

## A Look at $P(X > Y)$ in the Binomial Case

**J. Behboodian**

Islamic Azad University - Shiraz Branch

**Abstract:** In this article we consider  $P(X > Y)$  for two independent random variables  $X \sim B(n + m, p_1)$  and  $Y \sim B(n, p_2)$ . This is a useful measure in biomedical studies and engineering reliability. The calculation of this probability is discussed by using a combinatorial identity and the approximate value of that is given when  $n$  is large. Finally some special cases are discussed.

**AMS Subject Classification:** 62F10.

**Keywords and Phrases:** Binomial variable, combinatorial identities, conditional probability, central limit theorem.

### 1. Introduction

There are some interesting problems in probability and statistics regarding two independent random variables  $X$  and  $Y$ . One of them is about the exact value, or parametric estimate, or non-parametric estimate of  $P(X > Y)$ . This is a useful measure, for example, in biomedical studies where  $X$  represents the result of an old treatment and  $Y$  the result of a new treatment. This probability is also useful for measuring the reliability of engineering systems ([6; pp 27-30]).

Suppose that  $f$  and  $F$  are the density and distribution of  $X$ , respectively and  $g$  and  $G$  of  $Y$ . Since  $X$  and  $Y$  are independent, conditioning on  $Y$ , we have easily

$$P(X > Y) = 1 - P(X \leq Y) = \begin{cases} 1 - \int_{-\infty}^{\infty} F(z)g(z)dz & (\text{continuous case}) \\ 1 - \sum_z F(z)g(z) & (\text{discrete case}) \end{cases}$$

For example, if  $X \sim N(\mu_1, \sigma^2)$  and  $Y \sim N(\mu_2, \sigma^2)$  are independent, then

$$P(X > Y) = 1 - \Phi\left(\frac{\mu_2 - \mu_1}{\sigma\sqrt{2}}\right),$$

where  $\Phi$  is the standard normal distribution. As another example, if  $X \sim \text{Exp}(\theta_1)$  and  $Y \sim \text{Exp}(\theta_2)$  are independent, then

$$P(X > Y) = \frac{\theta_2}{\theta_1 + \theta_2}.$$

However, when  $X$  and  $Y$  are discrete, the calculation of  $P(X > Y)$  is not always straightforward or the result is not so simple. The purpose of this article is to study  $P(X > Y)$  when  $X \sim B(n + m, p_1)$  and  $Y \sim B(n, p_2)$  are independent.

In Section 2, a simple example is given to show the method of calculation. In Section 3, a general case is considered and the complexity of the problem is discussed. In Section 4, an approximate value for  $P(X > Y)$  is suggested when  $n$  is large by using the Bernoulli representation of  $X$  and  $Y$  and the Central Limit Theorem.

## 2. A Simple Example

We first look at the following simple example before we discuss about the general case.

**Example 1.** Let  $X \sim B(4, \frac{1}{2})$  and  $Y \sim B(3, \frac{1}{2})$  be two independent Bernoulli variables. We can easily find the joint probability table of  $X$  and  $Y$  as follows:

| $y \backslash x$ | 0               | 1                 | 2                  | 3                  | 4                  | $P(Y = y)$    |
|------------------|-----------------|-------------------|--------------------|--------------------|--------------------|---------------|
| 0                | $\frac{1}{128}$ | $\frac{4^*}{128}$ | $\frac{6^*}{128}$  | $\frac{4^*}{128}$  | $\frac{1^*}{128}$  | $\frac{1}{8}$ |
| 1                | $\frac{3}{128}$ | $\frac{12}{128}$  | $\frac{18^*}{128}$ | $\frac{12^*}{128}$ | $\frac{3^*}{128}$  | $\frac{3}{8}$ |
| 2                | $\frac{3}{128}$ | $\frac{12}{128}$  | $\frac{18}{128}$   | $\frac{12^*}{128}$ | $\frac{13^*}{128}$ | $\frac{3}{8}$ |
| 3                | $\frac{1}{128}$ | $\frac{4}{128}$   | $\frac{6}{128}$    | $\frac{4}{128}$    | $\frac{1^*}{128}$  | $\frac{1}{8}$ |
| $P(X = x)$       | $\frac{1}{16}$  | $\frac{4}{16}$    | $\frac{6}{16}$     | $\frac{4}{16}$     | $\frac{1}{16}$     | 1             |

Using this table, we find all the events for which  $X > Y$  and their probabilities (marked by \*). For example,  $P(X = 3, Y = 1) = 12/128$ . Therefore, we obtain

$$P(X > Y) = \frac{4^*}{128} + \frac{6^*}{128} + \dots + \frac{1^*}{128} = \frac{1}{2}.$$

Actually, we have

$$\begin{aligned} P(X > Y) &= \sum_{y=0}^3 \sum_{x=1}^{4-y} P(Y = y, X = y + x) \\ &= \sum_{y=0}^3 \sum_{x=1}^{4-y} \binom{3}{y} \binom{4}{y+x} \left(\frac{1}{2}\right)^7 = \frac{1}{2} \end{aligned}$$

In Section 3 we find a general formula for  $P(X > Y)$ .

### 3. A General Case

Let  $X \sim B(n + m, p_1)$  and  $Y \sim B(n, p_2)$  be two independent binomial random variables. Following the pattern of the above simple example, we obtain:

$$\begin{aligned}
 P(X > Y) &= \sum_{y=0}^n \sum_{x=1}^{m+n-y} P(Y = y, X = y + x) \\
 &= \sum_{y=0}^n \sum_{x=1}^{m+n-y} P(Y = y)P(X = y + x) \\
 &= \sum_{y=0}^n \sum_{x=1}^{m+n-y} \binom{n}{y} p_2^y q_2^{n-y} \binom{m+n}{y+x} p_1^{y+x} q_1^{m+n-y-x} \\
 &= \sum_{y=0}^n \sum_{x=1}^{m+n-y} \binom{n}{y} \binom{m+n}{y+x} p_2^y q_2^{n-y} p_1^{y+x} q_1^{m+n-y-x}
 \end{aligned}$$

This double sum is too complicated and it cannot be simplified. We consider some special cases.

(I) For  $p_1 = q_1 = p_2 = q_2 = \frac{1}{2}$ , we have:

$$P(X > Y) = \left(\frac{1}{2}\right)^{2n+m} \sum_{y=0}^n \sum_{x=1}^{m+n-y} \binom{n}{y} \binom{m+n}{y+x}$$

Here, we can reduce the double sum to a single sum. For this purpose, we use the fact that

$$\binom{N}{k} = 0 \quad ; \quad k > N$$

and we write

$$\sum_{y=0}^n \sum_{x=1}^{m+n-y} \binom{n}{y} \binom{m+n}{y+x} = \sum_{y=0}^n \sum_{x=1}^{m+n} \binom{n}{y} \binom{m+n}{y+x}.$$

Now, we are able to interchange the summation signs and to have

$$\sum_{x=1}^{m+n} \sum_{y=0}^n \binom{n}{y} \binom{m+n}{y+x}.$$

Next, we use the following combinatorial identity Number (10), given in

[5], page 217:

$$\sum_{k=0}^M \binom{M}{k} \binom{N}{R+k} = \binom{M+N}{M+R}.$$

This identity can be proved easily by the usual box-and-balls argument

if we replace  $\binom{M}{K}$  by  $\binom{M}{M-K}$ . Thus, we have:

$$P(X > Y) = \left(\frac{1}{2}\right)^{2n+m} \sum_{x=1}^{m+n} \binom{2n+m}{n+x}.$$

(II) It is interesting to observe that for  $m = 1$  and any integer  $n \geq 1$ ,

we have  $P(X > Y) = \frac{1}{2}$ . This follows from the two identities

$$\binom{N}{K} = \binom{N}{N-K}, \quad \sum_{K=0}^N \binom{N}{K} = 2^N$$

and the fact that

$$\begin{aligned} \sum_{x=1}^{n+1} \binom{2n+1}{n+x} &= \binom{2n+1}{n+1} + \binom{2n+1}{n+2} + \dots + \binom{2n+1}{2n+1} \\ &= \binom{2n+1}{n} + \binom{2n+1}{n-1} + \dots + \binom{2n+1}{0} \\ &= \frac{1}{2} (2^{2n+1}) = 2^{2n}. \end{aligned}$$

You could obtain this result by looking at the  $(2n+1)$  th row of a Pascal Triangle. For  $m = 2$  and  $m = 3$  some rather simple results are obtained by an argument similar to the case  $m = 1$ .

#### 4. Approximation of $P(X > Y)$

As we discussed in Section 3, we cannot simplify  $P(X > Y)$  in a general case. However, we can find an approximate value for this probability when  $n$  is large.

For this purpose we first consider the Bernoulli representation of  $X$  and  $Y$ . Then we apply conditional probability and the Central Limit Theorem.

It is well known that the independent random variables  $X \sim B(n + m, p_1)$  and  $Y \sim B(n, p_2)$  can be expressed in the following way:

$$\begin{aligned} X &= X_1 + X_2 + \dots + X_n + X_{n+1} + \dots + X_{n+m} \\ Y &= Y_1 + Y_2 + \dots + Y_n, \end{aligned}$$

where  $X_1, \dots, X_{n+m}$  are independent Bernoulli variables with success probability  $p_1$  and  $Y_1, \dots, Y_n$  are independent Bernoulli variables with success probability  $p_2$ ;  $X_i$  's are independent from  $Y_j$  's.

Now, let  $U = X_1 + X_2 + \dots + X_n$  and  $W = X_{n+1} + X_{n+2} + \dots + X_{n+m}$ .

It is clear that  $U \sim B(n, p_1)$ ,  $W \sim B(m, p_1)$ , and  $Y \sim B(n, p_2)$  are independent with  $X = U + W$ . We observe that

$$\begin{aligned} P(X > Y) &= P(U + W > Y) = P(Y - U < W) \\ &= \sum_{k=0}^m P(Y - U < W | W = k) P(W = k) \\ &= \sum_{k=0}^m P(Y - U < k) P(W = k) \\ &= \sum_{k=0}^m P(Y - U < k) \binom{m}{k} p_1^k q_1^{m-k}. \end{aligned}$$

Using the above Bernoulli representations, we can write

$$Y - U = (Y_1 - X_1) + (Y_2 - X_2) + \dots + (Y_n - X_n) = \sum_{i=1}^n V_i,$$

where  $V_1, V_2, \dots, V_n$  are independent and identically distributed as

$$\frac{V = v}{P(V = v)} \left| \begin{array}{ccc} -1 & 0 & 1 \\ p_1 q_2 & p_1 p_2 + q_1 q_2 & p_2 q_1 \end{array} \right.$$

with  $E(V) = p_2 q_1 - p_1 q_2 = a$  and  $Var(V) = p_1 q_1 + p_2 q_2 = b$ . Now, by the Central Limit Theorem an approximate value for  $P(Y - U < k)$  can be computed as follows:

$$\begin{aligned} P(Y - U < k) &\approx P(Y - U \leq k - 0.5) \\ &= P\left(\frac{Y - U - na}{\sqrt{nb}} \leq \frac{k - 0.5 - na}{\sqrt{nb}}\right) \\ &\approx P\left(Z \leq \frac{k - 0.5 - na}{\sqrt{nb}}\right) \end{aligned}$$

$$= \Phi\left(\frac{k - 0.5 - na}{\sqrt{nb}}\right) = h(k; a, b),$$

where  $Z \sim N(0, 1)$  has distribution  $\Phi$ . Thus, we have:

$$P(X > Y) \approx \sum_{k=0}^m h(k, a, b) \binom{m}{k} p_1^k q_1^{m-k}$$

The exact value of the probability, for  $m = 1$  and  $n \geq 1$ , and independent  $X \sim B(n+1, p)$  and  $Y = B(n, p)$  is

$$\begin{aligned} P(X > Y) &= qP(Y - U < 0) + pP(Y - U < 1) \\ &= qP(Y - U < 0) + p[1 - P(Y - U < 0)] \\ &= p + (q - p)P(Y - U < 0) < \frac{1}{2} \end{aligned}$$

This follows from the fact that  $Y - U$ , i.e., the difference of two independent random variables  $Y$  and  $U$  with common distribution  $B(n, p)$ , is symmetric about zero with positive probabilities at  $0, \pm 1, \pm 2, \dots, \pm n$ . For  $p = q = \frac{1}{2}$  we have  $P(X > Y) = \frac{1}{2}$ . This is the same answer we obtained in Section 3 by a combinatorial analysis.

It may be useful to observe that for two independent binomial variables  $Y \sim B(n_1, p_1)$  and  $U \sim B(n_2, p_2)$ , the probability function of  $Y - U$  with  $p_1 = p_2 = \frac{1}{2}$  is

$$P(Y - U = k) = \left(\frac{1}{2}\right)^{n_1+n_2} \binom{n_1+n_2}{n_2+k}, \quad k = 0, \pm 1, \pm 2, \dots, \pm n.$$

For obtaining this probability function, it is easy to show that  $Y - U + n_2$  has binomial distribution  $B(n_1 + n_2, \frac{1}{2})$ . This can be proved by using the moment generating function of  $Y - U + n_2$  or the fact that  $Y + n_2 - U$  is the sum of two independent binomial variables with distributions  $B(n_1, \frac{1}{2})$  and  $B(n_2, \frac{1}{2})$ . Now,  $P(Y - U = k) = P(Y - U + n_2 = k + n_2)$  gives the result. Of course, for a general case, we cannot find a simple expression ([1;p 55]).

### References

- [1] N. L. Johnson and S. Kotz, *Discrete distributions*, Houghton Mifflin Company, 1969.
- [2] M. S. Klamkin, A probability of more heads, *Mathematics Magazine*, 44 (1971), 146-149.
- [3] M. S. Klamkin, Symmetry in probability distributions, *Mathematics Magazine*, 61 (1988), 193-194.
- [4] S. M. Ross, *A first course in probability*, 6th edition, Prentice Hall, 2002.
- [5] A. Tucker, *Applied combinatorics*, 3rd edition, John Wiley, 1995.
- [6] D. A. Wolfe, and R. V. Hogg, Constructing statistics and reporting data, *The American Statistician*, (1971), 27-30.

### Javad Behboodian

Department of Mathematics  
Islamic Azad University - Shiraz Branch  
Shiraz, Iran  
Email: Behboodian@stat.susc.ac.ir