Probability Distributions

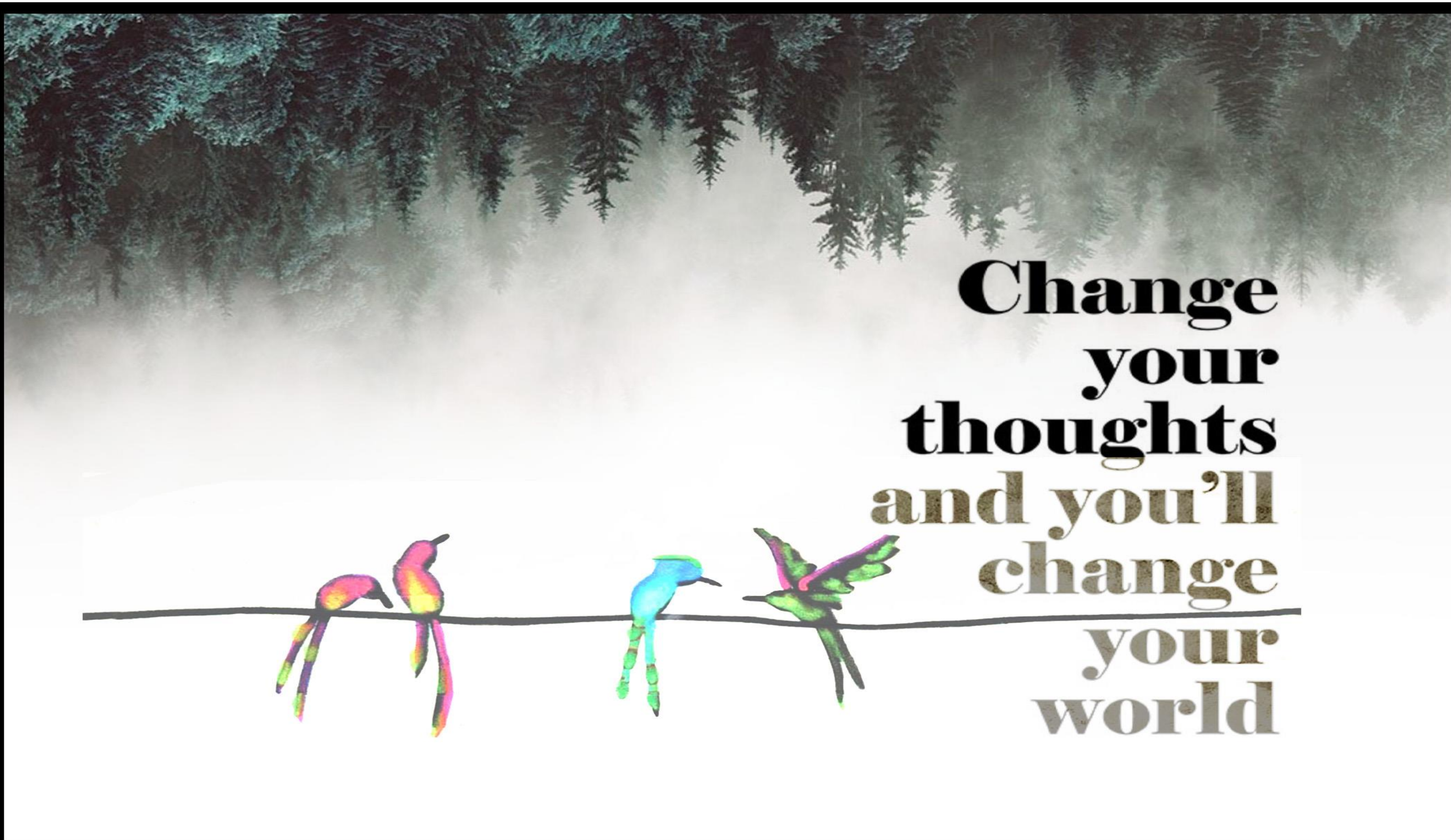
Course Taught at SUAD

Dr. Tanujit Chakraborty

Faculty @ Sorbonne

tanujitisi@gmail.com

Quote of the day..



NORMAN VINCENT PEALE, American - Clergyman

Today’s Topics…

- Probability vs. Statistics
- Concept of random variable
- Probability distribution concept

  - Discrete probability distribution
  - Continuous probability distribution

- Applications of Probability Distributions

Probability and Statistics

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

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:

- **Q1**: 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?
- **Q2**: 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 other words:

- 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.

Defining Random Variable

Definition: Random Variable

A random variable is a rule that assigns a numerical value to an outcome of interest.

Example : In "measles Study", we define a random variable X as the number of parents in a married couple who have had childhood measles.

This random variable can take values of 0, 1 and 2.

Note:

- Random variable is not exactly the same as the variable defining a data.
- The probability that the random variable takes a given value can be computed using the rules governing probability.
  - For example, the probability that $X = 1$ means either mother or father but not both has had measles is 0.32. Symbolically, it is denoted as **$P(X=1) = 0.32$.**

Probability Distribution

Definition : Probability distribution

A probability distribution is a definition of probabilities of the values of random variable.

Example : Given that 0.2 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

?

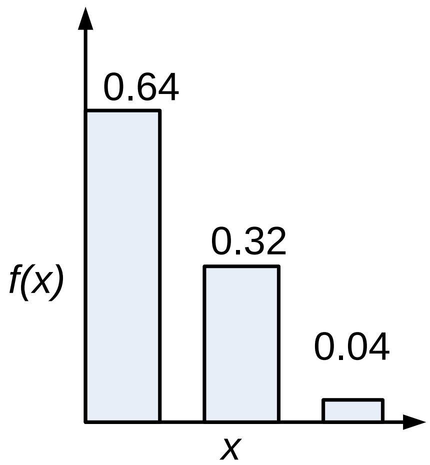
Probability Distribution

- 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

x	$x_1 \quad x_2 \ldots \ldots \ldots \ldots x_n$
$f(x) = P(X = x)$	$f(x_1) \quad f(x_2) \ldots \ldots f(x_n)$

Example: Measles Study

x	0	1	2
$f(x)$	0.64	0.32	0.04



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
- Lognormal distribution
- Weibull distribution

Usage 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:

1) 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.

2) Operation of ticketing system in a busy public establishment (e.g., airport), the arrival of passengers can be simulated using Poisson distribution.

Important note

- Probability mass function
  - A probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.
    - Sometimes it is also known as the discrete density function.

- Probability density function
  - A probability density function (PDF) in that is associated with continuous rather than discrete random variables.

- Note:
  - A PDF must be integrated over an interval to yield a probability.

Discrete Probability Distributions

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.

Example : Firing bullets to hit a target.

- Suppose, in a Bernoulli process, we define a random variable $X \equiv$ the number of successes in trials.
- Such a random variable obeys the binomial probability distribution, if the experiment satisfies the following conditions:

1) The experiment consists of n trials.
2) Each trial results in one of two mutually exclusive outcomes, one labelled a *"success"* and the other a *"failure"*.
3) The probability of a success on a single trial is equal to $\boldsymbol{p}$. The value of p remains constant throughout the experiment.
4) The trials are independent.

Defining Binomial Distribution

Definition: **Binomial distribution**

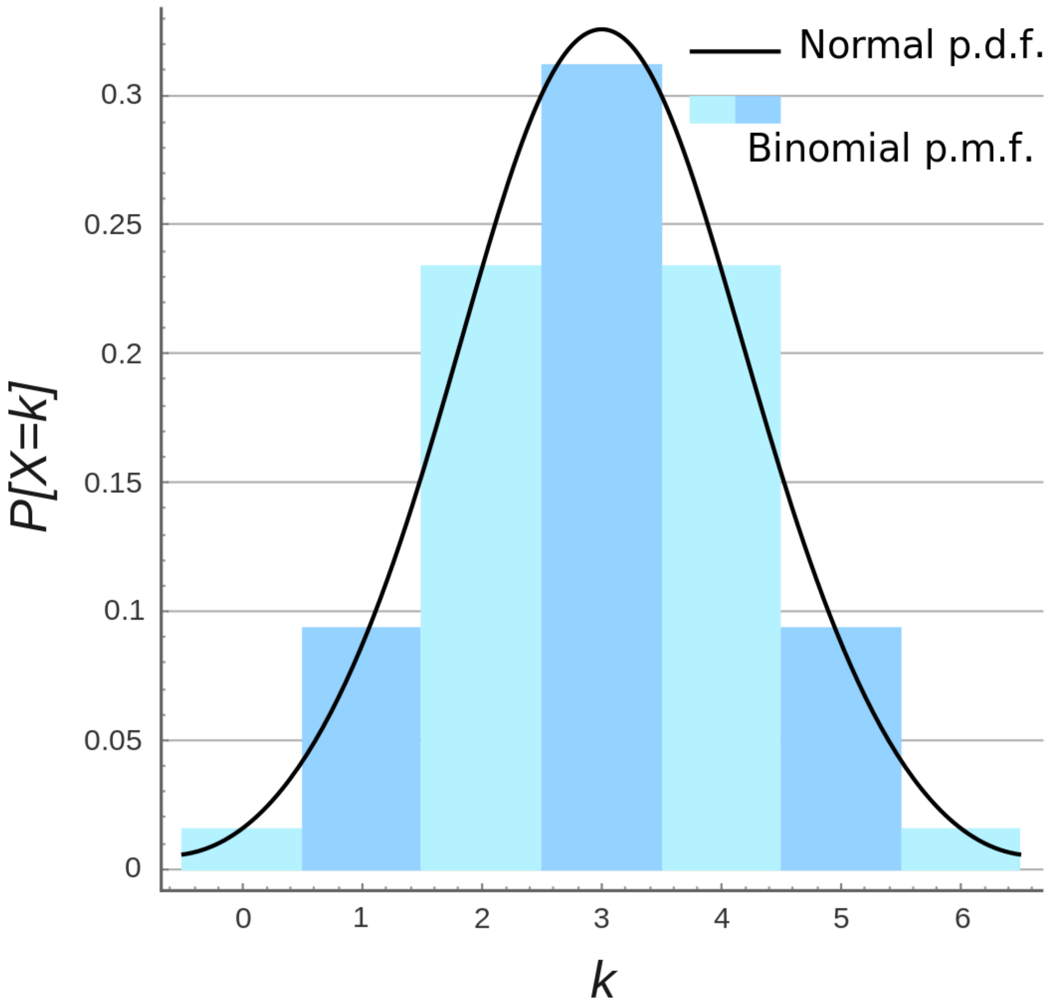
The function for computing the probability for the binomial probability distribution is given by

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

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

Here, $f(x) = P(X = x)$, where X denotes "the number of success" and $X = x$ denotes the number of success in n trials.

Binomial Distribution Curves



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 in a trial, and the number of trials is two.

Thus,

$$P(x=0)=\frac{2!}{0!(2-0)!}(0.2)^0(0.8)^{2-0} \quad =\mathbf{0.64}$$

$$P(x=1)=\frac{2!}{1!\,(2-1)!}(0.2)^1(0.8)^{2-1}=\mathbf{0.32}$$

$$P(x=2)=\frac{2!}{2!\,(2-2)!}(0.2)^2(0.8)^{2-2}=\mathbf{0.04}$$

X	Probability
0	0.64
1	0.32
2	0.04

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!

Exercise:

- The ***Los Angeles Times* (December 13, 1992)** reported that what airline passengers like to do most on long flights is rest or sleep; in a survey of 3697 passengers, almost 80% did so. Suppose that for a particular route the actual percentage is exactly 80%, and consider randomly selecting six passengers.

a. Calculate $p(4)$, and interpret this probability.

b. Calculate $p(6)$, the probability that all six selected passengers rested or slept.

c. Determine $P(X \geq 4)$.

Multinomial Distribution

The binomial experiment becomes a multinomial experiment, if we let each trial has more than two possible outcome.

Definition: **Multinomial distribution**

If a given trial can result in the $\boldsymbol{k}$ outcomes $E_1, E_2, \ldots \ldots, E_k$ with probabilities $p_1, p_2, \ldots \ldots, p_k$, then the probability distribution of the random variables $X_1, X_2, \ldots \ldots, X_k$ representing the number of occurrences for $E_1, E_2, \ldots \ldots, E_k$ in n independent trials is

$$f(x_1, x_2, \ldots \ldots, x_k) = \binom{n}{x_1, x_2, \ldots \ldots, x_k} {p_1}^{x_1} {p_2}^{x_2} \ldots \ldots {p_k}^{x_k}$$

where $\binom{n}{x_1, x_2, \ldots \ldots, x_k} = \frac{n!}{x_1! x_2! \ldots \ldots x_k!}$

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1$$

Hypergeometric Distribution

- Collection of samples with two strategies
  - With replacement
  - Without replacement
- A necessary condition of the binomial distribution is that all trials are **independent** to each other.
  - When sample is collected "with replacement", then each trial in sample collection is independent.

Example:

Probability of observing three red cards in 5 draws from an ordinary deck of 52 playing cards.

– You draw one card, note the result and then returned to the deck of cards

– Reshuffled the deck well before the next drawing is made

- The hypergeometric distribution *does not require* independence and is based on the sampling done **without** replacement.

Hypergeometric Distribution

- In general, the hypergeometric probability distribution enables us to find the probability of selecting x successes in n trials from N items.

Properties of Hypergeometric Distribution

- A random sample of size n is selected without replacement from N items.
- k of the N items may be classified as success and $N-k$ items are classified as failure.

Let X denotes a hypergeometric random variable defining the number of successes.

Definition: Hypergeometric Probability Distribution

The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labelled success and $N-k$ labelled as failure is given by

$$f(x) = P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

$$\max(0, n-(N-k)) \leq x \leq \min(n, k)$$

Multivariate Hypergeometric Distribution

The hypergeometric distribution can be extended to treat the case where the N items can be divided into k classes $A_1, A_2, \dots \dots, A_k$ with a_1 elements in the first class A_1, … and a_k elements in the k^{th} class. We are now interested in the probability that a random sample of size n yields x_1 elements from A_1, x_2 elements from $A_2, \dots \dots, x_k$ elements from A_k.

Definition: Multivariate Hypergeometric Distribution

If N items are partitioned into k classes $a_1, a_2, \dots \dots, a_k$ respectively, then the probability distribution of the random variables $X_1, X_2, \dots \dots, X_k$, representing the number of elements selected from $A_1, A_2, \dots \dots, A_k$ in a random sample of size n, is

$$f(x_1, x_2, \dots \dots, x_k) = P(X_1 = x_1, X_2 = x_2, \dots \dots, X_k = x_k) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2} \dots \dots \binom{a_k}{x_k}}{\binom{N}{n}}$$

with $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} a_i = N$

Poisson Process

There are some experiments, which involve the occurring of the number of outcomes during a given time interval (or in a region of space).

Such a process is called Poisson process.

The Poisson process is one of the most widely-used counting processes. It is usually used in scenarios where we are counting the occurrences of certain events that appear to happen at a certain rate, but completely at random

Example : Number of clients visiting a ticket selling counter in VOX Cinemas.

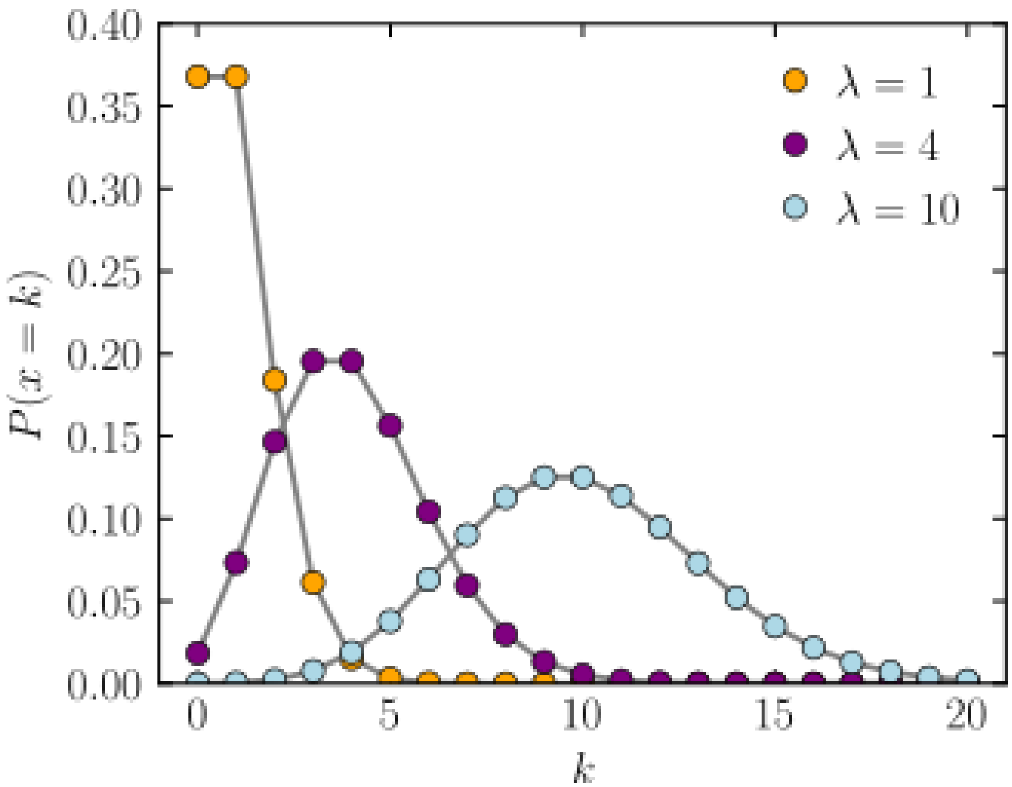
Poisson Process

Properties of Poisson process

- There is a discrete value, say x is the number of times an event occurs in an interval and x can take values 0, 1, 2, ...
- The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
- The average rate at which events occur assumed to be constant.
- Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur.
- Poisson process has no memory

If these conditions are true, then x is a Poisson random variable, and the distribution of x is a Poisson distribution.

Poisson Distribution Curves



Poisson Distribution

Definition: Poisson distribution

The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval t, is

$$f(x, \lambda t) = P(X = x) = \frac{e^{-\lambda t}.(\lambda t)^x}{x!}, x = 0, 1, \dots \dots$$

where λ is the average number of outcomes per unit time and $e = 2.71828 \dots$

What is $P(X = x)$ if $t = 0$?

Example:

- The number of customers arriving at a grocery store can be modelled by a Poisson process with intensity λ=10 customers per hour.
    1. Find the probability that there are 2 customers between 10:00 and 10:20.
    2. Find the probability that there are 3 customers between 10:00 and 10:20 and 7 customers between 10:20 and 11:00.

Examples

Suppose that the number of telephone calls coming into a telephone exchange between 10 A.M and 11 A.M. say, X_1 is a random variable with Poisson distribution with parameter 2. Similarly the number of calls arriving between 11 A.M. to 12 noon, say, X_2 has a Poisson distribution with parameter 6. If X_1 and X_2 are independent, what is the probability that more than 5 calls come in-between 10 A.M. and 12 noon?

Descriptive measures

Given a random variable X in an experiment, we have denoted $f(x) = P(X = x)$, the probability that $X = x$. For discrete events $f(x) = 0$ for all values of x except $x = 0, 1, 2, \ldots..$

Properties of discrete probability distribution

1. $0 \leq f(x) \leq 1$
2. $\sum f(x) = 1$
3. $\mu = \sum x . f(x)$ [is the mean]
4. $\sigma^2 = \sum (x - \mu)^2 . f(x)$ [is the variance]

In $2, 3$ *and* 4, summation is extended for all possible discrete values of x.

Note: For discrete uniform distribution, $f(x) = \frac{1}{n}$ with $x = 1, 2, \ldots \ldots, n$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

Descriptive Measures

1. Binomial distribution

The binomial probability distribution is characterized with p (the probability of success) and n (is the number of trials). Then

$$\mu = np\ ; \quad \sigma^2 = np(1-p)$$

2. Poisson Distribution

The Poisson distribution is characterized with λ where $\lambda = the\ mean\ of\ outcomes$ and $t = time\ interval.$

$$\mu = \lambda t\ ; \quad \sigma^2 = \lambda t$$

3. Hypergeometric distribution

The hypergeometric distribution function is characterized with the size of a sample (n), the number of items (N) and k labelled success. Then

$$\mu = \frac{nk}{N}$$

$$\sigma^2 = \frac{N-n}{N-1}.\mathrm{n}.\frac{k}{N}(1-\frac{k}{N})$$

Special Case: Discrete Uniform Distribution

- **Discrete uniform distribution**

  A random variable X has a discrete uniform distribution if each of the n values in the range, say $x_1, x_2, x_3, \ldots, x_n$ has equal probability. That is

$$f(x) = \frac{1}{n}$$

  Where $f(x)$ represents the probability mass function.

- **Mean and variance for discrete uniform distribution**

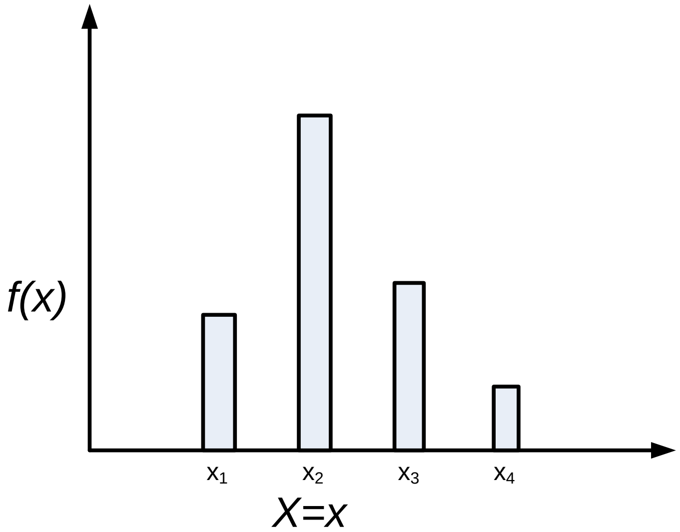
  Suppose, X is a discrete uniform random variable in the range $[a,b]$, *such that* $a \leq b$, then

$$\mu = \frac{b+a}{2}$$

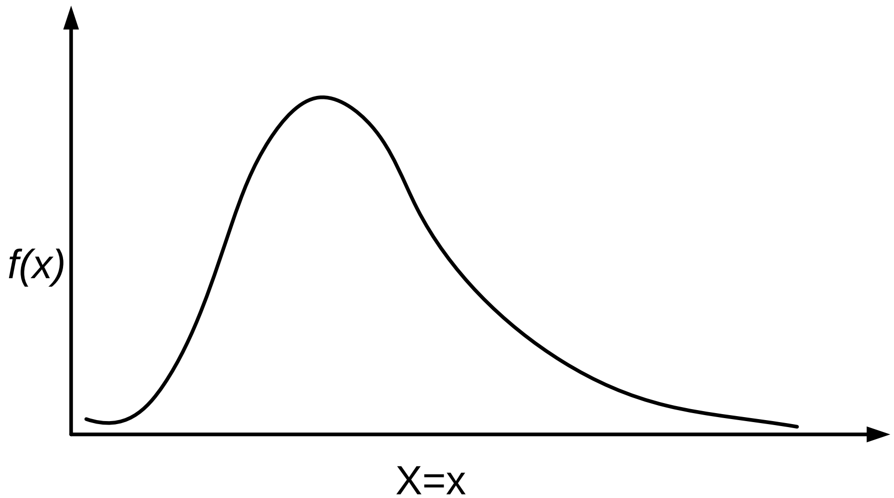
$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$$

Continuous Probability Distributions

Discrete Vs. Continuous Probability Distributions



Discrete Probability distribution



Continuous Probability Distribution

Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called continuous random variable.
  - Every continuous random variable has an infinite, uncountable number of possible values (i.e., any value in an interval)
- Consequently, continuous random variable differs from discrete random variable.

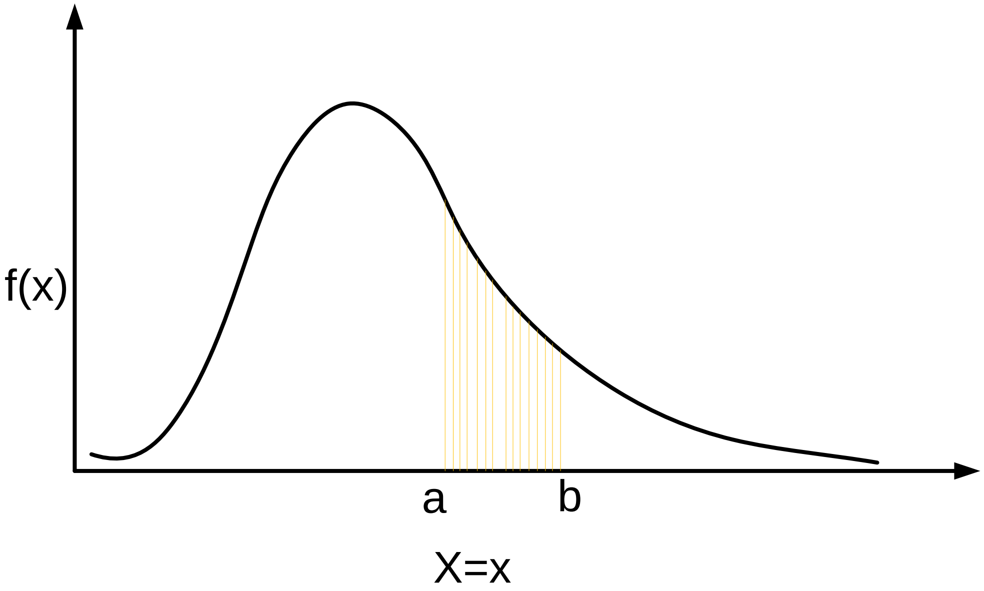
Examples:

1. Tax to be paid for a purchase in a shopping mall. Here, the random variable varies from 0 to $+\infty$.
2. Amount of rainfall in *mm* in a region.
3. Earthquake intensity in Richter scale.
4. Height of an earth surface. Here, the random variable varies from $-a$ to $+b$, $[a, b] \in R$, R is a set of real numbers.

Properties of Probability Density Function

The function $f(x)$ is a probability density function for the continuous random variable X, defined over the set of real numbers R, if

1. $f(x) \geq 0$, for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f(x)\,dx$
4. $\mu = \int_{-\infty}^{\infty} xf(x)\,dx$
5. $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$



Note: Probability is represented by area under the curve. The probability of a specific value of a continuous random variable will be zero because the area under a point is zero.

Continuous Probability Distributions

- **Example:**

  - Suppose bacteria of a certain species typically live 4 to 6 hours. The probability that a bacterium lives *exactly* 5 hours is equal to zero. A lot of bacteria live for approximately 5 hours, but there is no chance that any given bacterium dies at exactly 5.0000000000... hours.

  - However, the probability that the bacterium dies between 5 hours and 5.01 hours is quantifiable.

  - Suppose, the answer is 0.02 (i.e., 2%). Then, the probability that the bacterium dies between 5 hours and 5.001 hours should be about 0.002, since this time interval is one-tenth as long as the previous. The probability that the bacterium dies between 5 hours and 5.0001 hours should be about 0.0002, and so on.

Continuous Probability Distributions

- **Note:**

  - In these three examples, the ratio (probability of dying during an interval) / (duration of the interval) is approximately constant, and equal to 2 per hour (or 2 $hour^{-1}$). For example, there is 0.02 probability of dying in the 0.01-hour interval between 5 and 5.01 hours, and (0.02 probability / 0.01 hours) = 2 $hour^{-1}$. This quantity 2 $hour^{-1}$ is called the probability density for dying at around 5 hours.

  - Therefore, the probability that the bacterium dies at 5 hours can be written as (2 $hour^{-1}$) dt. This is the probability that the bacterium dies within an infinitesimal window of time around 5 hours, where dt is the duration of this window.

Continuous Uniform Distribution

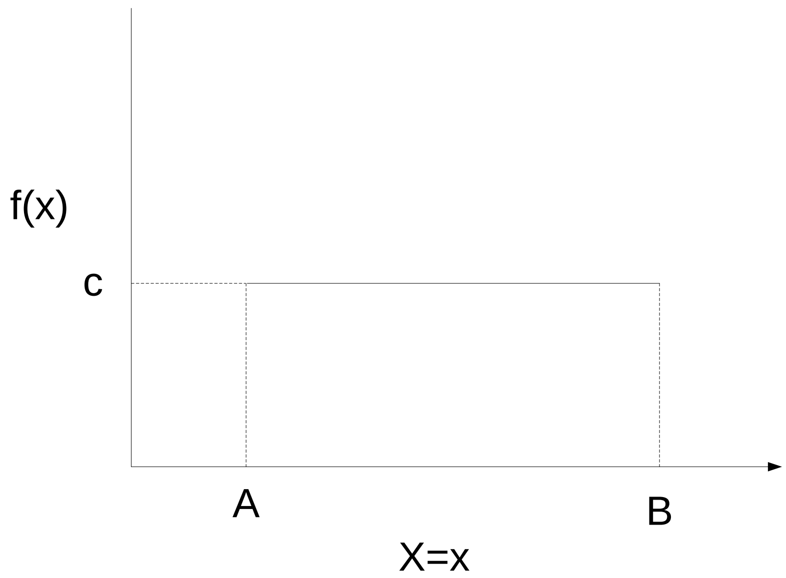
- One of the simplest continuous distribution in all of statistics is the continuous uniform distribution.

Definition : **Continuous Uniform Distribution**

The density function of the continuous uniform random variable X on the interval $[A, B]$ is:

$$f(x : A, B) = \begin{cases} \dfrac{1}{B - A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$$

Continuous Uniform Distribution



Note:

a) $\int f(x)dx = \frac{1}{B-A} \times (B-A) = 1$

b) $P(c < x < d) = \frac{d-c}{B-A}$ where both c and d are in the interval (A,B)

c) $\mu = \frac{A+B}{2}$

d) $\sigma^2 = \frac{(B-A)^2}{12}$

Example

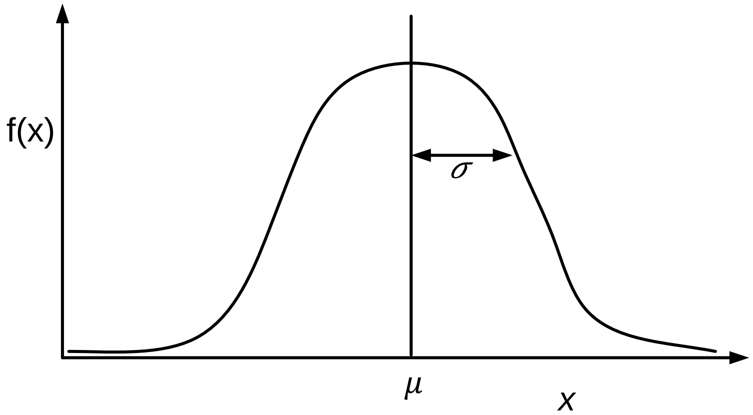
Suppose a train arrives at a subway station regularly every 20 min. If a passenger arrives at the station without knowing the timetable, then find the probability that the man will have to wait at least 10 min? What is the average waiting time ?

Normal Distribution

- The most often used continuous probability distribution is the normal distribution; it is also known as **Gaussian** distribution.
- It's graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
  - Physical measurement in areas such as meteorological experiments, rainfall studies and measurement of manufacturing parts are often more than adequately explained with normal distribution.
- A continuous random variable X having the bell-shaped distribution is called a normal random variable.

Normal Distribution

- The mathematical equation for the probability distribution of the normal variable depends upon the two parameters μ and σ, its mean and standard deviation.



Definition : Normal distribution

The density of the normal variable x with mean μ and variance σ^2 is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where $\pi = 3.14159 \ldots$ and $e = 2.71828 \ldots$, the Naperian constant

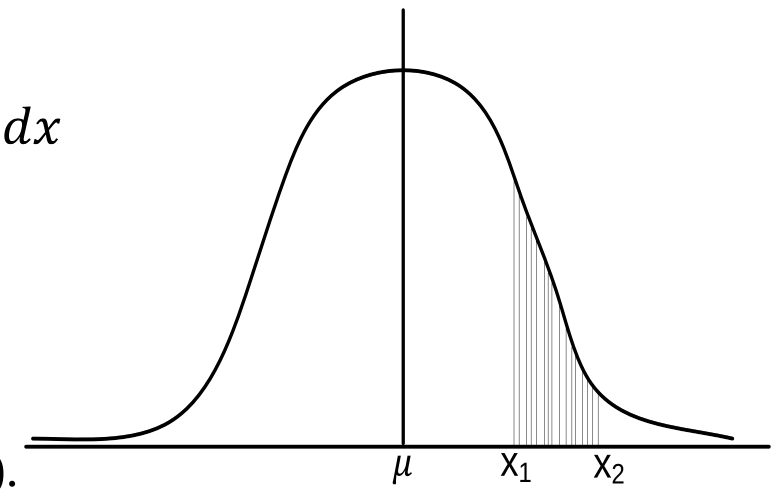
Properties of Normal Distribution

- The curve is symmetric about a vertical axis through the mean μ.
- The random variable x can take any value from $-\infty\ to\ \infty$.
- The most frequently used descriptive parameters define the curve itself.
- The mode, which is the point on the horizontal axis where the curve is a maximum occurs at $x = \mu$.
- The total area under the curve and above the horizontal axis is equal to 1.

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1$$

- $\mu = \int_{-\infty}^{\infty} x.f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x.e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$
- $\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 . e^{-\frac{1}{2}[(x-\mu)/\sigma^2]} dx$
- $P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

denotes the probability of x in the interval (x_1, x_2).

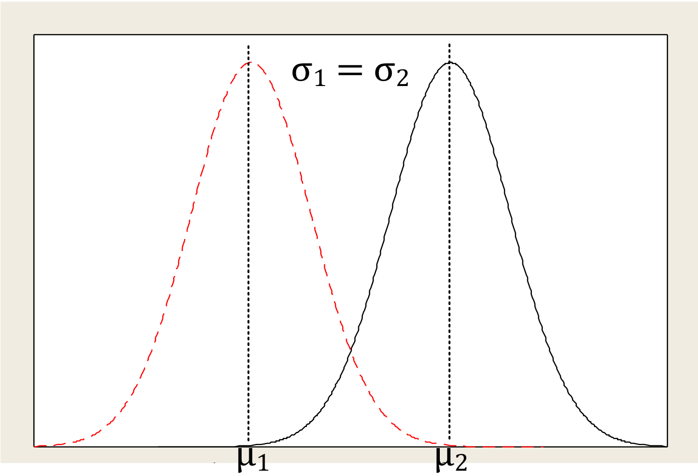


Examples

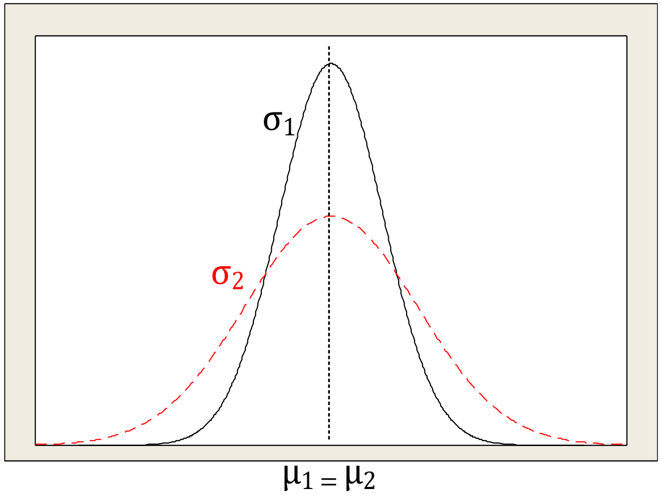
- There are 600 data science students in the under graduate classes of a university, and the probability for any student to need a copy of a particular book from the university library on any day is 0.05. How many copies of the book should be kept in the university library so that the probability may be grater than 0.90 that none of the students needing a copy from the library has to come back disappointed?

- Emissions of nitrogen oxides, which are major constituents of smog, can be modelled using a normal distribution. Let X denote the amount of this pollutant emitted by a randomly selected vehicle (in parts per billion). The distribution of X can be described by a normal distribution with mean 1.6 and standard deviation 0.4. Suppose that the EPA wants to offer some sort of incentive to get the worst polluters off the road. What emission levels constitute the worst 10% of the vehicles?

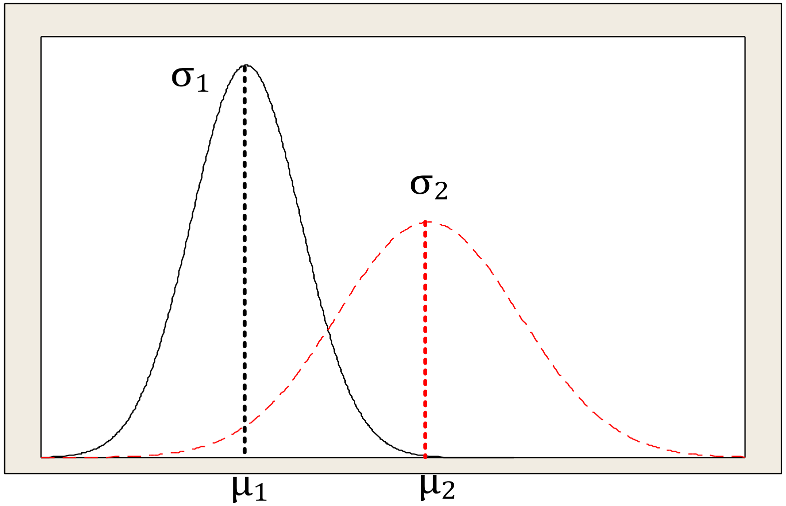
Normal Distribution



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$



Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

Chebyshev’s Rule

- The mean and standard deviation can be combined to make informative statements about how the values in a data set are distributed and about the relative position of a particular value in a data set.
- To do this, it is useful to be able to describe how far away a particular observation is from the mean in terms of the standard deviation.
- For example, we might say that an observation is 2 standard deviations above the mean or that an observation is 1.3 standard deviations below the mean.
- Sometimes in published articles, the mean and standard deviation are reported, but a graphical display of the data is not given.
- However, using a result called Chebyshev’s Rule, it is possible to get a sense of the distribution of data values based on our knowledge of only the mean and standard deviation.

Chebyshev's Inequality

If X is a r.v with mean μ and variance σ^2, then for any positive number k, we have

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}, or, P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

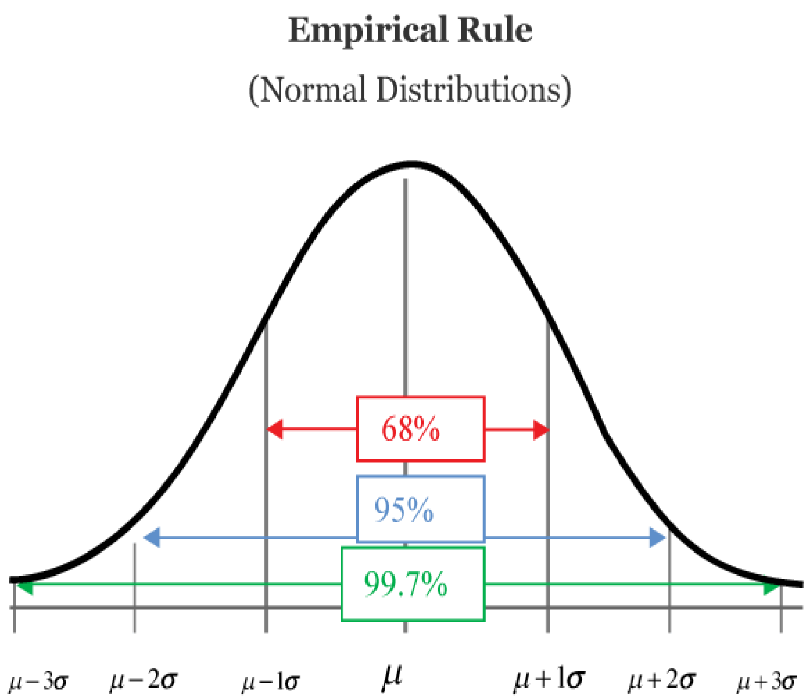
$$P(|X-\mu| < c) \geq 1-\frac{\sigma^2}{c^2}, or, P(|X-\mu| < k\sigma) \geq 1-\frac{1}{k^2}.$$

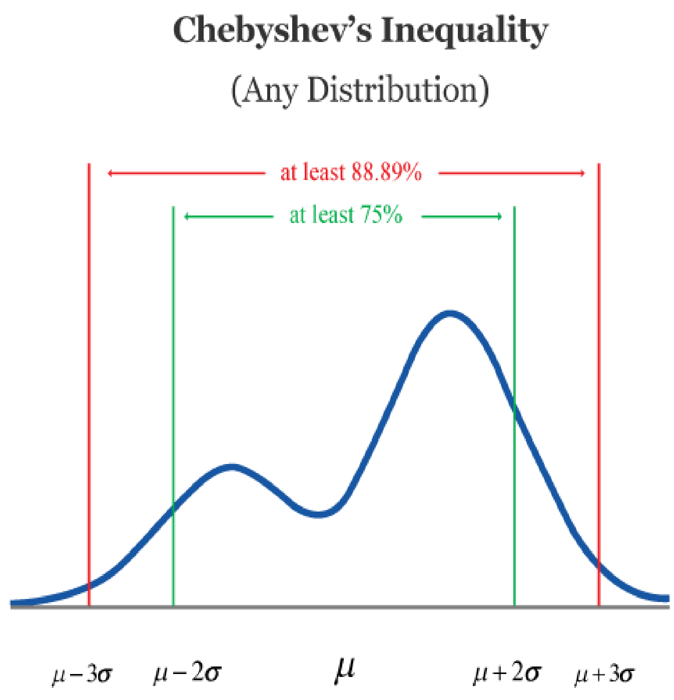
No. of S.D, k	$1-\frac{1}{k^2}$	**% within k S.D of the Mean**
2	$1-\frac{1}{4}$ =0.75	At least 75%
3	0.89	At least 89%
4	0.94	At least 94%
4.472	0.95	At least 95%
5	0.96	At least 96%
10	0.99	At least 99%

Empirical Rule

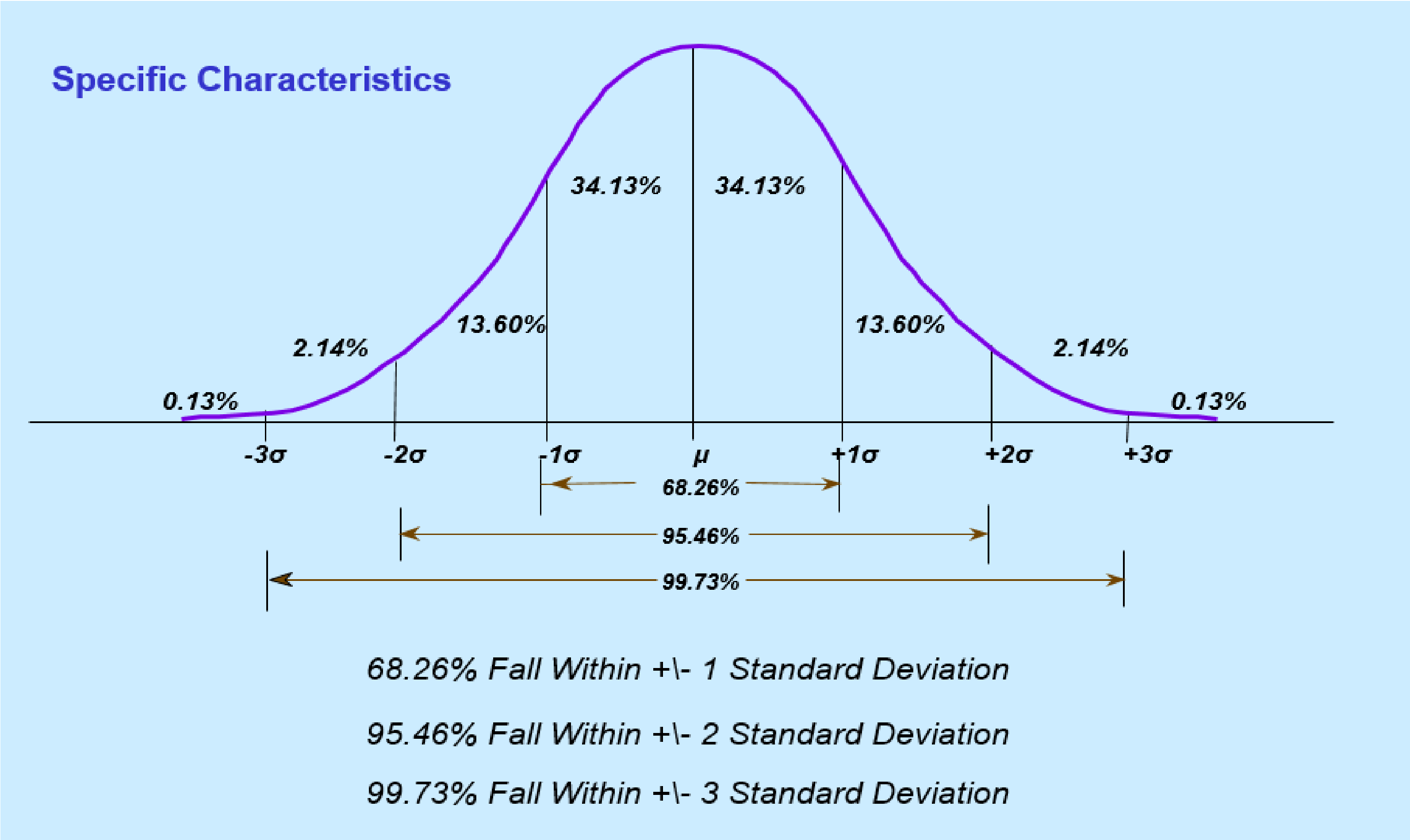
- The fact that statements based on Chebyshev’s Rule are frequently conservative suggests that we should look for rules that are less conservative and more precise.
- One useful rule is the **Empirical Rule**, which can be applied whenever the distribution of data values can be reasonably well described by a normal curve.
- The Empirical Rule : If the histogram of values in a data set can be reasonably well approximated by a normal curve, then
  - Approximately 68% of the observations are within 1 standard deviation of the mean.
  - Approximately 95% of the observations are within 2 standard deviations of the mean.
  - Approximately 99.7% of the observations are within 3 standard deviations of the mean.

Chebyshev's Inequality vs. Empirical Rule





Normal Curve (6-sigma)



What happens when X follows any continuous distribution? (Chebyshev's Inequality)

Z score

- The ***z* score** corresponding to a particular value is

$$z\ score = \frac{value - mean}{standard\ deviation}$$

- The *z* score tells us how many standard deviations the value is from the mean.
- It is positive or negative according to whether the value lies above or below the mean.
- The process of subtracting the mean and then dividing by the standard deviation is sometimes referred to as *standardization*, and a *z* score is one example of what is called a *standardized score.*

Examples

- A symmetric die is thrown 600 times. Find the lower bound for the probability of getting 80 to 120 sixes.

- Let $f(x) = \begin{cases} \frac{2}{3}x, 1 \le x \le 2 \\ 0, \quad elsewhere \end{cases}$

Give a bound using Chebyshev's for $P\left(\frac{10}{9} \le X \le 2\right)$. Calculate the actual probability. How do they compare?

Example

- A student took two national aptitude tests. The national average and standard deviation were 475 and 100, respectively, for the first test and 30 and 8, respectively, for the second test. The student scored 625 on the first test and 45 on the second test. Use ***z* scores** to determine on which exam the student performed better relative to the other test takers.
- A sample of concrete specimens of a certain type is selected, and the compressive strength of each specimen is determined. The mean and standard deviation are calculated as $\bar{x} = 3000$ and $s = 500$, and the sample histogram is found to be well approximated by a normal curve.
  - Approximately what percentage of the sample observations are between 2500 and 3500?
  - Approximately what percentage of sample observations are outside the interval from 2000 to 4000?
  - What can be said about the approximate percentage of observations between 2000 and 2500?
  - Why would you not use Chebyshev’s Rule to answer the questions posed in Parts (a)–(c)?

Standard Normal Distribution

- The normal distribution has computational complexity to calculate $P(x_1 < x < x_2)$ for any two (x_1, x_2) and given μ and σ
- To avoid this difficulty, the concept of z-transformation is followed.

$$\boxed{\mathrm{z} = \frac{x-\mu}{\sigma} \qquad \text{[Z-transformation]}}$$

- X: Normal distribution with mean μ and variance σ^2.
- Z: Standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.
- Therefore, if *f(x)* assumes a value, then the corresponding value of $f(z)$ is given by

$$f(x{:}\,\mu,\sigma):\;\; P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx$$

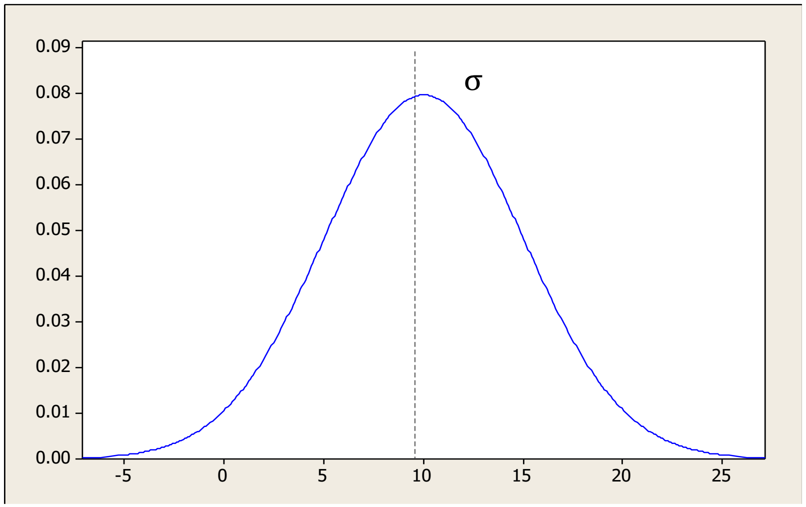
$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-\frac{1}{2}z^2}dz$$

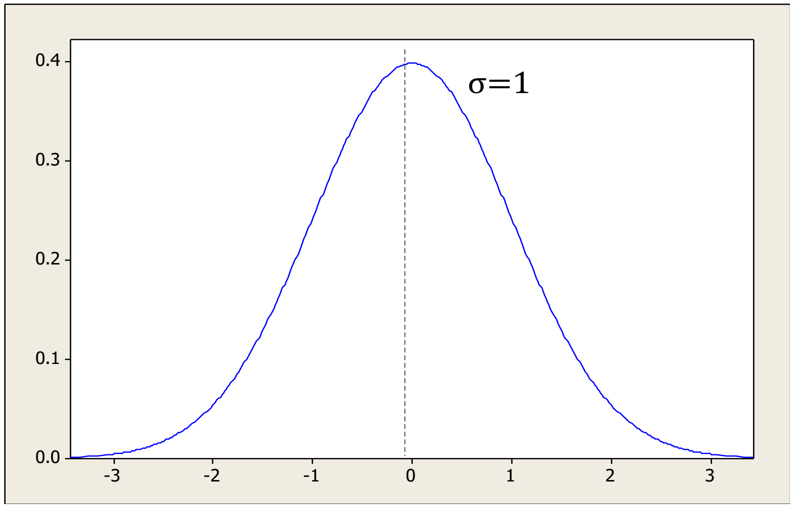
$$= f(z{:}\,0,\sigma)$$

Standard Normal Distribution

Definition : Standard normal distribution

The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.





$x=\mu$

$f(x: \mu, \sigma)$

$\mu=0$

$f(z: 0, 1)$

Gamma Distribution

The gamma distribution derives its name from the well known gamma function in mathematics.

Definition: **Gamma Function**

$$\Gamma(\alpha) = \int_0^{\alpha} x^{\alpha-1} e^{-x} dx \quad \text{for } \alpha > 0$$

Integrating by parts, we can write,

$$\begin{aligned}\Gamma(\alpha) &= (\alpha-1)\int_0^{\alpha} x^{\alpha-2} e^{-x} dx \\ &= (\alpha-1)\Gamma(\alpha-1)\end{aligned}$$

Thus Γ function is defined as a recursive function.

Gamma Function

When $\alpha = n$, we can write,

$$\Gamma(n) = (n-1)(n-2) \ldots \ldots \ldots . \Gamma(1)$$

$$= (n-1)(n-2) \ldots \ldots \ldots .3.2.1$$

$$= (n-1)!$$

Further, $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$

Note:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$ [An important property]

Gamma Distribution

Definition: **Gamma Distribution**

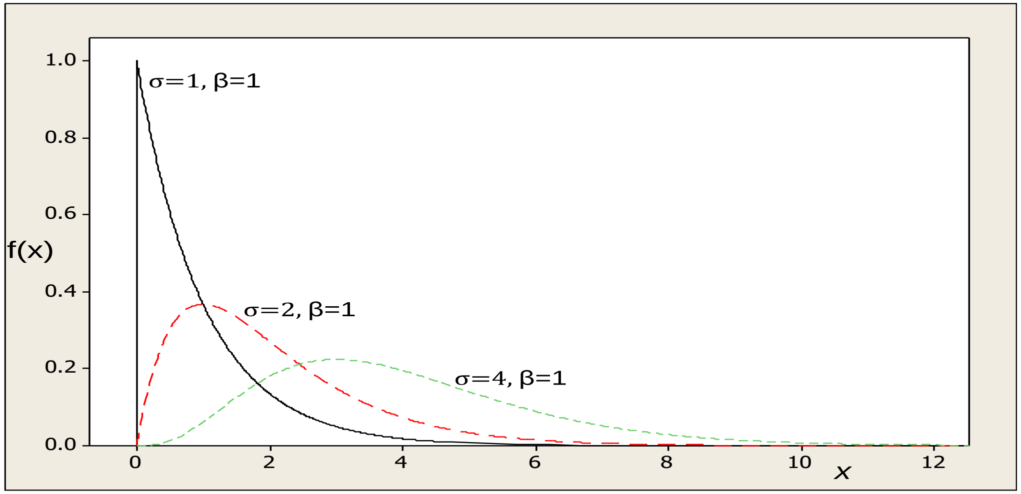
The continuous random variable x has a gamma distribution with parameters α and β such that:

$$f(x:\alpha,\beta)=\begin{cases}\dfrac{1}{\beta^{\alpha}\,\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}} & x>0\\ 0 & Otherwise\end{cases}$$

where $\alpha>0$ and β>0

Note: The mean and variance of gamma distribution are

$$\mu=\alpha\beta$$
$$\sigma^2=\alpha\beta^2$$



Exponential Distribution

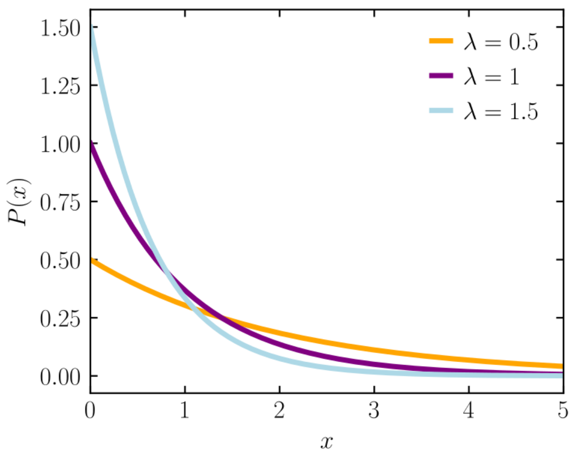
Definition: Exponential Distribution

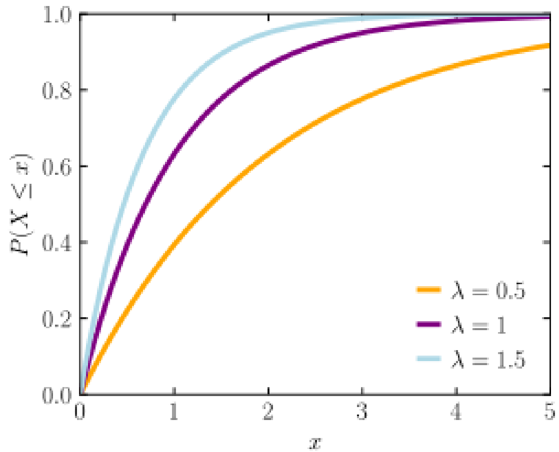
The continuous random variable X has an exponential distribution with parameter β, where:

$$f(x:\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{where } \beta > 0 \\ 0 & \end{cases}$$

Note: The mean and variance of exponential distribution are:

$$\mu = \beta$$
$$\sigma^2 = \beta^2$$





Lognormal Distribution

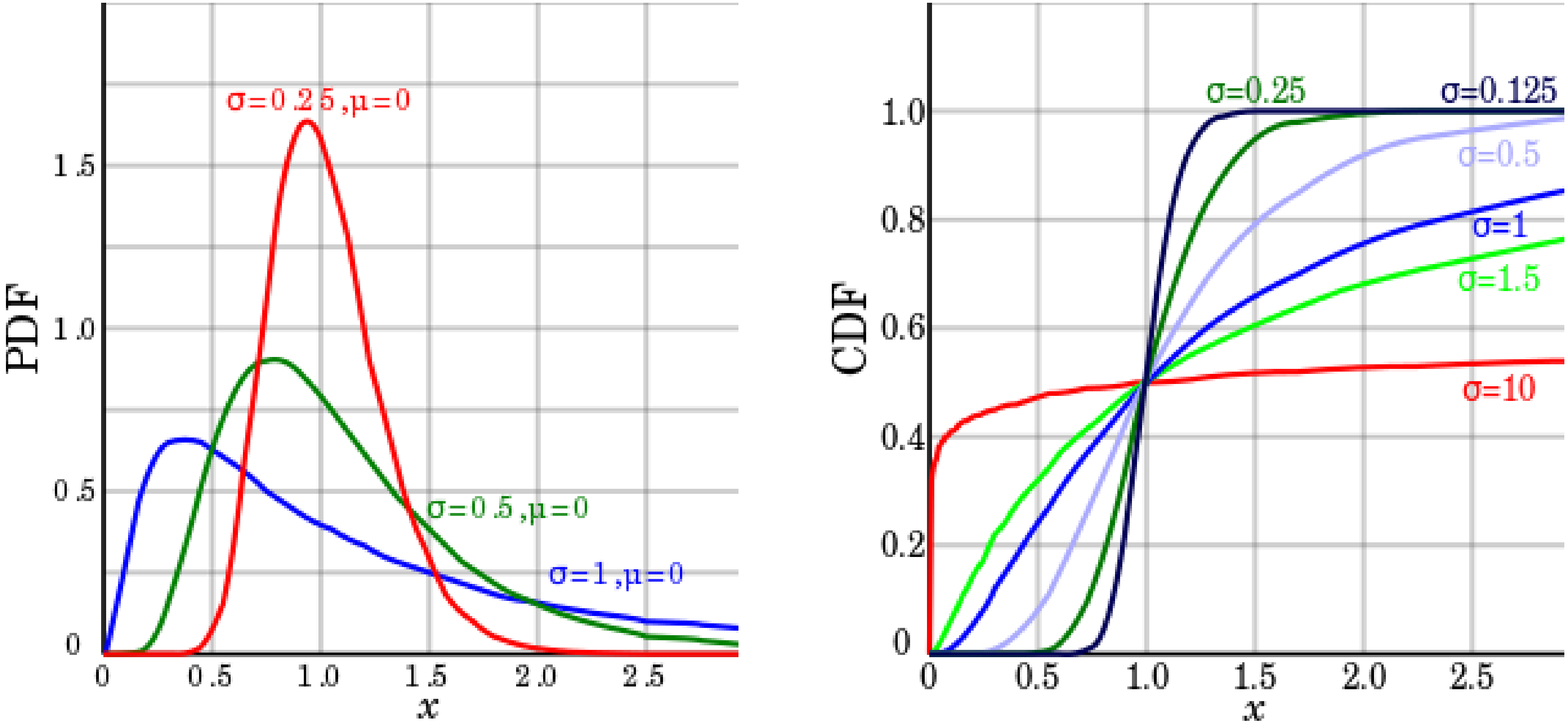
The lognormal distribution applies in cases where a natural log transformation results in a normal distribution.

Definition: Lognormal distribution

The continuous random variable x has a lognormal distribution if the random variable $y = \ln(x)$ has a normal distribution with mean μ and standard deviation σ. The resulting density function of x is:

$$f(x:\mu,\sigma) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Lognormal Distribution curves



Weibull Distribution

Definition: Weibull Distribution

The continuous random variable x has a Weibull distribution with parameter α and β such that.

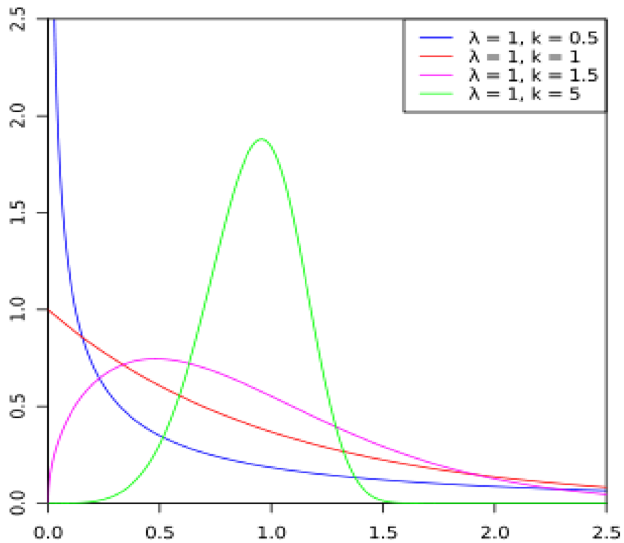
$$f(x:\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}} & x > 0 \\ 0 & Otherwise \end{cases}$$

where $\alpha > 0$ and $\beta > 0$

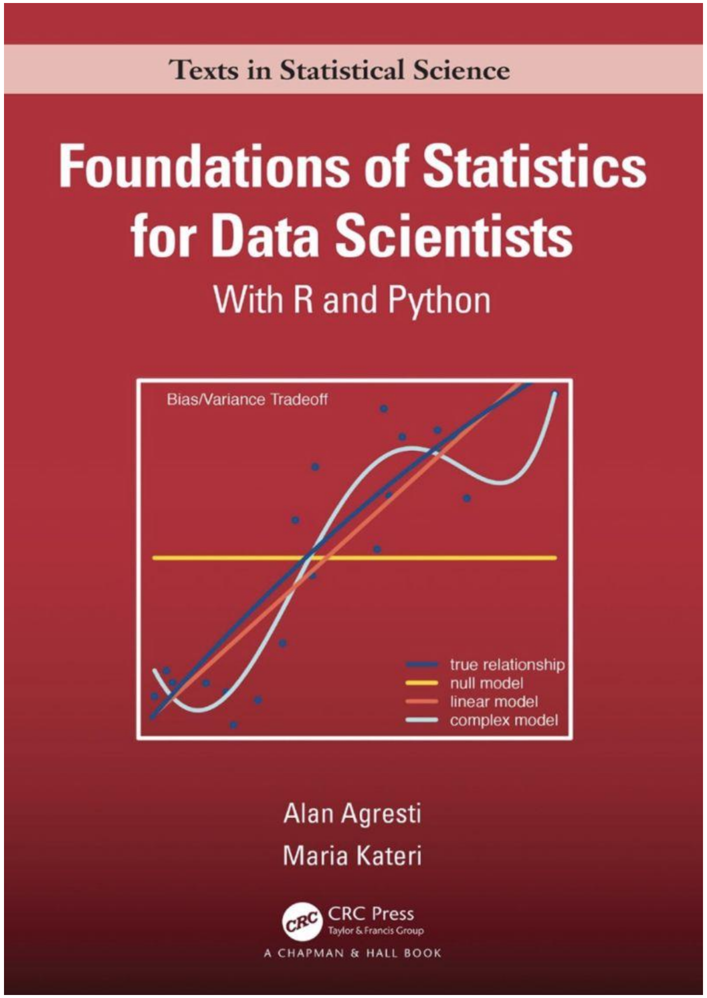
The mean and variance of Weibull distribution are:

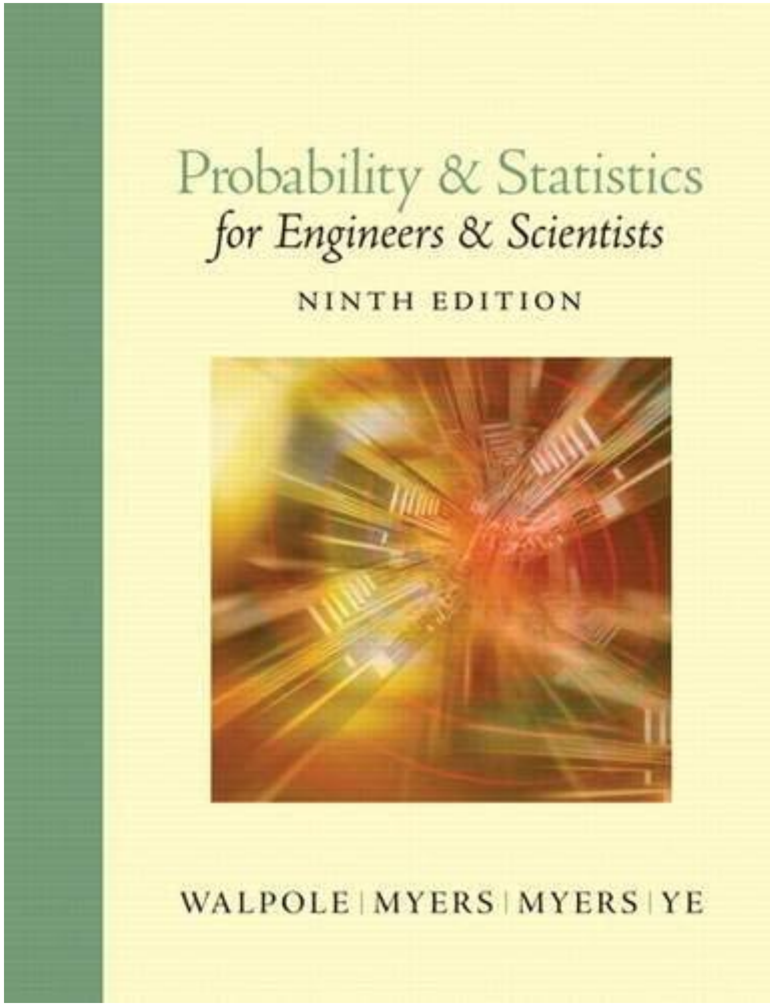
$$\mu = \alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right)$$

$$\sigma^2 = \alpha^{-2/\beta}\{\Gamma\left(1+\frac{2}{\beta}\right) - [\Gamma(1+\frac{1}{\beta})]^2\}$$



References:


Texts in Statistical Science
Foundations of Statistics for Data Scientists
With R and Python
Bias/Variance Tradeoff
true relationship
null model
linear model
complex model
Alan Agresti
Maria Kateri
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK


Probability & Statistics for Engineers & Scientists
NINTH EDITION
WALPOLE | MYERS | MYERS | YE