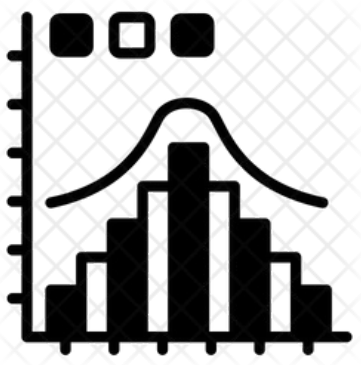
Probability Distributions

Tushar B. Kute,
http://tusharkute.com



Probability

- Probability is the chance of an outcome in an experiment (also called event).
  - Event: Tossing a fair coin
  - Outcome: Head, Tail

- Probability deals with predicting the likelihood of future events.
- Statistics involves the analysis of the frequency of past events

Probability

- Example: Consider there is a drawer containing 100 socks: 30 red, 20 blue and 50 black socks.
- We can use probability to answer questions about the selection of a random sample of these socks.
  - PQ1. What is the probability that we draw two blue socks or two red socks from the drawer?
  - PQ2. What is the probability that we pull out three socks or have matching pair?
  - PQ3. What is the probability that we draw five socks and they are all black?

Statistics

- Instead, if we have no knowledge about the type of socks in the drawers, then we enter into the realm of statistics. Statistics helps us to infer properties about the population on the basis of the random sample.
- Questions that would be statistical in nature are:
  - SQ1: A random sample of 10 socks from the drawer produced one blue, four red, five black socks. What is the total population of black, blue or red socks in the drawer?
  - SQ2: We randomly sample 10 socks, and write down the number of black socks and then return the socks to the drawer. The process is done for five times. The mean number of socks for each of these trial is 7. What is the true number of black socks in the drawer? etc.

Probability vs. Statistics

- In probability, we are given a model and asked what kind of data we are likely to see.
- In statistics, we are given data and asked what kind of model is likely to have generated it.
- Example: Measles Study
  - A study on health is concerned with the incidence of childhood measles in parents of childbearing age in a city. For each couple, we would like to know how likely, it is that either the mother or father or both have had childhood measles.
  - The current census data indicates that 20% adults between the ages 17 and 35 (regardless of sex) have had childhood measles.
  - This give us the probability that an individual in the city has had childhood measles.

Probability Distribution

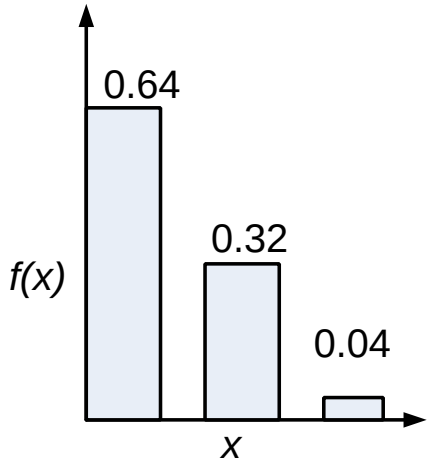
- A probability distribution is a definition of probabilities of the values of random variable.
- Given that is the probability that a person (in the ages between 17 and 35) has had childhood measles. Then the probability distribution is given by

X	Probability
0	0.64
1	0.32
2	0.04

Probability Distribution

- In data analytics, the probability distribution is important with which many statistics making inferences about population can be derived .
- In general, a probability distribution function takes the following form,

	0	**1**	**2**
	0.64	0.32	0.04



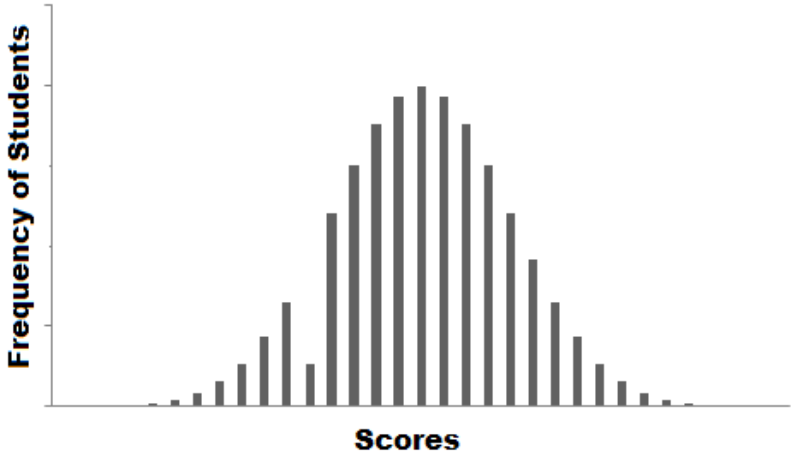
Probability Distribution

- Suppose you are a teacher at a university. After checking assignments for a week, you graded all the students.
- You gave these graded papers to a data entry guy in the university and tell him to create a spreadsheet containing the grades of all the students.
- But the guy only stores the grades and not the corresponding students.

S. No.	Scores
1	25
2	27
3	38
4	42
5	
6	16
7	35
8	46
9	48
10	31

Probability Distribution

- He made another blunder, he missed a couple of entries in a hurry and we have no idea whose grades are missing. Let's find a way to solve this.
- One way is that you visualize the grades and see if you can find a trend in the data.



Probability Distribution

- The graph that you have plot is called the frequency distribution of the data. You see that there is a smooth curve like structure that defines our data, but do you notice an anomaly?
- We have an abnormally low frequency at a particular score range. So the best guess would be to have missing values that remove the dent in the distribution.
- This is how you would try to solve a real-life problem using data analysis. For any Data Scientist, a student or a practitioner, distribution is a must know concept. It provides the basis for analytics and inferential statistics.

Use of Probability Distribution

- Distribution (discrete/continuous) function is widely used in simulation studies.
- A simulation study uses a computer to simulate a real phenomenon or process as closely as possible.
- The use of simulation studies can often eliminate the need of costly experiments and is also often used to study problems where actual experimentation is impossible.
- Examples:
  - A study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use such a drug approximately follows a binomial distribution.
  - Operation of ticketing system in a busy public establishment (e.g., airport), the arrival of passengers can be simulated using Poisson distribution.

Taxonomy of Probability Distributions

- **Discrete probability distributions**
  - Binomial distribution
  - Multinomial distribution
  - Poisson distribution
  - Hypergeometric distribution
- **Continuous probability distributions**
  - Normal distribution
  - Standard normal distribution
  - Gamma distribution
  - Exponential distribution
  - Chi square distribution
  - Lognormal distribution
  - Weibull distribution

Binomial Distribution

- In many situations, an outcome has only two outcomes: success and failure.
  - Such outcome is called dichotomous outcome.
- An experiment when consists of repeated trials, each with dichotomous outcome is called Bernoulli process. Each trial in it is called a Bernoulli trial.

Binomial Distribution

- Example: Firing bullets to hit a target.
  - Suppose, in a Bernoulli process, we define a random variable X number of successes in trials.
  - Such a random variable obeys the binomial probability distribution, if the experiment satisfies the following conditions:
    - The experiment consists of n trials.
    - Each trial results in one of two mutually exclusive outcomes, one labelled a “success” and the other a “failure”.
    - The probability of a success on a single trial is equal to . The value of remains constant throughout the experiment.
    - The trials are independent.

Defining Binomial Distribution

- The binomial random variable represents the number of successes(r) in n successive independent trials of a Bernoulli experiment.
- Probability of achieving r success and n-r failure is :

$$p^r * (1-p)^{n-r}$$

- The number of ways we can achieve r successes is :

$$\frac{n!}{(n-r)! * r!}$$

- Hence, the probability mass function(pmf), which is the total probability of achieving r success and n-r failure is :

$$\frac{n!}{(n-r)! * r!} * p^r * (1-p)^{n-r}$$

Defining Binomial Distribution

- The function for computing the probability for the binomial probability distribution is given by

$$for\ x = 0, 1, 2, \ldots., n$$

- Here, where denotes “the number of success” and denotes the number of success in trials.

Defining Binomial Distribution

- An example illustrating the distribution :
- Consider a random experiment of tossing a biased coin 6 times where the probability of getting a head is 0.6.
- If ‘getting a head’ is considered as ‘success’ then, the binomial distribution table will contain the probability of r successes for each possible value of r.

r	0	1	2	3	4	5	6
P(r)	0.004096	0.036864	0.138240	0.276480	0.311040	0.186624	0.046656

Example: Binomial Distribution

- Example: Measles study
  - X = having had childhood measles a success
  - p = 0.2, the probability that a parent had childhood measles
  - n = 2, here a couple is an experiment and an individual a trial, and the number of trials is two. Thus,

Example: Binomial Distribution

- Example : Verify with real-life experiment

  Suppose, 10 pairs of random numbers are generated by a computer (Monte-Carlo method)

  15 38 68 39 49 54 19 79 38 14
- If the value of the digit is 0 or 1, the outcome is “had childhood measles”, otherwise, (digits 2 to 9), the outcome is “did not”.
- For example, in the first pair (i.e., 15), representing a couple and for this couple, x = 1. The frequency distribution, for this sample is

x	0	1	2
$f(x)=P(X=x)$	0.7	0.3	0.0

- Note: This has close similarity with binomial probability distribution!

Binomial Distribution

- Suppose that you won the toss today and this indicates a successful event.
- You toss again but you lost this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow.
- Let's assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin.
- There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.

Binomial Distribution

- A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.
- The outcomes need not be equally likely. Remember the example of a fight between me and Undertaker?
- So, if the probability of success in an experiment is 0.2 then the probability of failure can be easily computed as q = 1 – 0.2 = 0.8.

Binomial Distribution

- Each trial is independent since the outcome of the previous toss doesn’t determine or affect the outcome of the current toss.
- An experiment with only two possible outcomes repeated n number of times is called binomial.
- The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

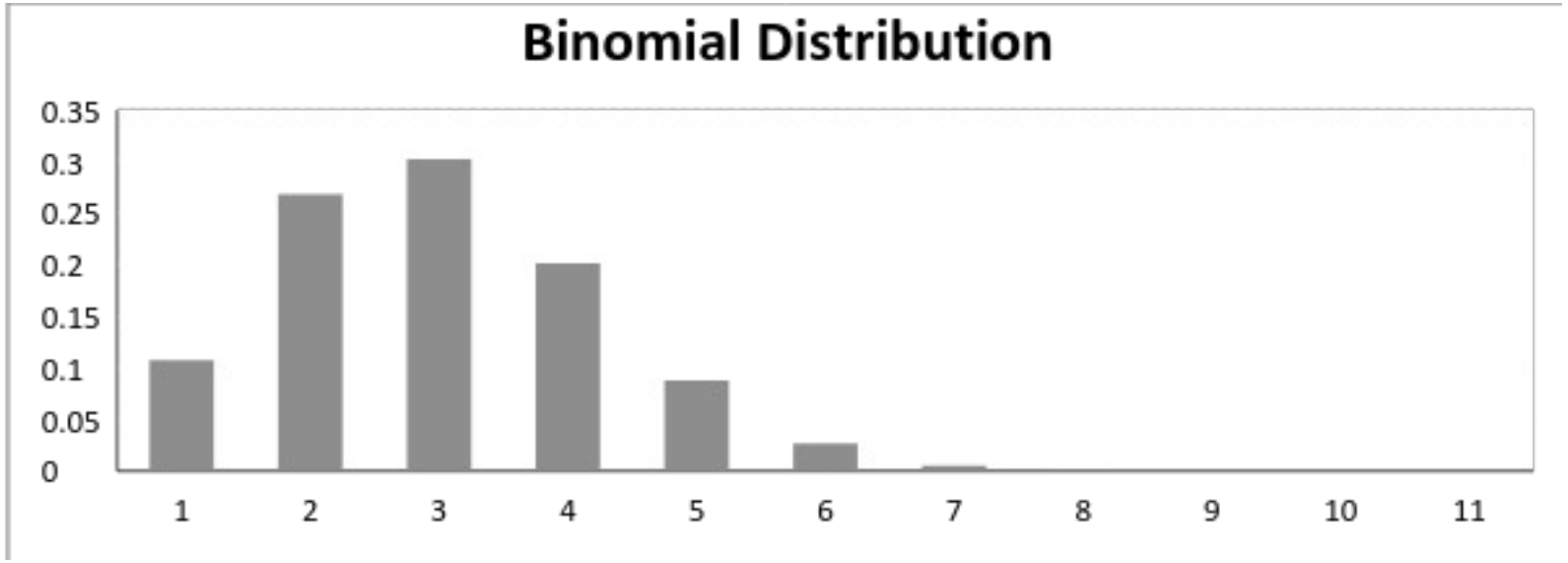
Binomial Distribution

- The properties of a Binomial Distribution are
  - Each trial is independent.
  - There are only two possible outcomes in a trial- either a success or a failure.
  - A total number of n identical trials are conducted.
  - The probability of success and failure is same for all trials. (Trials are identical.)
- The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

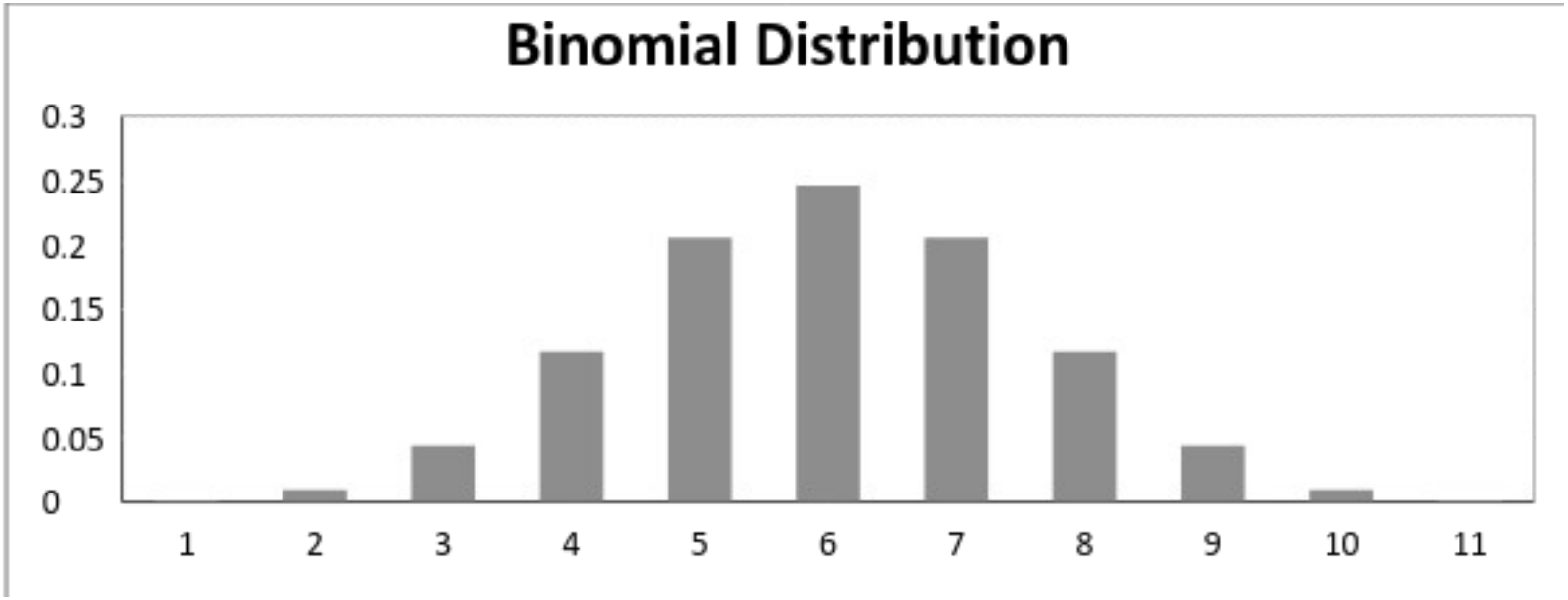
Binomial Distribution

- A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Binomial Distribution

- Now, when probability of success = probability of failure, in such a situation the graph of binomial distribution looks like,



Binomial Distribution

- The mean and variance of a binomial distribution are given by:

  Mean -> $\mu = n*p$

  Variance -> $Var(X) = n*p*q$

Binomial Distribution in Python

- You can generate a binomial distributed discrete random variable using scipy.stats module's binom.rvs() method which takes n (number of trials) and p (probability of success) as shape parameters.
- To shift distribution use the loc parameter. size decides the number of times to repeat the trials.
- If you want to maintain reproducibility, include a random_state argument assigned to a number.

Binomial Distribution

- The mean and variance of a binomial distribution are given by:

  Mean -> $\mu = n*p$

  Variance -> $Var(X) = n*p*q$

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com