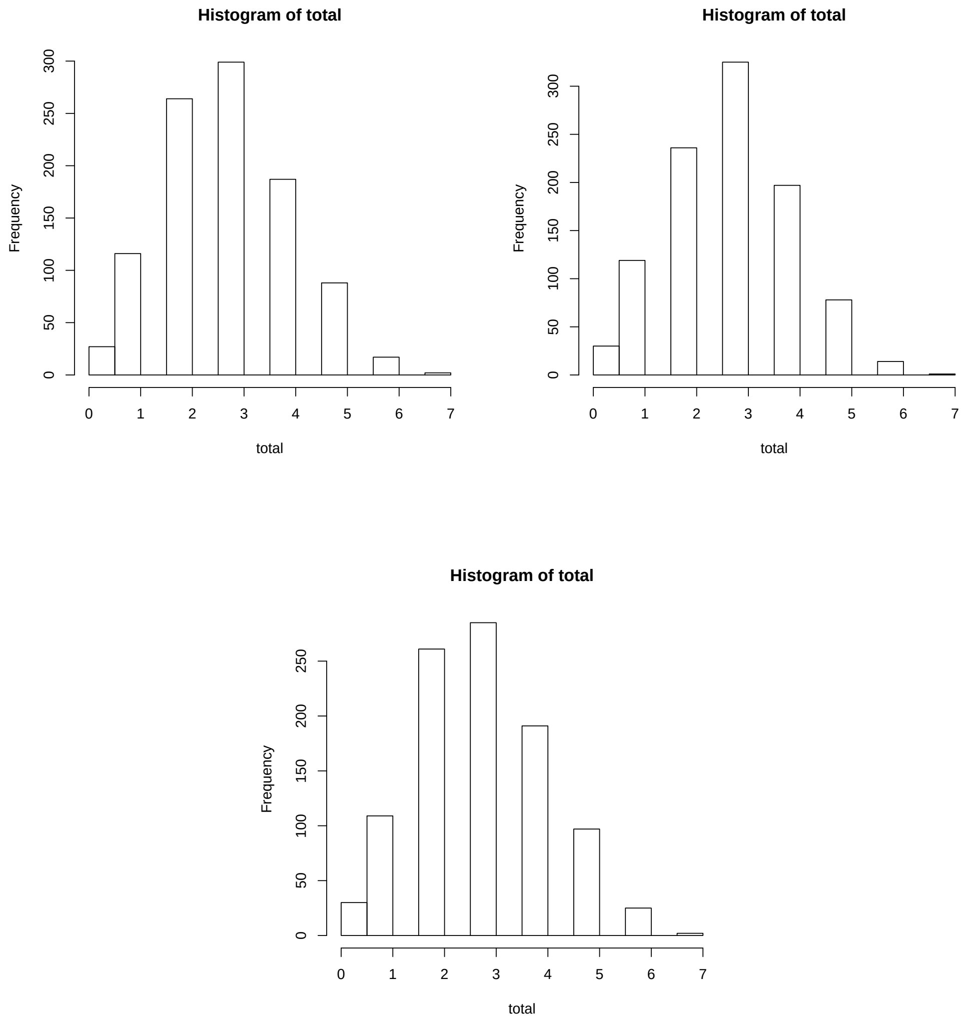


Figure 4: Histograms of Beta Distribution (35,50)

The next set of trials uses a left skewed Beta distribution over 0.7 with $\alpha = 50$ and $\beta = 35$. This distribution was established through trial and error based on Table 1 and is shown in Figure 5. The average value for this Beta distribution is $E(P_i) = \frac{50}{50+35} = 0.625$ and $Var(P_i) = \frac{50*30}{(50+35)^2(50+35+1)} = 0.002816$. It represents the situation where most of the n=7 positions are for disciplines that have a higher proportion of women earning a PhD, as in the the social and behavioral sciences.

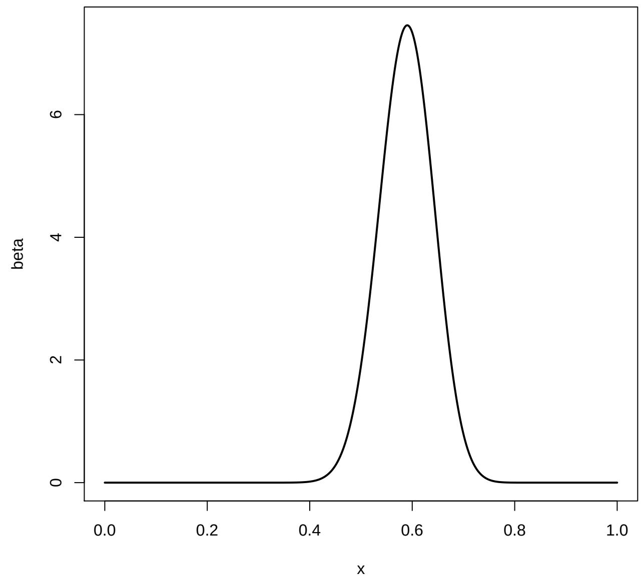


Figure 5: Beta(50,35)

The simulation was run using this Beta distribution and three example histograms for the total number of women hired in $n = 7$ positions are shown below in Figure 6. These histograms indicate that the number of women hired is skewed left with a lower probability that one or no women are hired. In this case, since a low number of women hired is unlikely, one or no women are hired for $n = 7$ academic positions could provide evidence of a gender bias in hiring.

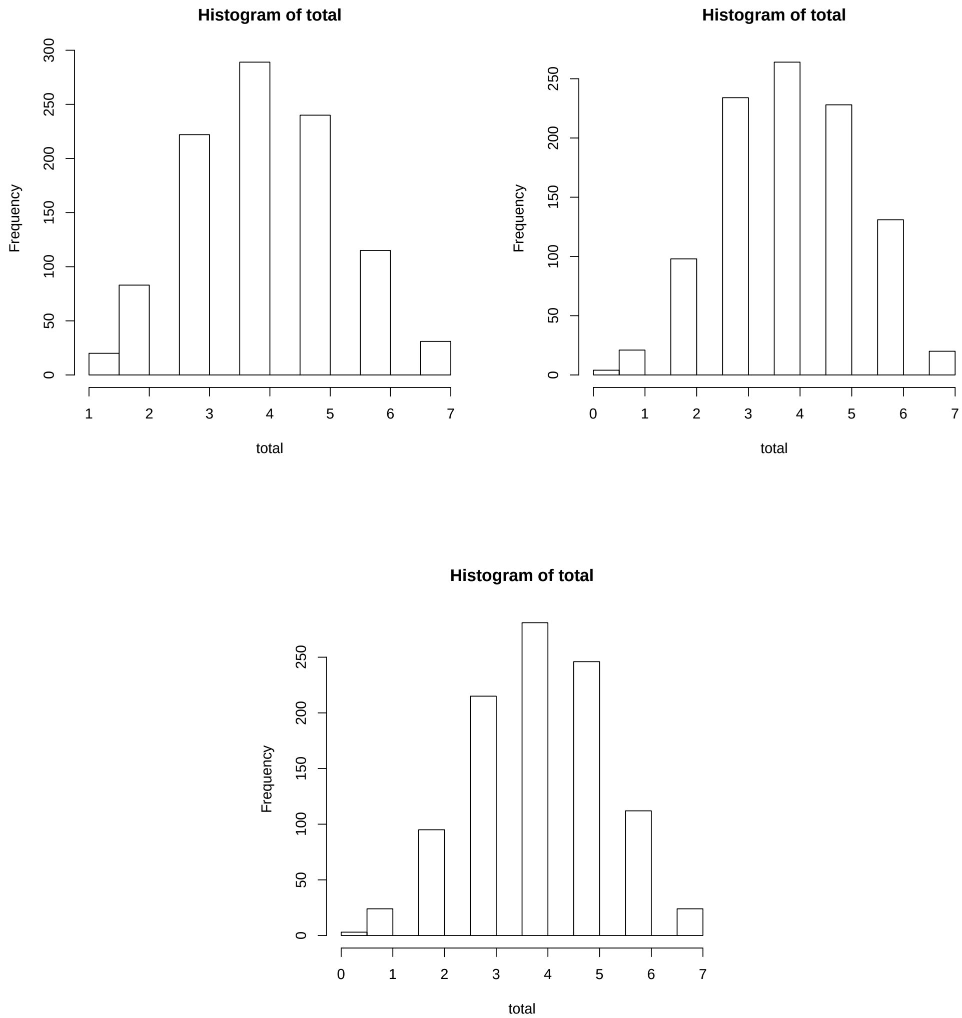


Figure 6: Histograms of Beta Distribution (50,30)

The final set of trials uses the Uniform distribution with a minimum value of 0.3, a maximum value of 0.7, and symmetric about 0.5. The average value for this Uniform distribution is $E(P_i) = \frac{0.3+0.7}{2} = 0.5$ and $Var(P_i) = \frac{(b-a)^2}{12} = \frac{(0.7-0.3)^2}{12} = 0.01333$. It represents the situation where the proportion of men and women hired is symmetrical, about at 0.5 and there is no gender bias. This distribution is shown below in Figure 7.

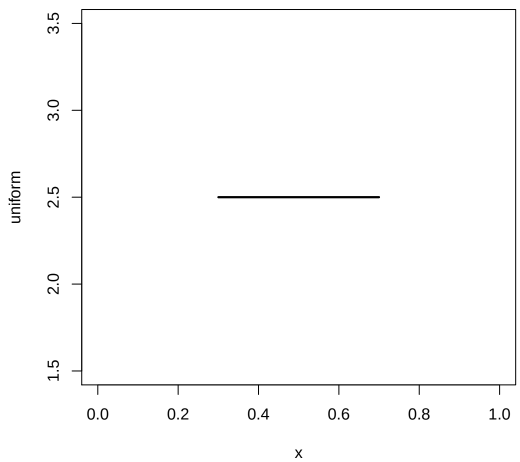
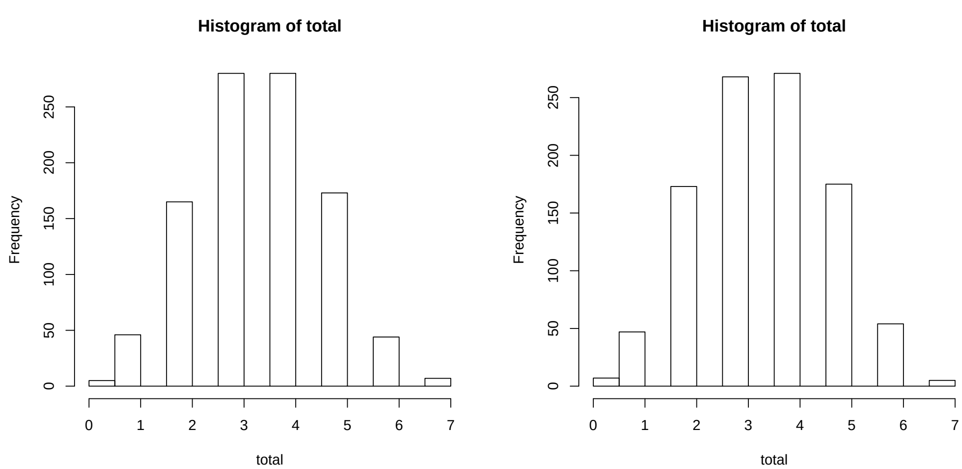


Figure 7: Uniform(0.3,0.7)

The simulation was run using this Uniform distribution and three example histograms for the total number of women hired in n = 7 positions are shown below in Figure 8. These histograms show that the number of women hired is symmetric around 3.5 which is expected if the proportion of women applicants is 0.5 and there is no gender bias. The histograms and results from this distribution should be similar to that of the Beta distribution $Beta(12, 12)$ with an average of 3.5 women hired because the distribution for generating the P_is is centered at 0.5.



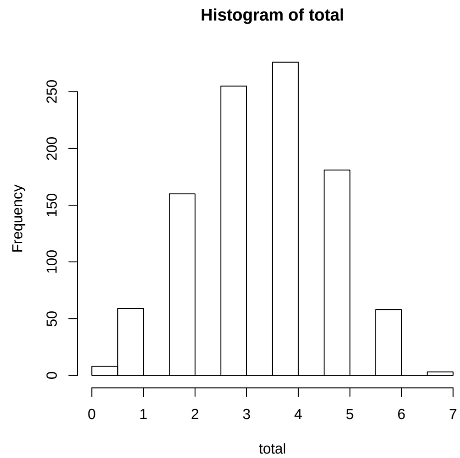


Figure 8: Histograms of Uniform(0.3,0.7)

For comparison, since the Binomial distribution's PGF is $G_X^0(s) = ((1-p)+ps)^n$, it is possible to solve for the probability that one or no women will be hired when $p = 0.3$, $p = 0.5$, and $p = 0.7$. Because the values of n and p are known, it is much simpler to find the formula needed.

$G_7^0(s) = ((1-p)+ps)^7$

$G_7^0(0) = (1-p)^7$

$G_7'(s) = 7((1-p)+ps)^6 p$

$G_7'(0) = 7p(1-p)^6$

When $p = 0.3$ and $n = 7$: $G_7'(0) = 7(0.3)(1-0.3)^6$ the probability of one or no women being hired is 0.2471 or approximately 24.71 percent. This would correspond to the situation where all seven positions are in the disciplines with a low proportion of PhDs are earned by women. When $p = 0.5$ and $n = 7$: $G_7'(0) = 7(0.5)(1-0.5)^6$ the probability of one or no women being hired when there is no gender bias is 0.0547 or approximately 5.47 percent. When $p = 0.7$ and $n = 7$: $G_7'(0) = 7(0.7)(1-0.7)^6$ the probability of one or no women is 0.00357 or approximately 0.357 percent.This corresponds to the situation where all seven positions are in disciplines with a high proportion of PhDs that are earned by women.

The probability that one or no women are hired and the mean and standard deviation of the number of women hired in $n = 7$ positions are provided in Table 2 below. These results are compared for the different distribution used for the proportion of women PhD recipients, P_i.

Distribution of X	Distribution for P_i	$E(P_i)$	$P(X \leq 1)$	Mean	Standard Deviation
Sum of Independent Bernoullis	$P_i \sim beta(35, 50)$	0.41	0.1453	2.867	1.303
	$P_i \sim beta(12, 12)$	0.5	0.0573	3.472	1.326
	$P_i \sim beta(50, 35)$	0.625	0.0221	4.134	1.268
	$P_i \sim uniform(.3, .7)$	0.5	0.0573	3.527	3.137
Binomial	$P_i = 0.3$ for each i	0.3	0.247	2.1	1.212
	$P_i = 0.5$ for each i	0.5	0.0547	3.5	1.323
	$P_i = 0.7$ for each i	0.7	0.00357	4.9	1.212

Table 2: Probability of One or Fewer Females Hired for Seven Faculty Positions

With the results given in Table 2, the chances that one or no women will be hired when seven positions are available tends to be low. The probability of one or no women hired in seven positions depends on the proportion of PhD's earned by women in the disciplines involved. When looking at the sum of the independent Bernoulli random variables as we expected, the largest probability for one or no women to be hired when $Beta(35, 50)$ and the smallest probability for one or no women to be hired when $Beta(50, 35)$. We assumed this because there are fewer number of women applicants, such as the fields of mathematics and physical sciences when compared to the number of women hired when there are more female applicants such as in the humanities. In the scenario where $Beta(35, 50)$ on average out of seven vacant positions, three women will be hired while when $Beta(35, 50)$ an average of four women are hired out of seven open positions. As predicted, the $Beta(12, 12)$ distribution and the $Uniform(0.3, 0.7)$ distribution have the same values for $P(X \leq 1)$ and the $E(X)$ as well as similar values of the average number of about three to four women being hired out of seven positions.

There is a 0.357% chance of one or no women being hired if all of the positions are in the field of humanities, education, and social sciences, for example. This corresponds to the Binomial case with $P_i = 0.7$. The chance of one or no women being hired when the disciplines have equal proportions of men and women earning a PhD is 5.47%. Finally, for the number of women hired with doctorate degrees in disciplines such as math, computer science, and business there exists a 24.7% chance that one or no women will be hired in 7 positions. As expected, when the number of women hired on average is larger, $P(X \leq 1)$ is much smaller.

The probability that one or no women will be hired using the sum of independent Bernoulli random variables to be overestimated by the Binomial distribution when P_i is equal to the lower bound of 0.3.

3.3 Approximating the Distribution of the Sum of Independent Bernoulli Trials

Program 2 gives the results of when the the Binomial distribution is a good fit for the sum of independent Bernoulli random variables. Simulation results for $P(X \leq 1)$ when $E(P_i) = 0.5$ for the Beta and Uniform distributions are close in value to $P(X \leq 1)$ for the Binomial distribution with $p = 0.5$. This leads to the question under what circumstances the Binomial distribution could be used to approximate $P(X \leq 1)$? The goodness of fit is assessed by the Chi-Squared test with the hypothesis that the Binomial distribution is a good fit at a significance level $\alpha = 0.05$. The results are shown in Tables 3, 4, 5, and 6. In each table, the central tendency, $\hat{p}$, standard deviation of the estimated values, and the average number of times H_0 is rejected out of three simulations of one thousand trials are given. Table 3 shows the results of the simulations using the $Beta(12, 12)$ distribution to generate values for the $7P_i$s below.

Central Tendency	Average $\hat{p}$	Standard Deviation	Average Number of Times Reject H_0
Mean	0.5023639	0.1010679	35
10% Trimmed Mean	0.50003599	0.09899328	45
20% Trimmed Mean	0.5005884	0.1016173	41
Median	0.4987311	0.1012174	57

Table 3: Binomial Approximation of Bernoulli Distribution using Beta(12,12)

From these results, the average value of $\hat{p}$ is approximately 0.5 which is expected since the the proportion of women applicants is 0.5 and there is no gender bias for these simulations. Table 4 below shows the results of the same central tendencies for the situation represented by Beta(35,50). This situation occurs when there is a lower proportion of women applicants.

Central Tendency	Average $\hat{p}$	Standard Deviation	Average Number of Times Reject H_0
Mean	0.41031773	0.0531371	46.333
10% Trimmed Mean	0.4108255	0.0533522	44.5
20% Trimmed Mean	0.4099467	0.053633685	41
Median	0.40975985	0.0529993	53.5

Table 4: Binomial Approximation of Bernoulli Distribution using Beta(35,50)

Similarly, from these results, the average value of $\hat{p}$ is approximately 0.41 which is expected since the average proportion of women applicants is 0.41 according to Beta(35,50). Table 5 below shows the results of the same central tendencies for the situation represented by Beta(50,35). This situation occurs when there is a higher proportion of women applicants.

Central Tendency	Average $\hat{p}$	Standard Deviation	Average Number of Times Reject H_0
Mean	0.5584799	0.0533417	44.333
10% Trimmed Mean	0.5871799	0.053381	50
20% Trimmed Mean	0.5907242	0.051996	46.667
Median	0.5898385	0.05293793	49.333

Table 5: Binomial Approximation of Bernoulli Distribution using Beta(50,35)

These results are also as expected. The average value of $\hat{p}$ is approximately 0.58 which is expected since the proportion of women applicants is 0.625 according to Beta(50,35). This result is not as close to the expected average. The last distribution to check is the Uniform distribution with a minimum of 0.3 and a maximum at 0.7. This situation should be similar to that of Beta(12,12), as both situations are centered over 0.5 and again represents the situation where there is an equal proportion of men and women applicants.

The results from Table 6 indicate the average value of p is approximately 0.5 as expected. Compared to Beta(12,12) the standard deviation tends to be a little higher when using the Uniform data to generate values of P_i. The average $\hat{p}$ values are about the same and number of times H_O is rejected appears to be larger for the Uniform distribution approximations.

When comparing the central tendency measures used, the median tends to reject the null hypothesis the most. However, the number of rejections is still relatively low. The Binomial approximation of the Bernoulli distribution when $Beta(50, 35)$ is the only exception, however it is only 0.66667 away from having

Central Tendency	Average $\hat{p}$	Standard Deviation	Average Number of Times Reject H_0
Mean	0.4996742	0.1168897	42.5
10% Trimmed Mean	0.5015904	0.1158968	39.667
20% Trimmed Mean	0.502433367	0.1152253	41
Median	0.49847819	0.1163431	64

Table 6: Binomial Approximation of Bernoulli Distribution using Uniform(0.3,0.7)

the highest rejection rate in that situation. For the lowest number of rejections of H_0, between the four simulations it varies from being the actual mean and the 10% trimmed mean. This should not be surprising since 10% trimmed mean is the closest to the mean of all four scenarios.

Overall, the Binomial distribution with p equal to a central tendency tends to be a good approximation for the sum of independent Bernoulli random variables in the varying situations of women applicants being studied here. Based on the simulation results, it is not possible to conclude if this study provides strong evidence of gender bias among hiring practices. This is due to the fact that there are varying proportions of men and women earning their doctorate degrees in most academic disciplines.

In conclusion, the Binomial distribution appears to be a useful approximate for the sum of seven independent Bernoulli random variables. This result could be useful for approximating the probability of 1 or 0 women hired in n positions when the disciplines are not all the same, particularly for lower values of n.

4 Conclusion

We found that when the probability of a woman being hired is directly proportional to the percentage of women who earned their doctorates in a specific discipline, the chances that one or no women will be hired of seven available positions tends to be low. As expected, there is a larger probability that one or no women will be hired when there are fewer women with doctorate degrees in that field. This also given our results that there is a smaller probability that one or no women will be hired in the case were there is a higher proportion of women earning their degrees within a discipline. Similarly, when trials were ran for no gender bias, the $Beta(12, 12)$ and $Uniform(0.3, 0.7)$ distributions, on average, three to four women were hired giving desired results.

After applying the Binomial distribution with P_i set to the minimum proportion of women hired in each discipline for comparison, the question of can the Binomial distribution approximate the results of the sum of independent Bernoulli random variables in the various situations of women applicants in this study was raised. Using the central tendencies, we found that the Binomial distribution seems to be a useful estimate for the sum of seven independent Bernoulli random variables. This result would be beneficial to expand this study by approximating the probability of 1 or 0 women hired for varying n positions open in various disciplines, particularly for lower values of n.

In conclusion, if the probability of women hired is based on the proportion of women earning their doctorate degree in each discipline, we cannot conclude if there is evidence of gender bias. This is a possibility; however, since there are fewer women earning their degrees in disciplines such as mathematics and physical sciences, there are fewer women who are available for hiring.

5 Future Research Topics

Although the main focus of this study was gender bias in hiring practices, there are many topics of research that can stem from this project. The simplest way to continue research would be to determine how the results change from when $n < 7$ and when $n > 7$. As there are many diverse reasons for why one person is hired for a job over others, bias in some of them can be tested using the same methods. For example, the same simulations can be run with different data based upon the different levels of educational attainment. Or, instead of the probabilities representing employment based on gender, they could similarly represent another type of bias in employment such as race. Additionally, one could determine the circumstances, if any, when the binomial approximation would not be a good fit. Possibilities for this include increasing the range of values for P_i perhaps or examining values of P_i closer to 0 or 1.

6 Acknowledgements

I would like to thank Dr. Debra Hydorn from the University of Mary Washington for her guidance and assistance on this project. I would also like to thank the Jepson Summer Science Institute for giving me the opportunity to begin this project.

I would also like to thank Dr. Melody Denhere and Dr. Julius Esunge for being committee members for my thesis. And, Dr. Randall Helmstutler and Dr. Julius Esunge again for their help with Latex.

7 Appendix

7.1 R code

7.1.1 Program 1

```
count=0                       # to count the number of times 1 or no women are hired in 1000 trials
p=numeric(7)                  # to store the randomly generated proportions of women applicants for
each trial
data=numeric(7)               # to store the Bernoulli results (0 =male, 1 =female)
total=numeric(1000)           # to store the simulated results, the number of women hired in 7 positions
for each trial
for (i in 1:1000){
    for(j in 1:7){
        p[j]=rbeta(1,a,b,ncp=0)          # generates the random proportions using Beta distribution
        data[j]=rbinom(1,1,p[j]) }       # generates the Bernoulli Random Variable
    total[i]=sum(data)                   # the number of women hired in 7 positions
    if(total[i]<=1) {
        count=count+1 } }
cat(count)
sum=summary(total)
cat(”min. 1Q med mean 3Q max”)
cat(sum)
stand=sd(total)
cat(stand)
hist(total)
```

7.1.2 Program 2

```
chicount=0                         #record the number of times H0 is rejected
sim=1000
n=7                                #the number of positions
t=10                               #the number of trials
#lower proportion of women hired; higher probability P¡=1 rbeta(1,35,50,ncp=0) – centered over .3/.4
#equal proportion of women hired rbeta(1,12,12,ncp=0) – centered over .5
#higher proportion of women hired rbeta(1,50,35,ncp=0) – centered over .6/.7
#uniform dist: runif(1,min=.3,max=.7)
p = numeric(7)                     #probability vector
m = numeric(sim)                   #to store the mean values of p
pbi=numeric(8)                     #vector for binomial probabilities
databer=numeric(7)                 #vector for bernoulli data
for (i in 1:sim)

totalber = numeric(t)
for (j in 1:7)

p[j]=rbeta(1,alpha,beta,ncp=0)
#generates prob values based on beta distribution
#p[j]=runif(1,.45,.55)
m[i] =p[j]

meanp = mean(p)
#mean of p
#medp = median(p)
#median of p
#cat(”meanp”,meanp)
for (k in 1:t)

for (j in 1:7)
```

```
databer[j]=rbinom(1,1,p[j])
# cat(”prob”,p[j])

for (j in 1:8)

pbi[j]=dbinom(j-1,7,meanp)          #Change mean; median; trimmed mean

totalber[k]= sum (databer)

sumber¡-numeric(8)          #holds frequencies of number of women hired using bernoulli pro-
portions, 0 to 7
for (j in 1:t)
if(totalber[j]==0)
sumber[1]=sumber[1]+1

if(totalber[j]==1)
sumber[2]=sumber[2]+1

if(totalber[j]==2)
sumber[3]=sumber[3]+1

if(totalber[j]==3)
sumber[4]=sumber[4]+1

if(totalber[j]==4)
sumber[5]=sumber[5]+1

if(totalber[j]==5)
sumber[6]=sumber[6]+1

if(totalber[j]==6)
```

```
sumber[7]=sumber[7]+1

if(totalber[j]==7)
sumber[8]=sumber[8]+1


#cat(”totalber ”, totalber)                    #cat commands to make sure program is working
#cat(”totalbi”, totalbi)
#cat(”pbi”,pbi)
#cat(”sumber ”, sumber)
#cat(”sumbi ”, sumbi)
chi¡-chisq.test(sumber,p=pbi,simulate.p.value=T)
chipv¡-chi$p.value
#cat(”pvalue: ”, chipv)
if(chipv<0.05)

chicount=chicount+1



cat(”alpha=”, alpha)
cat(”beta=”, beta)
cat(”the number of trials =”,t)
cat(”the mean of p is used for the binomial distribution”)
cat(”Out of 1000, we rejected H0 ”,chicount, ” times.”)
cat(”The average value of p is”, mean(m))
cat(”The standard deviation of p is”, sd(m))
```

7.2 Example Output of $P(X \leq 1)$

Out of one thousand schools with seven faculty positions available for hiring, what is the mean and standard deviation?

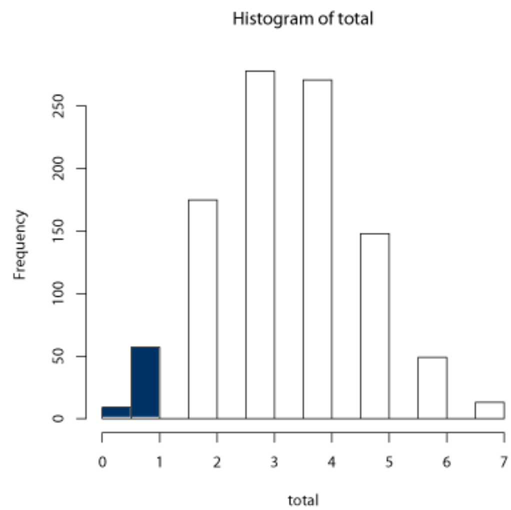
- **Mean** = 3.45

- **Standard Deviation** = 1.3399

– Total represents the one thousand schools

  – Frequency is the number of times each number of women in seven was represented

  – Shaded region: Probability of one or no women hired



7.3 Example Output of Approximating the Distribution of the Sum of Independent Bernoulli Trials

alpha= 35

beta= 50

the number of trials = 10

the mean of p is used for the binomial distribution

Out of 1000, we rejected H0 41 times.

The average value of p is 0.5047159

The standard deviation of p is 0.1156977

alpha= 35

beta= 50

the number of trials = 10

the 10% trimmed mean of p is used for the binomial distribution

Out of 1000, we rejected H0 57 times.

The average value of p is 0.4117338

The standard deviation of p is 0.05221911

alpha= 35

beta= 50

the number of trials = 10

the 20% trimmed mean of p is used for the binomial distribution

Out of 1000, we rejected H0 43 times.

The average value of p is 0.40879

The standard deviation of p is 0.05306582

alpha= 12

beta= 12

the number of trials = 10

the mean of p is used for the binomial distribution

Out of 1000, we rejected H0 31 times.

The average value of p is 0.5077085

The standard deviation of p is 0.1140594

alpha= 12

beta= 12

the number of trials = 10

the 10% trimmed mean of p is used for the binomial distribution

Out of 1000, we rejected H0 60 times.

The average value of p is 0.5002887

The standard deviation of p is 0.1038152

alpha= 12

beta= 12

the number of trials = 10

the 20% trimmed mean of p is used for the binomial distribution

Out of 1000, we rejected H0 44 times.

The average value of p is 0.5016004

The standard deviation of p is 0.1023007

alpha= 12

beta= 12

the number of trials = 10

the median of p is used for the binomial distribution

Out of 1000, we rejected H0 69 times.

The average value of p is 0.4989417

The standard deviation of p is 0.1161674

alpha= 50

beta= 35

the number of trials = 10

the mean of p is used for the binomial distribution

Out of 1000, we rejected H0 47 times.

The average value of p is 0.5001136

The standard deviation of p is 0.1141362

alpha= 50

beta= 35

the number of trials = 10

the 10% trimmed mean of p is used for the binomial distribution

Out of 1000, we rejected H0 48 times.

The average value of p is 0.5874332

The standard deviation of p is 0.05282253

alpha= 50

beta= 35

the number of trials = 10

the 20% trimmed mean of p is used for the binomial distribution

Out of 1000, we rejected H0 47 times.

The average value of p is 0.5879196

The standard deviation of p is 0.05299328

alpha= 50

beta= 35

the number of trials = 10

the median of p is used for the binomial distribution

Out of 1000, we rejected H0 43 times.

The average value of p is 0.5024579

The standard deviation of p is 0.1159749

Uniform (0.3,.07)

the number of trials = 10

the mean of p is used for the binomial distribution

Out of 1000, we rejected H0 47 times.

The average value of p is 0.5009915

The standard deviation of p is 0.1130887

References

[1] Gribakin, Gleb. Probability Generating Functions. Queen's University Belfast, 2002. <http://www.am.qub.ac.uk/users/g.gribakin/sor/Chap3.pdf >

[2] Jaschik, Scott. Women Lead in Doctorates — Inside Higher Ed. Inside Higher Ed, 14 Sept. 2010. Web. June 2013. Queen's University Belfast, 2002. <http://www.insidehighered.com/news/2010/09/14/doctorates >

[3] Rice, John A. Mathematical Statistics and Data Analysis. 3rd ed. Belmont, CA: Thomson/BrooksCole, 2007. Print.