

Binomial Distribution

If X represents the number of successes in n independent Bernoulli trials (each with probability p of success), then the probability distribution function of X is the Binomial distribution function with parameters p and n .

$$P(X = x) = \binom{n}{x} [p^x (1-p)^{n-x}]$$

where

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

is the **binomial coefficient**

Note that $x! = x \cdot (x-1) \cdot (x-2) \cdots 1$ and $0! = 1$.

$P(X = x)$ is also written as $f_X(x)$ or $f_X(x; n, p)$.

Recall the experiment to select three individuals and record their smoking status. Let Y_i denote the smoking status of the i^{th} person ($i = 1, 2, 3$); assume the Y_i are independent.

Let X denote the number of smokers in the pair. Then $X = 0, 1, 2, 3$ are possible outcomes. X is Binomially distributed with population parameters $n = 3$ and $p = 0.3$. I use $p = 0.3$ here so that we can look up probabilities in the Appendix.

I simulated this experiment 10 times to get an idea about the average value of X .

Simulated Data:

Y_1	Y_2	Y_3	Number of Smokers
1	1	0	2
0	0	0	0
0	0	0	0
1	0	1	2
1	0	0	1
1	0	0	1
0	1	0	1
0	0	1	1
0	0	0	0
1	0	0	1

Simulated Data:

Y_1	Y_2	Y_3	Number of Smokers
1	1	0	2
0	0	0	0
0	0	0	0
1	0	1	2
1	0	0	1
1	0	0	1
0	1	0	1
0	0	1	1
0	0	0	0
1	0	0	1

Sample Mean:

$$\bar{x} = \frac{1}{10}(2+0+0+2+1+1+1+1+0+1) = 0.9$$

Another way to write this is

$$\begin{aligned} \bar{x} &= \frac{1}{10}(3 \cdot (0) + 5 \cdot (1) + 2 \cdot (2) + 0 \cdot (3)) \\ &= \frac{3}{10} \cdot (0) + \frac{5}{10} \cdot (1) + \frac{2}{10} \cdot (2) + \frac{0}{10} \cdot (3) \\ &= 0.3 \cdot (0) + 0.5 \cdot (1) + 0.2 \cdot (2) + 0.0 \cdot (3) \\ &\approx 0.343 \cdot (0) + 0.441 \cdot (1) + 0.189 \cdot (2) + 0.0270 \cdot (3) \\ &= f_X(x=0) \cdot (0) + f_X(x=1) \cdot (1) + f_X(x=2) \cdot (2) + f_X(x=3) \cdot (3) \end{aligned}$$

Summary measures

Given a random variable X and its probability distribution function, $f_X(x)$, we can find quantities that summarize the behavior of X . Such summaries include the expected value of X (population mean) and the population variance of X .

The **expected value** of a discrete random variable X is:

$$E[X] = \mu_X = \sum_{i=1}^N x_i f_X(x_i)$$

where the x_i 's are distinct values that the random variable can assume.

The expected value represents the average value of the random variable. It is obtained by multiplying each possible value x_i by its respective probability $f_X(x_i)$ and summing these products over all the values that the random variable can assume.

The **population variance** of a discrete random variable X is:

$$\text{var}[X] = \sigma_X^2 = \sum_{i=1}^N (x_i - \mu)^2 f_X(x_i)$$

where the x_i 's are defined as above.

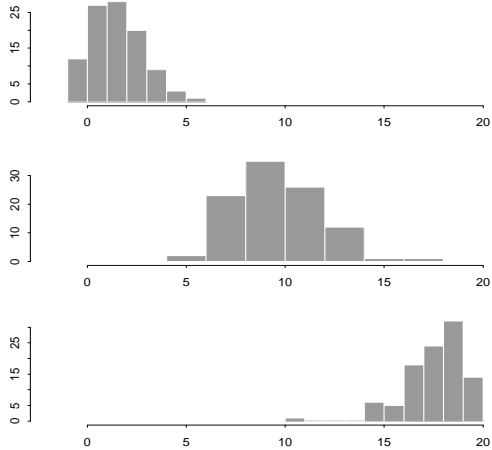
The variance represents the spread, relative to the expected value, of all values of the random variable that have positive probability. The **standard deviation**, σ , is the positive square root of the variance.

Binomial Distribution Summary Measures

If X is Binomially distributed with population parameters n and p , then

$$E[X] = np \quad \text{and} \quad \text{var}[X] = n \cdot p \cdot (1 - p)$$

Each histogram shows 100 simulations of a Binomial random variable X with population parameters $n = 20$ and $p = 0.1$ (upper), $p = 0.5$ (middle), and $p = 0.9$ (lower). Note that X can take on values $0, 1, 2, \dots, 20$



Note that, $f(k) = F(k) - F(k - 1)$, for discrete variables that take only integer values.

Another useful object is the **cumulative distribution function** (cdf), (or distribution function) which is usually denoted by $F_X(x)$. By definition $F_X(x) = P(X \leq x)$ where X refers to the random variable, and x refers to a specific value that X can assume.

$$F(x) = \sum_{y \leq x} f(y)$$

When is $F_X(x)$ useful ?

Suppose 20 individuals are treated surgically and there is a 60% chance of success for each surgery. What is the chance of a total of 3 or fewer unsuccessful surgeries?

$$P[3 \text{ or fewer failures}] = P[17 \text{ or more successes}] =$$

$$f(17) + f(18) + f(19) + f(20) = \sum_{k=17}^{20} \frac{n!}{(n-k)!k!} [p^k(1-p)^{n-k}]$$

You could also use the distribution function $F_X(k) = P[X \leq k]$.

$$P[17 \text{ or more successes}] = 1 - F(16)$$

The distribution function F is tabulated in some textbooks.

In a Binomial experiment, we count the number of successes in n trials. Each trial could be an experiment in one unit of time, an experiment in one person, an experiment in one unit of space. We know (or estimate) that an event happens with a given probability per trial (p) (per unit of time, per person, per unit of space) and the number of trials is known (n).

Examples:

1. An investigator is studying 20 family households where both parents smoke and notices that 3 out of the 20 households contain one or more children with chronic bronchitis. What is the probability of observing 3 such households, considering the prevalence rate of chronic bronchitis in children is 0.05? What is the probability of observing at least 3 such households?
2. A common lab test during a medical exam is a blood count. The two main aspects are 1. counting the number of white blood cells and 2. differentiating white blood cells into five categories. One of the categories is neutrophils. Suppose 60% of white blood cells are neutrophils in a healthy person. What is the probability of seeing 2 neutrophils in a sample of 5 white blood cells from a healthy person? What is the probability of seeing 2 or fewer?
3. Suppose it is known that a group of people has on average 0.25 cases of skin cancer per year. What is the probability that you will see 1 cancer case per year in a similar group?

The following are useful properties of Poisson distributions:

- The Poisson can be used to approximate the binomial when n is very large and p is very small. To do this, set $\lambda = np$; then

$$P[X = x] \approx \frac{e^{-\lambda} \lambda^x}{x!}.$$

- The sum of several Poisson random variables is also a Poisson random variable. Specifically, if Y_i , $i = 1, 2, \dots, m$, are independent Poisson random variables with respective parameters λ_i , $i = 1, 2, \dots, m$, then $Y_t = Y_1 + Y_2 + \dots + Y_m$ is a Poisson random variable with parameter $\lambda_t = \lambda_1 + \lambda_2 + \dots + \lambda_m$.

A Poisson random variable is often used to represent counts - the number of times an event occurs in some interval (time, space, etc...). The Poisson distribution is often associated with rare events.

If X is Poisson (representing the number of occurrences of some event in an interval of length t), then, $\Omega = \{0, 1, 2, \dots\}$, and the probability density function is given by,

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}.$$

λ represents the average rate of occurrence in an interval of length t . The parameter λ is both the mean and variance of the Poisson distribution.

The Poisson distribution is an appropriate model for counting events in an interval (say of length t) if the following three criteria are met (here, h is much smaller than t):

1. The probability that a single event occurs within a small interval of length h is proportional to h .
2. Within a small interval of length h , an infinite number of occurrences of the event are possible in principle. However, the probability that more than one event occurs is VERY small compared to the probability that a single event occurs.
3. The events occur independently between consecutive intervals.

Poisson Conditions in Practice

Consider the number of cancer cases observed in a year.

If we can assume the following, then the counts could be Poisson:

1. The probability of observing a case over a short time period is proportional to the length of the time period. i.e., a nurse who works for 2 days is twice as likely to see a case as a nurse who works for 1 day.
2. For a small interval of time (1 day), the probability of seeing two cases is negligible.
3. Observing a case in one interval (1 day in June) is independent of observing a case in a non-overlapping interval (a different day in June).

If these assumptions seem reasonable, you could use a Poisson distribution with λ (for the last example, being the average number of cases in a year).

To think about the last example (3) as a Binomial r.v., you had to think about breaking the year into n pieces ($n = 365$ days) and estimating the probability that a case is observed in one of the pieces (one day).

This seems much less intuitive.

Two things make the Poisson distribution seem appropriate or reasonable to describe the last example (3):

The events are assumed to be random and independent of one another (observing a skin cancer case one day does not affect the probability that you will observe a case on a different day).

Instead of the probability of the event, we know the average number of occurrences over some interval of measure.

Similarities: Binomial and Poisson Distributions

If X counts the number of successes in n independent Bernoulli trials (each with probability p of success), then the probability distribution function of X is the Binomial distribution function with parameters p and n .

$$P(X = x) = \binom{n}{x} [p^x (1-p)^{n-x}]$$

If X counts the number of occurrences of some event in an interval of length t (and X satisfies the conditions described last time) and λ represents the average rate of occurrence in the interval of length t , then X is a Poisson random variable with population parameter λ :

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}.$$

In a Binomial experiment, we count the number of successes in n trials. Each trial could be an experiment in one unit of time, an experiment in one person, an experiment in one unit of space.

We know (or estimate) that an event happens with a given probability per trial (p) (per unit of time, per person, per unit of space) and the number of trials is known (n).

In a Poisson experiment, we count the number of events in an interval (could be time, group of people, space). We know (or estimate) the average number of events in the interval (λ). We often don't know n .