Binomial distribution

Suppose you play a purely random game that exactly has 40 % chance of success (i.e., 60 % chance of failure). In other words, the outcome of your game doesn't depend on your ability, but only on luck. If you play this game 4 times, what is the probability that you will succeed 2 times? See Table 1. There are six possible cases for the 2 success out of 4 games. In other words, $\binom{4}{2}$ ways.

What is the probability for each case? In the last column of the table, you see the probability. For example, in case of the 4th case, you fail the 1st game, which has the probability of 0.4, and you succeed the 2nd game, which has a probability of 0.6, and you succeed the 3rd game, which has a probability of 0.6, and you fail the 4th game, which has a probability of 0.4. So, the probability is given by $0.4 \times 0.6 \times 0.6 \times 0.4$ as in the last column. However, notice that all the probabilities for each case is the same; they are all given by $(0.6)^2(0.4)^2$, because they are the cases for 2 success and 2 failure. It doesn't matter, in which order you succeed or fail, as the order doesn't matter in multiplications.

In conclusion, the answer to our question is

$$\binom{4}{2}(0.6)^2(0.4)^2 = 6(0.6)^2(0.4)^2 = 0.3456 \tag{1}$$

In other words, 34.56 %.

	1st game	2nd game	3rd game	4th game	probability
1	○	○	×	×	$0.6 \times 0.6 \times 0.4 \times 0.4$
2	○	×	○	×	$0.6 \times 0.4 \times 0.6 \times 0.4$
3	○	×	×	○	$0.6 \times 0.4 \times 0.4 \times 0.6$
4	×	○	○	×	$0.4 \times 0.6 \times 0.6 \times 0.4$
5	×	○	×	○	$0.4 \times 0.6 \times 0.4 \times 0.6$
6	×	×	○	○	$0.4 \times 0.4 \times 0.6 \times 0.6$

Table 1: There are six possible ways in which you win 2 games out of 4. ○ denotes a succcess and × denotes a failure.

Problem 1. A game has exactly 75 % chance of success. If you play this game 4 times, what is the probability that you will succeed 3 times?

More generally, if a game has the probability of p for success and the probability of

$q(= 1 - p)$ of failure, the probability that you succeed k times out of n games is given by

$$f(k, n, p) = \binom{n}{k} p^k q^{n-k} = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

This is called "binomial distribution." Notice that, if you play a game n times, you will succeed either 0 times, 1 time, 2 times, $\cdots$, n times. In other words, if you add up the probability that you will succeed 0 times, 1 time, 2 times, $\cdots$, n times, it will be 1. Thus, we necessarily have

$$\sum_{k=0}^{n} f(k, n, p) = 1 \tag{3}$$

What is the expectation value of your number of success? The success probability for one game is p. Thus, if you play n times, you will succeed np times on average. If you recall our earlier article "Standard deviation of the sample means," this is the "easy" way to solve. What would be the "difficult" way to solve? Recall that the expectation value of x is given by

$$\langle x \rangle = \sum_i x_i p_i \tag{4}$$

i.e., the sum of each value multiplied by its probability. Thus, $\langle k \rangle_n$ the expectation value of number of success for n games is

$$\langle k \rangle_n = \sum_{k=0}^{n} k f(k, n, p) \tag{5}$$

Now, you have to plug in (2) into the above formula, so this is indeed a difficult calculation, but you are guaranteed to obtain np at the end.

At this point, it might be useful to see some examples of the binomial distribution.

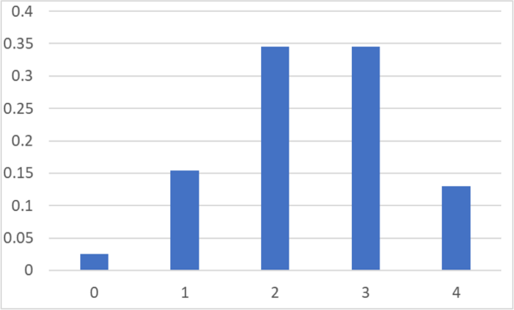

Figure 1: $n = 4$, $p = 0.6$

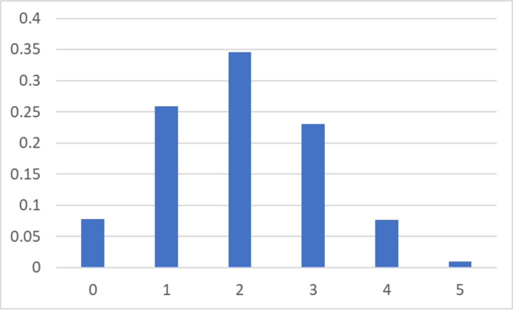

Figure 2: $n = 5$, $p = 0.4$

In Fig. 1, we have the binomial distribution for $n = 4$ and $p = 0.6$. Therefore, the expectation value np is given by 2.4. So, we expect the probability will be the highest around this number. I say "around this number," because the value k, the number of success, can be only non-negative integer. Indeed, the probability is biggest for $k = 2$ and 3. Notice also that the probability for $k = 4$ is bigger than the one for $k = 0$. It's because p(=0.6) is bigger than $q(= 1 - p = 1 - 0.6 = 0.4)$; it is easier to get 4 successes in a row than 4 failures in a row, because a success (60%) is easier than a failure (40%).

In Fig. 2, we have the binomial distribution for $n = 5$ and $p = 0.4$. The expectation value np is given by 2. So, we expect the probability will be the highest around this number. Indeed, this is the case. $k = 2$ has the probability of almost 35 %. It actually sounds very reasonable that the k for the highest probability must be np, if np happens to be an integer. Actually, it can be shown rigorously that this is always the case by doing some calculations. Also, the probability for $k = 5$ is smaller than the one for $k = 0$. It is because the success probability $p = 0.4$ is smaller than the failure probability $q = 1 - p = 0.6$.

Now, what is the standard deviation of the binomial distribution (2)? To obtain the standard deviation, we must first calculate the variance. Again, there are an easy way and a difficult way to calculate the variance. The difficult way is straightforward. Only the calculation is difficult. You subtract the mean value np from k, square them, and average them. In other words,

$$\mathrm{Var}(k)_n = \sum_{k=0}^{n} (k - np)^2 f(k, n, p) \tag{6}$$

Now, the easy way. Recall that this is a purely random game based on luck. The outcome of the next game doesn't depend on the outcome of the previous games. Therefore, the variance can be added for each game. In other words, the variance of k for n games is the variance of k for 1 game multiplied by n, i.e.,

$$\mathrm{Var}(k)_n = n\mathrm{Var}(k)_1 \tag{7}$$

So, let's calculate the variance of k for 1 game. We have

$$\mathrm{Var}(k)_1 = \langle k^2 \rangle_1 - (\langle k \rangle_1)^2 \tag{8}$$

Problem 2. Calculate this by using (4) or (5)

If you correctly calculate you will get

$$\mathrm{Var}(k)_1 = p(1 - p) = pq \tag{9}$$

Thus, we conclude the variance of binomial distribution is $npq = np(1 - p)$. In other words, the standard deviation of binomial distribution is given by

$$\sigma = \sqrt{np(1 - p)} \tag{10}$$

Binomial distribution had wide applications. For example, suppose you want to know what proportion of the villagers support the construction of nuclear power plant in a certain village. For simplicity, I will assume the villagers either support or object the construction of nuclear power plant. There is no third way. Anyhow, to know the proportion exactly, you need to do a referendum, i.e., asking every villager, but that costs too much money. So, you randomly choose a certain number of people, a number big, but much smaller than the total population of the village, ask them, and estimate the true proportion of nuclear power plant construction supporters. If the true proportion is p and the number of villagers that answered the survey is n, the number of respondents that support the nuclear power plant construction, k will follow the binomial distribution $f(k, n, p)$. Putting it slightly differently, the proportion of nuclear power plant supporters among respondents, k/n will follow the distribution of mean with $p(= np/n)$ with the standard deviation of

$$\frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}} \tag{11}$$

The bigger the number of survey respondents (i.e., n), the better we can estimate the proportion, as the standard deviation of the proportion is smaller for bigger n, which you can see from the above formula. For example, if the true p is 0.5 (i.e., 50%) and n is 2,500, the standard deviation of the proportion obtained from the survey will be 0.5 ± 0.01 or $(50 \pm 1)\%$. So, if you perform the survey, you will obtain the value somewhere around 50 % but deviates from 50 % by around 1 %.

Suppose, we do not know the true p, but obtained the proportion of 50.3% for the construction from a survey. Then, can we say that there are more people who support the construction than object it? The proportion of respondents who objected it was 49.7%, so the difference is 0.6%, which is a value smaller than the error 1%, so we can't be quite sure. If the proportion of supporter among respondents was 50.7%, the difference is 1.4%, so we can be quite sure that there are more supporters than non-supporters, but we can't be 100 % sure, because more proportion of supporters could have been included in the survey by chance. So, if you read an article on survey, you will find the expression that "the error is so and so percent for the confidence level of 95 %". It means that there can be 5(=100-95) % chance that the proportion obtained is bigger or smaller than the true proportion by so and so percent.

How can we calculate such probabilities? We can of course calculate them from the binomial distribution $f(k, n, p)$, but there is an easier way to calculate them by using an approximation. We will talk more about it in the next article.

Final comment. Suppose you decided to call a success by a "failure" and a "failure" by a success from now on. Then, the probability of success now will be the probability of failure earlier, i.e., $1 - p(= q)$, and the probability of failure now will be the probability of success earlier, i.e., p. Also, the probability that you now succeed $n - k$ times out of n times must be equal to the probability that you fail $n - k$ times out of n times earlier, i.e., the probability that you succeed k times out of n times earlier. Thus, we conclude

$$f(n-k, n, 1-p) = f(k, n, p) \tag{12}$$

Problem 2. Explicitly check the above relation is satisfied by using (2).

Problem 3. In Fig. 2, we have the binomial distribution of $n = 5$ and $p = 0.4$. How will the figure of the binomial distribution of $n = 5$ and $p = 1 - 0.4 = 0.6$ look like compared to

Fig. 2? Explain why the standard deviation of the both cases will be the same, and check this from (10).

Summary

- Let's say a game, whose outcome purely depends on the luck, has a probability of p of success. The binomial distribution tells you what is the probability that you succeed k games out of n games.
- This probability is given by

$$f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- The expectation value of k is given by np and the standard deviation is given by $\sqrt{np(1-p)}$.