Probability Distribution

Ms Swapnil Shrivastava
swapnil@cdac.in

Content

- Random Variables and Probability Distribution
- Discrete Probability Distribution
  - Bernoulli Distribution
  - Binomial Distribution
  - Poisson Distribution
- Continuous Probability Distribution
  - Uniform Distribution
  - Normal Distribution
  - Exponential Distribution

Random Variables

- A random variable is a function that assigns a real number to each element of a sample space.
- It is a numerical description of the outcome of an experiment.
- To have a consistent mathematical basis for dealing with probability.
- Random variable may be:
  - Discrete: is one for which the number of possible outcomes can be counted.
  - Continuous: has outcomes over one or more continuous intervals of real numbers.

Probability Distribution

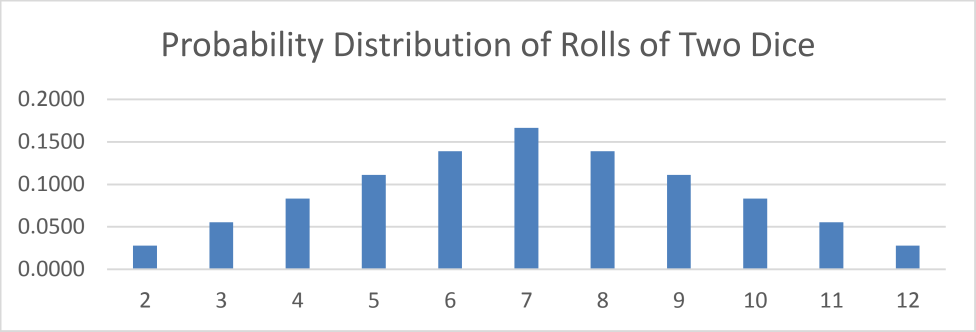
- It is a characterization of the possible values that a random variable may assume along with the probability of assuming these values.
- It can be either discrete or continuous, depending on the nature of the random variable it models.
- Can be developed using any of the three perspectives of probability.
  - Theoretical Probability Distribution
  - Empirical Probability Distribution
  - Subjective Probability Distribution

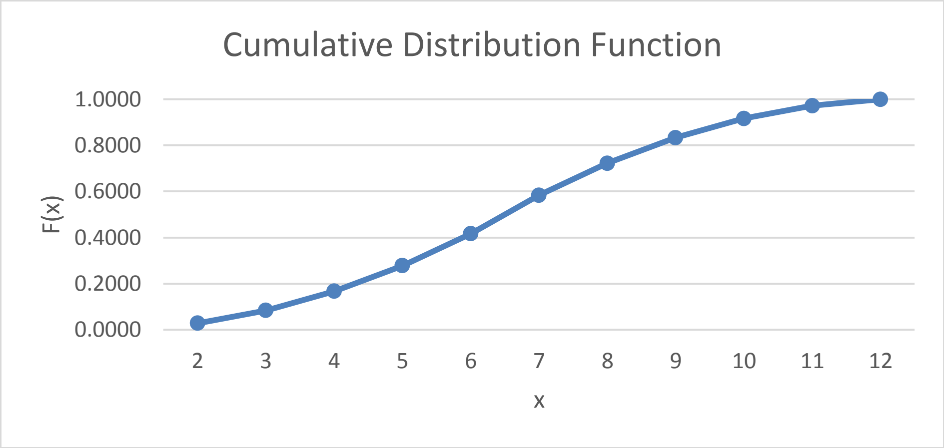
Discrete Probability Distributions

- Probability Mass Function f(x) represents probability distribution of the discrete outcomes for a discrete random variable X.
- If x_i represents the i^{th} value of X and $f(x_i)$ is the probability, the properties of f(x) are
  - $0 <= f(x_i) <= 1$ for all i
  - $\Sigma_i f(x_i) = 1$
- Cumulative Mass Function F(x) specifies the probability that the random variable X assumes a value less than or equal to a specified value, x.
  - $P(X<=x)$ is the probability that the random variable X is less than or equal to x.

Rolling Two Dice

Values of X	Outcomes	Probability Mass Function f(x)	Cumulative Distribution Function F(x)
x_1=2	1	1/36	1/36
x_2=3	2	1/18	1/12
x_3=4	3	1/12	1/6
x_4=5	4	2/18	5/18
x_5=6	5	5/36	5/12
x_6=7	6	1/6	7/12
x_7=8	5	5/36	13/18
x_8=9	4	2/18	10/12
x_9=10	3	1/12	11/12
x_{10}=11	2	1/18	35/36
x_{11}=12	1	1/36	1





Expected Value of X

- The expected value of a random variable corresponds to the notion of the mean, or average, for a sample.
- For a discrete random variable X, the expected value, denoted E[X], is weighted average of all possible outcomes.
- Expected value is long run average. It is appropriate for decisions that occur on a repeated basis.

Outcome, x	Probability, f(x)	x*f(x)
2	0.0278	0.0556
3	0.0556	0.1667
4	0.0833	0.3333
5	0.1111	0.5556
6	0.1389	0.8333
7	0.1667	1.1667
8	0.1389	1.1111
9	0.1111	1.0000
10	0.0833	0.8333
11	0.0556	0.6111
12	0.0278	0.3333
	Expected value	7.0000

Variance of X

- The variance Var[X] of a discrete random variable X as a weighted average of the squared deviations from the expected value.
- The variance measure the uncertainty of the random variable.
- The higher the variance, the higher the uncertainty of the outcome.

Outcome, x	Probability, f(x)	x*f(x)	(x - E[X])	(x - E[X])^2	(x - E[X])^2*f(x)
2	0.0278	0.0556	-5.0000	25.0000	0.6944
3	0.0556	0.1667	-4.0000	16.0000	0.8889
4	0.0833	0.3333	-3.0000	9.0000	0.7500
5	0.1111	0.5556	-2.0000	4.0000	0.4444
6	0.1389	0.8333	-1.0000	1.0000	0.1389
7	0.1667	1.1667	0.0000	0.0000	0.0000
8	0.1389	1.1111	1.0000	1.0000	0.1389
9	0.1111	1.0000	2.0000	4.0000	0.4444
10	0.0833	0.8333	3.0000	9.0000	0.7500
11	0.0556	0.6111	4.0000	16.0000	0.8889
12	0.0278	0.3333	5.0000	25.0000	0.6944
	Expected value	7.0000		Variance	5.8333

Bernoulli Distribution

- The Bernoulli distribution characterizes a random variable having two possible outcomes, each with a constant probability of occurrence.
- Typically these outcomes represent "success" (x=1), having probability p and "failure" (x=0), having probability 1-p.
- Example: Booting of a computer
- The probability mass function

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

where p represents the probability of success

- The expected value is p, and the variance is p(1-p)

Using the Bernoulli Distribution

A Bernoulli distribution could be used to model whether an individual responds positively (x=1) or negatively (x=0) to a telemarketing promotion.

For example, if we estimate that 20% of customers contacted will make a purchase, the probability distribution that describes whether or not a particular individual makes a purchase is Bernoulli with p=0.2.

Binomial Distribution

- It models n independent replications of a Bernoulli experiment, each with a probability p of success.
- The random variable X represents the number of successes in these n experiments.
- Using binomial distribution, we can calculate the probability that exactly x customers out of the n will make a purchase for any value of x between 0 and n.
- The expected value is np and the variance is $np(1-p)$.

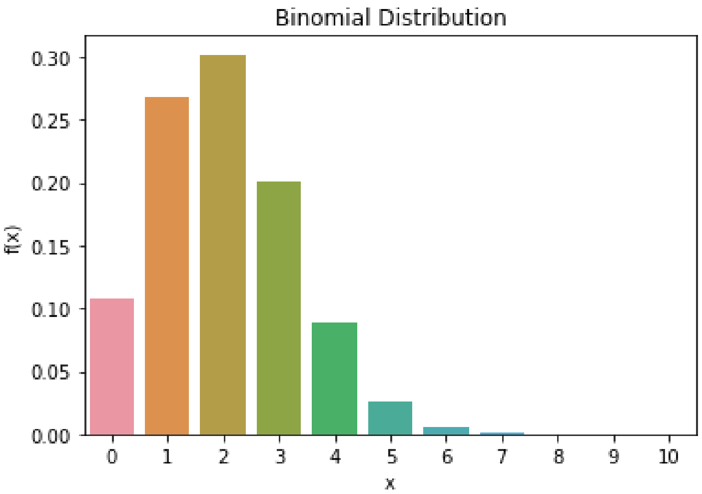
Binomial Distribution

In the telemarketing example, suppose that we call $n=10$ customers, each of which has a probability $p=0.2$ of making a purchase. Then the probability distribution of the number of positive responses obtained from 10 customers is binomial.

Using Binomial Distribution, we can calculate the probability that exactly x customers, maximum x customers or more than x customers out of 10 will make a purchase.

Binomial Distribution

```
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sbn

##probability that exactly 5 customers will make a purchase
print(stats.binom.pmf(5,10,0.2))

##probability that a maximum of 5 customers will make a purchase
print(stats.binom.cdf(5,10,0.2))

##probability that more than 5 customers will make a purchase

##Expected value and variance of binomial distribution
mean,var=stats.binom.stats(10,0.2)
print("Mean ",mean," Variance ",var)

#Binomial Distribution
bd_df=pd.DataFrame({"x" : range(0,11),"f(x)" :
list(stats.binom.pmf(range(0,11),10,0.2))})
sbn.barplot(x=bd_df["x"], y=bd_df["f(x)"])
plt.title("Binomial Distribution")
```



Poisson Distribution

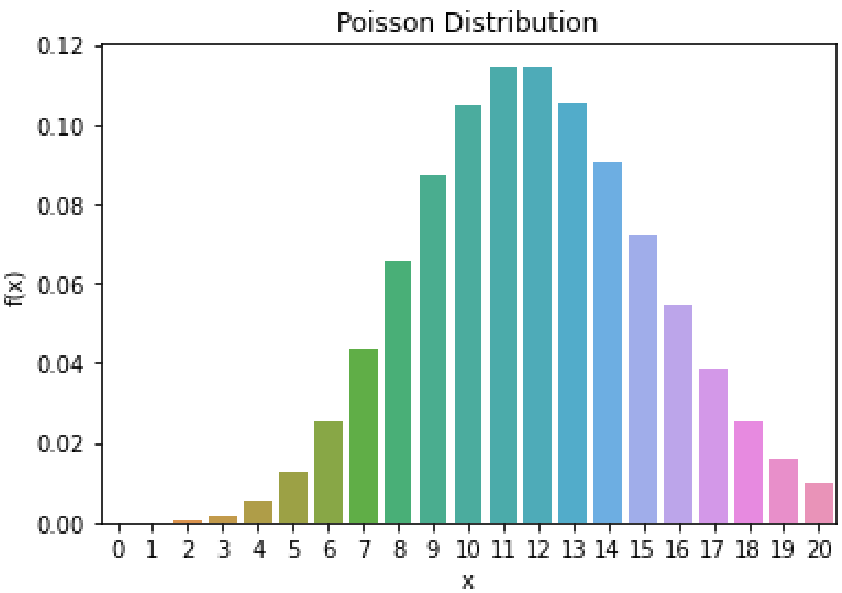
- It is a discrete distribution used to model the number of occurrences in some unit of measure.
  - Number of failures of a machine during a month
  - Number of customers arriving at a Subway store during a weekday lunch hour.
- The random variable X can assume any nonnegative integer value. The occurrences are independent. The constant ʎ is the average number of occurrences per unit.
- The expected value and variance is ʎ.

Using Poisson Distribution

- Suppose that on average the number of customers arriving at Subway during lunch hour is 12 customers per hour. The probability that exactly x customers will arrive during the hour is given by Poisson distribution with a mean of 12.
- Using Computing Poisson Probabilities Excel file

Poisson Distribution

```
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sbn

#probability that 6 customers will arrive during the hour
print(stats.poisson.pmf(6,12))

#probability that maximum 6 customers will arrive during the hour
print(stats.poisson.cdf(6,12))

#poisson distribution
pd_df=pd.DataFrame({"x" : range(0,21),"f(x)" : list(stats.poisson.pmf(range(0,21),12))})
sbn.barplot(x=pd_df["x"], y=pd_df["f(x)"])
plt.title("Poisson Distribution")
```



Continuous Probability Distribution

- A continuous random variable is defined over one or more intervals of real numbers. It has infinite number of possible outcomes.
- Probability Density Function f(x) is a curve that characterizes outcomes of a continuous random variable.
- Properties of f(x) are as follows:
  - F(x) >=0 for all values of x.
  - The total area under the density function above the x-axis is 1.0.
  - P(X=x) = 0 probability for a specific value of x doesn’t make sense.
  - P(a<= X<=b) is the area under the density function between a and b

Continuous Probability Distribution…contd

- Probabilities of continuous random variables are only defined over intervals.
  - $P(a<=X<=b)$: probabilities between two numbers a and b
  - $P(X<c)$ and $P(X>c)$: to the left or right of a number c
- Cumulative Distribution function F(x) represents the probability that the random variable X is less than or equal to x.
  - $F(x)= P(X <= x)$
- The probability that X is between a and b is equal to the difference of the cumulative distribution function evaluated at these two points:
  - $P(a <=X<=b) = P(X<=b) – P(X<=a)= F(b) – F(a)$

Uniform Distribution

- It characterizes a continuous random variable for which all outcomes between some minimum and maximum value are equally likely.
- Assumed when little is known about a random variable other than reasonable estimates for minimum and maximum values (a and b).
- Density function

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a <= x <= b \\ 0, & \text{otherwise} \end{cases}$$

Uniform Distribution

- Cumulative Distribution Function

$$F(x) = \begin{cases} 0, & \text{if } x<a \\ \frac{x-a}{b-a}, & \text{if } a<= x<=b \\ 1, & \text{if } b<x \end{cases}$$

- Expected Value

$$EV[X] = \frac{a+b}{2}$$

- Variance

$$Var[X] = \frac{(b-a)^2}{12}$$

Example

- Suppose that sales revenue, X, for a product varies uniformly each week between a=$1000 and b=$2000.
- Calculate
  - Density function
  - Probability that sales revenue would be less than x=$1,300
  - Probability that revenue will be between $1,500 and $1,700.

Python Code

```
# import uniform distribution
from scipy.stats import uniform

# random numbers from uniform distribution
n = 10000
start = 10
width = 20
data_uniform = uniform.rvs(size=n, loc = start,
scale=width)
```

Normal Distribution

- It is a continuous distribution that is described by the familiar bell shaped curve.
- The normal distribution is observed in many natural phenomena.
  - Human height and weight, test score
- Characterized by two parameters: mean(μ) and standard deviation(σ)
  - As μ changes the location of distribution on x-axis also changes
  - As σ is increased or decreased, the distribution becomes narrower or wider, respectively.
- Use stats.norm.cdf method for cumulative distribution of normal distribution.

Properties

- The distribution is symmetric, so its measure of skewness is zero.
- The mean, median and mode are all equal. Thus, half the area falls above the mean and half falls below it.
- The range of X is unbounded, meaning that the tails of the distribution extend to negative and positive infinity.
- The empirical rules apply for the normal distribution:
  - The area under the density function within +/- 1 standard deviation is 68.3%
  - The area under the density function within +/- 2 standard deviation is 95.4%
  - The area under the density function within +/- 3 standard deviation is 99.7%

Example

Suppose that a company has determined that the distribution of customer demand (X) is normal with a mean of 750 units/month and standard deviation of 100 units/month.

The company would like to know the following:

1. What is the probability that demand will be at most 900 units?
2. What is the probability that demand will exceed 700 units?
3. What is the probability that demand will be between 700 and 900 units?

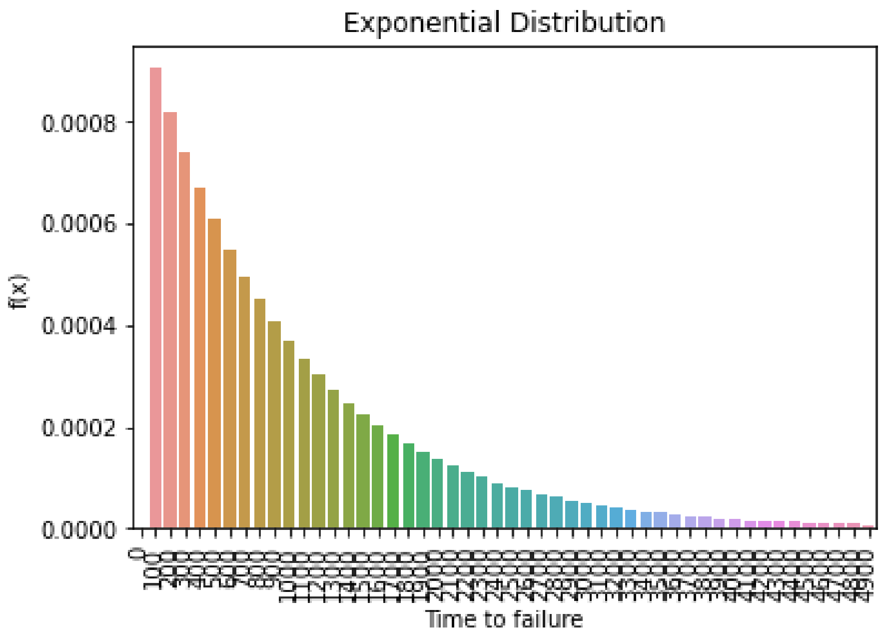
Exponential Distribution

- It is a continuous distribution that models the time between randomly occurring events.
- It is used in applications as
  - Modeling the time between customer arrivals to a service system.
- The exponential distribution has one parameter i.e. lambda.
- The Exponential distribution is closely related to the Poisson distribution.

Exponential Distribution

Suppose that the mean time to failure of a critical component of and engine is μ = 8,000 hours. Therefore λ = 1/μ = 1/8,000 failures/hour.

```
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sbn

#probability of failing before 5,000 hours
print(stats.expon.cdf(5000,loc=1/8000,scale=1000))

#exponential distribution
ed_df=pd.DataFrame({"x" : range(0,5000,100),"f(x)" :
list(stats.expon.pdf(range(0,5000,100),loc=1/8000,scale
=1000))})
sbn.barplot(x=ed_df["x"], y=ed_df["f(x)"])
plt.title("Exponential Distribution")
plt.xticks(rotation=90)
plt.xlabel("Time to failure")
```



Random Sampling from Probability Distribution

- Sampling from Discrete Probability Distributions
- Sampling from Common Probability Distributions
  - Uniform
  - Normal

Summary

- Probability quantifies the uncertainty that we encounter and is an important building block for business analytics applications.
- Many applications in business analytics require random samples from specific probability distribution.
- Prediction of probability of occurrence of an event, testing a hypothesis, building models to explain variation in key performance indicator.

References

- James Evans, Business Analytics: Methods, Models and Decisions, Second Edition, Pearson Publication, 2017.
- Manaranjan Pradhan and U Dinesh Kumar, Machine Learning using Python, Wiley Publication, 2019.
- U Dinesh Kumar, Business Analytics- The Science of Data-Driven Decision Making, Wiley Publication, 2017.
- Python pandas: https://pandas.pydata.org/
- Matplotlib: Visualization with Python https://matplotlib.org/