Binomial Distribution

Nikos Apostolakis

March 19, 2024

1 Binomial Distribution

Formulas: n is the number of trials p the probability of one of the outcomes, called *success*, and $q = 1 - p$ the probability of the other outcome, the *failure*.

$$p(X = r) = \binom{n}{r} p^r q^r,$$

where,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

We have formulas for the mean μ, and the standard deviation σ.

$$\mu = np, \quad \sigma^2 = npq, \quad \sigma = \sqrt{npq}. \tag{1}$$

How to use the tables

There are two sets of tables. One set shows the *probability function*, and the other shows the *cumulative probability function*. The pages of the first set have $p(X = r)$ on top center, while the pages of the second have $p(X \leq)$. The first table gives gives the values of what is called the *PDF* (*Probability Density Function*), and the second what is called the *CDF*, (*Cumulative Density Function*). The PDF is what we called the *probability distribution* of X. For any value of r the pdf-table gives the probability that X will take that particular value, i.e. $p(X = r)$. On the other hand the CDF gives the probability that X will take the value r or less, i.e. $p(X \leq r)$. We call that probability *the cumulative probability up to* r, because it is the sum of the probabilities of all the values up to r. Look at Figure 1, where the histogram of the binomial distribution for $n = 10$ and $p = .5$ is shown. For $r = 6$ the PDF file will give you $p(X = 6)$, which in the histogram is represented by the area of the box at $r = 6$ (the left side); on the other hand the CDF table gives the value $p(X \leq 6)$, that is the area of all the boxes from 0 up to 6 (the right side).

In principle we don't need two different tables. If we know the PDF tables we can get the value of CDF table at the column r by adding all the values up to r of the PDF. In practice though that's a lot of work, we don't really want to add 10 values of the PDF table if we want to find the probability $p(X \leq 10)$.

It turns out that if we have the CDF table we can get the PDF. To see how that happens, look at Figure 2 we have the corresponding rows of the PDF and CDF tables of the binomial distribution for $n = 14$ and $p = 0.30$. As we go down the table the values of each cell in the CDF is the sum of its previous row of the the row in the PDF next to it.

The first rows of the the tables are that same, both are 0.0024. Then the second row in the CDF (on the right) is the first row of the CDF plus the second row of the PDF

$$0.0024 + 0.0181 = 0.0205.$$

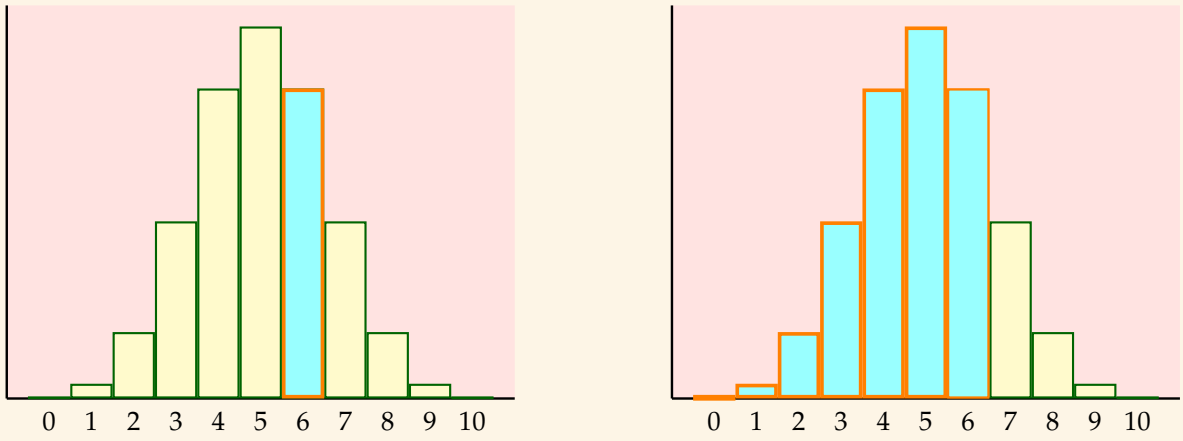


Figure 1: PDF (left) and CDF (right).

If we add to the second row of the the CDF the third row of the PDF we get the third row of the CDF.

$$0.0205 + 0.0634 = 0.0839.$$

If we add to the third row of the the CDF the fourth row of the PDF we get the fourth row of the CDF.

$$0.0024 + 0.0181 = 0.0205.$$

And this patterns continues, until the end.

	PDF		CDF
r	.30		.30
0	0.0024		0.0024
1	0.0181		0.0205
2	0.0634		0.0839
3	0.1366		0.2205
4	0.2022		0.4227
5	0.2178	⋮	0.6405
6	0.1759		0.8164
7	0.1082		0.9247
8	0.0510		0.9757
9	0.0183	⋮	0.9940
10	0.0049		0.9989
11	0.0010		0.9999
12	0.0001		1.0000
13	0.0000		1.0000
14	0.0000		1.0000

Figure 2: The columns of the PDF and CDF tables for Binomial Distribution with $n = 14$, $p = .30$.

This means that if we have the CDF tables we can get the PDF tables by subtracting consecutive rows. The first rows are the same, the second row of the PDF is the second row minus the first row of the CDF.

So *in principle* we need only one table, the PDF or the CDF because we can calculate the other table if we need it. In practice though, each table is best for different kind of problems. That said, generally speaking, the CDF tables are more useful.

Using the PDF tables

The PDF tables are most useful when we want to find the probability that X takes one particular value, or a few isolated values.

Basic examples of using the PDF tables

Let X be a binomial distribution with $n = 30$ and $p = .35$. Calculate the following probabilities.
(a) $p(X = 11)$.

(b) $p(X = 20)$.

(c) $p(X = 8$ **or** $X = 14)$.

Answer. We use the PDF table for $n = 30$, $p = .35$, and look up the values.
(a) $p(X = 11) = 0.1471$.

(b) $p(X = 20) = 0.0003$.

(c) We have two isolated values. So we look the probabilities for $X = 8$ and $X = 14$.

$$p(X = 8 \text{ or } X = 14) = p(X = 8) + p(X = 14) = 0.1009 + 0.0611 = 0.162.$$

□

Using CDF tables

We use the CDF tables when we want to calculate the probability that X takes a value within a range. The basic uses are shown below:

- $p(X \leq r)$. This is given by the CDF table.
- $p(X > r) = 1 - p(x \leq r)$.
- $p(a < X \leq b) = p(X \leq b) - p(X \leq a)$.

Note: Sometimes the questions involve different types of inequalities. Because the values of X are always natural numbers we can convert any question to one of the three types above. For example:

- $p(X < 5) = p(X \leq 4)$.
- $p(X \geq 8) = p(X > 7)$.
- $p(9 \leq X \leq 16) = p(8 < X \leq 16)$.
- $p(5 < X < 11) = p(5 < X \leq 10)$.
- $p(7 \leq X < 13) = p(6 < X \leq 12)$.

Example. Let X be a binomial random variable with $n = 20$, and $p = .70$. Find the following probabilities:
(a) $p(X \leq 9)$.

Answer. We can get this directly from the CDF table. $p(X \leq 9) = 0.0171$. □

(b) $p(x \geq 14)$.

Answer. The *complementary* event of $X \geq 4$ is $X \leq 13$. From the table we see that $p(X \leq 13) = 0.3920$. So

$$p(X \geq 14) = 1 - 0.3920 = 0.608$$

□

(c) $p(9 < X \leq 18)$.

Answer. We have:

$$p(9 < X \leq 18) = p(X \leq 18) - p(X \leq 9) = 0.9924 - 0.0171 = 0.9753.$$

□

(d) $p(12 \leq X \leq 16)$.

Answer. $12 \leq X \leq 16$ is the same as $11 < X \leq 16$. So we have

$$p(12 \leq X \leq 16) = p(11 < X \leq 16) = 0.8929 - 0.1133 = 0.7796.$$

□

Solved problems

1. A biologist is studying a new hybrid tomato. It is known that the seeds of this hybrid have 75% germination rate, that is the probability that a seed will germinate is 0.75. She plants 12 seeds.
   (a) What is the probability that *exactly* 9 seeds will germinate?
   (b) What is the probability that all 12 seeds will germinate?
   (c) What is the probability that no more than 7 seeds will germinate?
   (d) What is the probability that *at least* 6 will germinate?
   (e) What is the probability that more than 4 but less than 9 seeds will germinate?

   Answer. We have $n = 12$ independent trials (the 12 seeds) of an experiment that has probability of success, $p = 0.75$. So we have a random variable X with a binomial probability distribution with parameters $n = 12$ and $p = 0.75$. We will use the tables to answer the questions.

   The first two questions ask for the probability that X takes a particular value *exactly*. So we will use the tables that give the probabilities $p(X = r)$. The last three question ask for the probability that the value of X is in a *range*, so we will use the tables that give the *cumulative* probabilities $p(X \leq r)$.

   (a) From the tables we see that $p(X = 9) = 0.2581$.
   (b) From the same table we have $p(X = 12) = 0317$.
   (c) *No more than* 7 means $X \leq 7$. This probability is given by the tables of the *cumulative* probability. From those tables we see that $p(X \leq 7) = 0.1576$.
   (d) *At least* 6 means $X \geq 6$, and this probability is not given by the table. So we will use the *complementary* event of $X \geq 6$. If it's *not* the case that $X \geq 6$ then $X \leq 5$, and we car read the probability of the later for the table. We have,

$$\begin{aligned} p(X \geq 6) &= 1 - p(X \leq 5) \\ &= 1 - 0.0143 \\ &= 0.9857. \end{aligned}$$

(e) We are asked to calculate $p\,(4 < X < 9)$ or equivalently $p\,(4 < X \leq 8)$. Now

$$p\,(4<X\leq 8) = p\,(X \leq 8) - p\,(X \leq 4)\,.$$

So from the cumulative tables we have

$$p\,(4 < X \leq 8) = 0.3512 - 0.0028 = 0.3484.$$

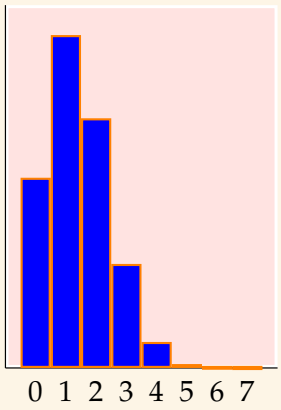
□

2. According to a study published by *USA Today*, about 20% of all people in United States are illiterate. Suppose that you interview seven people chosen at random in the United States.
   (a) Use the binomial distribution tables to construct a histogram that shows the probability distribution of the random variable X that counts the number of illiterate people out of those seven.
   (b) Find the mean and the standard deviation of X. How many people out of those seven should we expect to be illiterate?

Answer. For each of the randomly chosen people we have that the probability of being illiterate is $p = 0.20$. We chose seven people so we have a binomial distribution with $n = 7$ and $p = 0.20$. From the tables, for $n = 7$ and $p = .20$, we can get the probability distribution.

r	0	1	2	3	4	5	6	7
$p(r)$	0.2097	0.3670	0.2753	0.1147	0.0287	0.0043	0.0004	0.0000

So we get the histogram:



From the formulas we have $\mu = np = 0.2 \cdot 7 = 1.4$, and $\sigma = \sqrt{npq} = \sqrt{7 \cdot 0.2 \cdot 0.8} \approx 1.058$. So we should expect 1.4 people. □

Remark 1.0.1. It sounds silly to say 1.4 people. One way to understand statements like that is that if we do this, seven times say, then we will get $5 \cdot 1.4 = 5$ people.

3. A fair coin is tossed 20 times. Let X by the random variable that counts how many times it comes up heads.
   (a) What is the expected value of X?
   (b) What is the probability that X will come up heads exactly 10 times?
   (c) What is the standard deviation of X?
   (d) What is probability that X will come up heads more than five but less than fifteen times?
   (e) Compare the results above with the results you get from the Empirical Rule. Explain why we are justified to apply the Empirical rule.

Answer. Since the coin is *fair* the probability of coming with heads up is 50% or as a decimal 0.5. So We have a binomial distribution with $n = 20$, and $p = .50$. We also have $q = 1 - 0.50 = 0.50$.

(a) By the formulas (1) we have that

$$mu = n \cdot p = 20 \cdot 0.50 = 10.$$

(b) From the table we see that

$$p(X = 10) = 0.1593.$$

(c) Again fro the formulas (1) we have

$$\sigma = \sqrt{npq} = \sqrt{20 \cdot 0.5 \cdot 0.5} \approx 2.236.$$

(d) We have

$$\begin{aligned} p\,(5 < X < 15) &= p\,(5 < X \leq 14) \\ &= p(X \leq 14) - p(X \leq 5) \\ &= 0.9423 - 0.0207 \\ &= 0.9216. \end{aligned}$$

(e) The histogram of any binomial distribution with $p = .50$ is bell shaped. The histogram for $n = 20$ and $p = 0.50$ is shown in Figure 3.

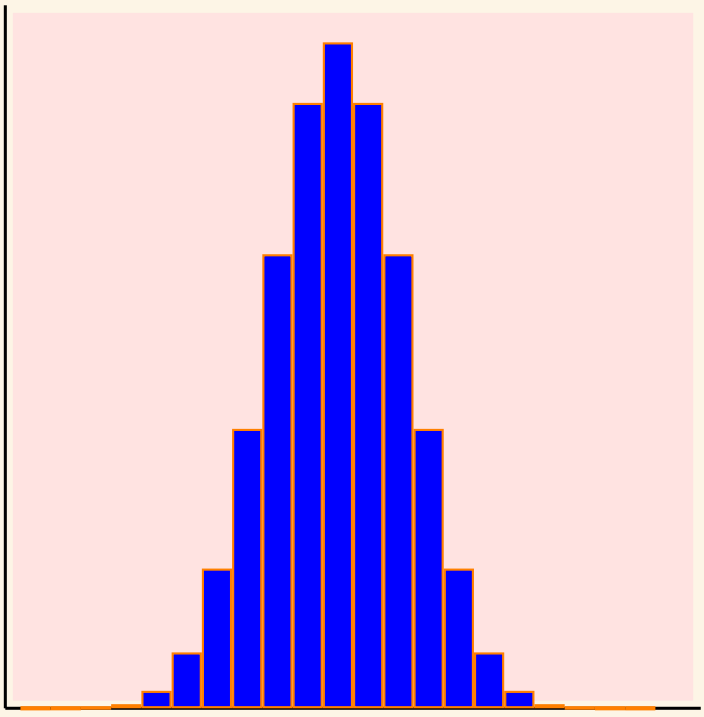


Figure 3: The histogram of a binomial variable with $n = 20$ and $p = 0.5$.

So since the histogram is bell shaped we are justified to apply the Empirical rule.
The Empirical Rule says that *approximately* 95% of values will have z-scores between -2 and 2. In our case we have $\mu = 10$ and $\sigma = 2.236$, so $2\sigma = 4.472$. So we get the raw scores

$$z = -2 \implies x = 10 - 4.472 = 5.528, \quad z = 2 \implies x = 10 + 4.472 = 14.472.$$

So for our distribution the Empirical rule says that about 95% of the values are between 5.528 and 14.472. We saw above that $p\,(5 < X < 15) = 0.9216$, in other words, 92.16% of values are between 5 and 15. So the prediction of the Empirical Rule is consistent with our results.

□

1.1 Exercises for you

1. A fair coin is tossed 30 times. Let X be the random variable that counts how many times the coin comes up tails. Compute the following probabilities:
   (a) $p(X = 10)$.
   (b) $p(X \leq 11)$.
   (c) $p(X > 16)$.
   (d) $p(10 < X < 18)$.

2. A coin has been fixed in such a way that instead of 50% the probability the it will come Heads up when tossed is 70%. That coin is tossed 30 times. Let X be the random variable that counts how many times the coin comes up tails. Compute the following probabilities:
   (a) $p(X = 10)$.
   (b) $p(X \leq 11)$.
   (c) $p(X > 16)$.
   (d) $p(10 < X < 18)$.

3. About 5% of all eggs are sold by a supermarket chain are cracked. Assume that you buy a dozen (that is 12) eggs from a store of that chain.
   (a) How many cracked eggs you expect to find?
   (b) What's the probability that you'll find 6 or less cracked eggs?
   (c) what's the probability that you'll find more than 6 cracked eggs?
   (d) what's the probability that the number r of cracked eggs is more than 1 but no more than 6 (i.e. $1 < X \leq 6$)?
   (e) What is the standard deviation of the number of cracked eggs?

4. Lebron James makes about 75% of the free throws he shoots. If he shoots 30 free throws, how many of those we *expect* to be successful?
   (a) What is the probability that he will make exactly 23 of them?
   (b) What is the probability that he will make between no less than 22 but no more than 24? (This is $p(22 \leq X \leq 24)$.)

5. When a tack is tossed the probability that it will come with the pointy end up is about 0.65. We toss a tack 8 times. Let X be the random variable that counts the number of times that it comes with the pointy end up.
   (a) Use the tables of the binomial distribution to construct a histogram for X.
   (b) What is the mean value of X?
   (c) What is the standard deviation of X?
   (d) What percentage of values has z–score, $-1 \leq z < 2$?

6. In June of 2017, Washington Post reported that according to an survey commissioned b Innovation Center of U.S., 7% of American adults believe that chocolate milk comes from brown cows. Our tables don't cover the $p = 0.07$ case, so use the tables for $p = 0.10$, to answer the questions below.

   In a randomly selected group of 25 American adults what is the probability that

(a) exactly 5 of them believe that chocolate milk comes from brown cows?

(b) more than 2 but less than 8 believe that chocolate milk comes from brown cows?

(c) What is the expected number of people that believe that chocolate milk comes from brown cows in that group?

7. The histograms of the binomial distribution for $n = 7$ with $p = .15, .25, .5, .75$ and $.85$ are shown in Figure 4, not necessarily in order. Determine what histogram corresponds to each p.

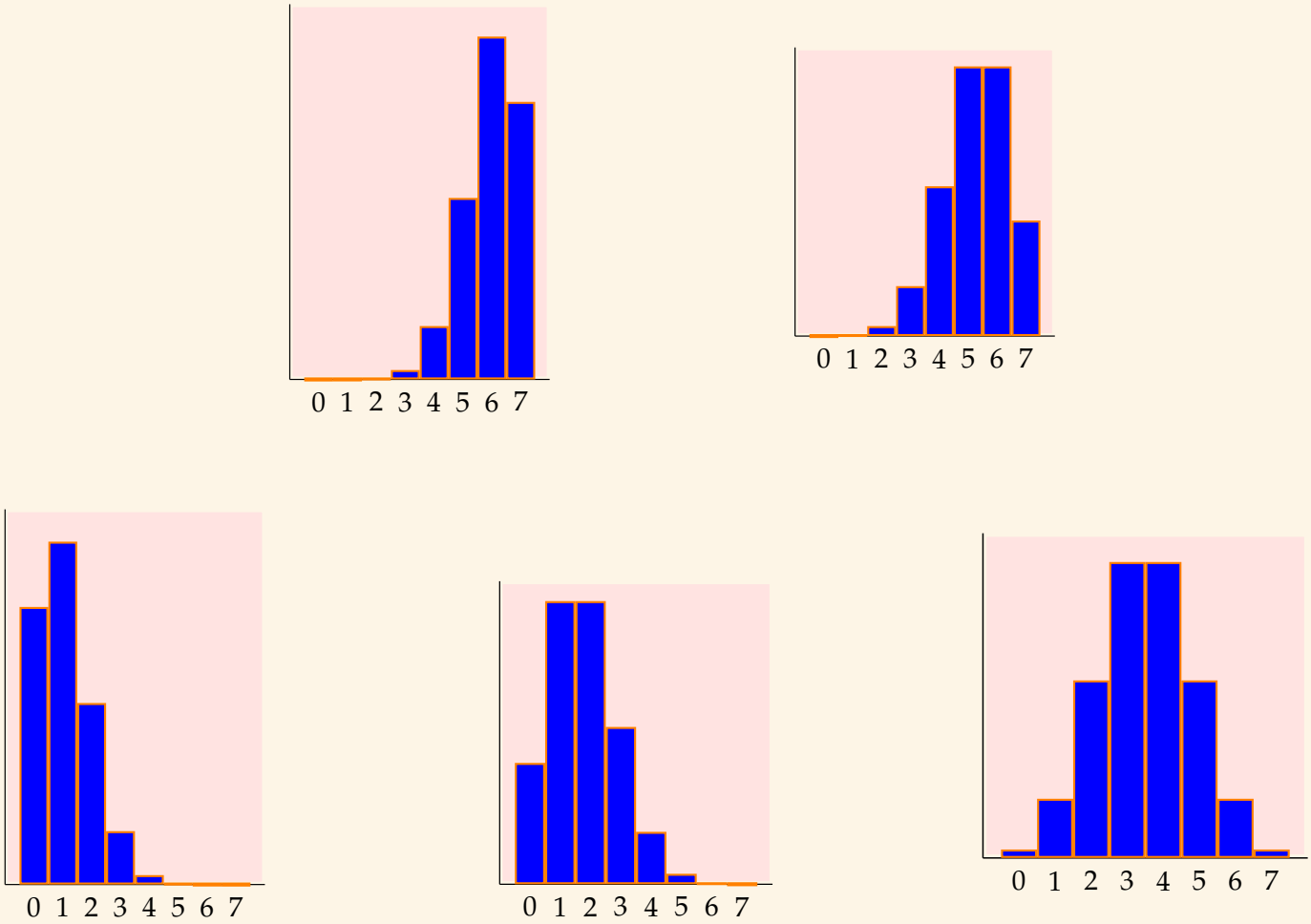


Figure 4: The histograms of five binomial distributions.

8. About 75% of an orange crop is good, the other 25% have rotten centers that can't be detected unless we cut the orange open. Oranges are sold in sacks of 10. Let X be the random variable that counts the number of good oranges in a sack.

(a) Make a histogram of the probability distribution of X.

(b) What is the probability that there is no more than one bad orange in a sack?

(c) What is the probability that there is at least one good orange in a sack?

(d) What is the expected number of good oranges in a sack?

(e) What is the standard deviation of X?

9. Alice and Bob play the following game. They toss a fair coin 10 times. If it comes ups Heads more than 3 but less than 7 times then Bob pays Alice 5 dollars. Otherwise Alice pays Bob 10 dollars. Determine how many dollars is Alice *expected* to win for one run of the game.