Introduction to Descriptive Statistics and Probability for Data Science

- Sandeep Chaurasia

Random Variable

A random variable is a variable whose value is not known. It can either be discrete (having a specific value) or continuous (any value in a continuous range). All possible values that a random variable accepts is also called a sample space.

Binomial Random Variable: A binomial random variable is a number of successes in an experiment consisting of N trails. Some of the examples are:

- The number of successes (tails) in an experiment of 100 trials of tossing a coin. Here the sample space is {0, 1, 2, …100}
- The number of successes (four) in an experiment of 100 trials of rolling a dice. Here the sample space is {0, 1, 2, …100}

Binomial Distribution

Binomial distribution is a discrete probability distribution that represents the probabilities of binomial random variables. The binomial distribution is a probability distribution associated with a binomial experiment in which the binomial random variable specifies the number of successes or failures that occurred within that sample space.

- Example. Suppose you flipped a coin. The probability of getting heads or tails is equal. But what will be the probability of getting six heads in ten flips of coins? This is where you will need binomial distribution. You can calculate the probability of getting six heads in ten flips of a coin.

- The binomial distribution formula for any random variable X is given by

$$P(x, n, P) = nCx * P^x * (1 - P)^{n-x}$$

- Where, n = the number of experiments, x = 0, 1, 2, 3, 4, … (total number of successes), p = Probability of success in a single experiment.

Ex: Let's calculate the probability of getting exactly six heads when a coin is tossed ten times.

$$P(x=6) = 10C6 * 0.5^6 * 0.5^4 = ?$$

Mean and Variance of Binomial Distribution: The mean and variance of the binomial distribution are:

- Mean = np
- Variance = npq where,
- p is the probability of success
- q is the probability of failure (1-p)
- n is the number of trials.

Properties of a binomial distribution are:

1.There are only two possible outcomes: True or False, Yes or No.

2.There are N number of independent trials.

3.The probability of success and failure varies in each trial.

4.Only the number of successes are taken into account out of N independent trials.

Examples:

#1: 80% of people who purchase pet insurance are women. If 9 pet insurance owners are randomly selected, find the probability that exactly 6 are women.

#2: 60% of people who purchase sports cars are men. If 10 sports car owners are randomly selected, find the probability that exactly 7 are men.

Bernoulli Distribution

- A discrete probability distribution wherein the random variable can only have 2 possible outcomes is known as a Bernoulli Distribution. If in a Bernoulli trial the random variable takes on the value of 1, it means that this is a success.
- The probability of success is given by p. Similarly, if the value of the random variable is 0, it indicates failure. The probability of failure is q or 1 - p

$$f(x, p) = p^x (1 - p)^{1 - x}, x \in \{0, 1\}$$

Bernoulli Distribution	Binomial Distribtuion
Bernoulli distribution is used when we want to model the outcome of a single trial of an event.	If we want to model the outcome of multiple trials of an event, Binomial distribution is used.
It is represented as X ~ Bernoulli (p). Here, p is the probability of success.	It is denoted as X ~ Binomial (n, p). Where n is the number of trials.
Mean, E[X] = p	Mean, E[X] = np
Variance, Var[X] = p(1-p)	Variance, Var[X]= np(1-p)
Example: Suppose the probability of passing an exam is 80% and failing is 20%. Then the Bernoulli distribution can be used to model the passing or failing in such an exam.	Example: Suppose the probability of passing an exam is 80% and failing is 20%. Then if we want to find the probability that a student will pass in exactly 4 out of 5 exams, we use the Binomial Distribution.

The mean or average of a Bernoulli distribution is given by the formula
$E[X] = p$

To find the variance formula of a Bernoulli distribution we use $E[X^2] - (E[X])^2$ and apply properties.
Thus, Var[x] = p(1-p) of a Bernoulli distribution.

Bernoulli distribution is a case of binomial distribution when only 1 trial has been conducted. A binomial distribution is given by X ~~ Binomial (n, p). When n = 1, it becomes a Bernoulli distribution.

- Example 1: A basketball player can shoot a ball into the basket with a probability of 0.6. What is the probability that he misses the shot?

We know that success probability P (X = 1) = p = 0.6: Thus, probability of failure is P (X = 0) = 1 - p = 1 - 0.6 = 0.4

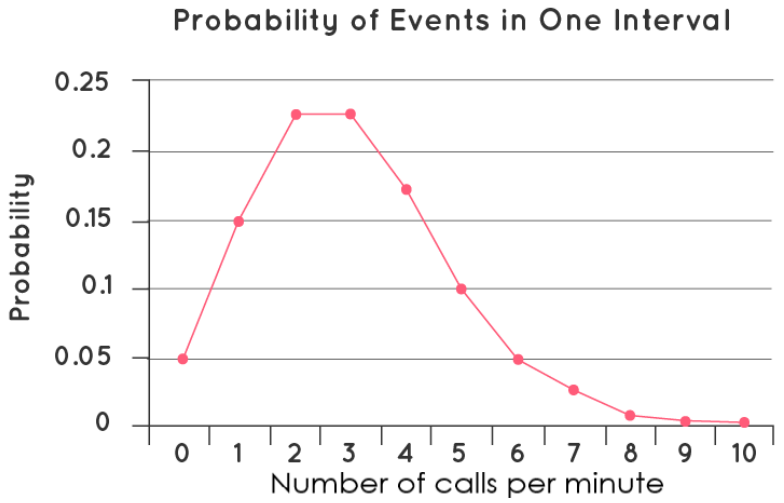
- Example 2: If a Bernoulli distribution has a parameter 0.45 then find its mean.
  Solution: X ∼∼ Bernoulli (p) or X ∼∼ Bernoulli (0.45).
  Mean E[X] = p = 0.45

- Example 3: If a Bernoulli distribution has a parameter 0.72 then find its variance.
  Solution: X ∼∼ Bernoulli (p) or X ∼∼ Bernoulli (0.72).
  Variance Var[X] = p (1-p) = 0.72 (0.28) = 0.2016

Poisson Distribution

- Poisson distribution is used to estimate how many times an event is likely to occur within the given period of time. λ is the Poisson rate parameter that indicates the expected value of the average number of events in the fixed time interval. Poisson distribution has wide use in the fields of business as well as in biology.

- Example, customer care center receives 100 calls per hour, 8 hours a day. As we can see that the calls are independent of each other. The probability of the number of calls per minute has a Poisson probability distribution. There can be any number of calls per minute irrespective of the number of calls received in the previous minute. Below is the curve of the probabilities for a fixed value of λ of a function following Poisson distribution:



For a random discrete variable X that follows the Poisson distribution, and λ is the average rate of value, then the probability of x is given by:

$$f(x) = P(X=x) = (e^{-\lambda} \lambda^{x})/x! \quad \text{, where}$$

x = 0, 1, 2, 3... ; e is the Euler's number(e = 2.718) & λ is an average rate of the expected value and λ = variance, also λ>0

For Poisson distribution, which has λ as the average rate, for a fixed interval of time, then the mean of the Poisson distribution and the value of variance will be the same. So, for X following Poisson distribution, we can say that λ is the mean as well as the variance of the distribution.
Hence: E(X) = V(X) = λ

Example 1: In a cafe, the customer arrives at a mean rate of 2 per min. Find the probability of arrival of 5 customers in 1 minute using the Poisson distribution formula.

Given: λ = 2, and x = 5.

Example 2: Find the mass probability of function at x = 6, if the value of the mean is 3.4.

Given: λ = 3.4, and x = 6.