CENG4513
MODELLING & SILMULATION

LECTURE #4

Statistical Models

Statistical Models

Discrete Probability Distributions

- Binomial Distribution
- Bernoulli Distribution
- Geometric Distribution
- Discrete Poisson Distribution/Poisson Process

Continuous Probability Distributions

- Uniform
- Exponential
- Normal

Empirical Distributions

Discrete and Continuous Random Variables

Discrete Random Variable	Continuous Random Variable
Finite Sample Space e.g. {0, 1, 2, 3}	Infinite Sample Space e.g. [0,1], [2.1, 5.3]
Probability **Mass** Function (PMF) $p(x_i) = P(X = x_i)$ 1. $p(x_i) \geq 0$, for all i 2. $\sum_{i=1}^{\infty} p(x_i) = 1$	Probability **Density** Function (PDF) $f(x)$ 1. $f(x) \geq 0$, for all x in R_X 2. $\int_{R_X} f(x)dx = 1$ 3. $f(x) = 0$, if x is not in R_X
Cumulative Distribution Functions (CDF) $p(X \leq x)$	
$p(X \leq x) = \sum_{x_i \leq x} p(x_i)$	$p(X \leq x) = \int_{-\infty}^{x} f(t)dt = 0$ $p(a \leq X \leq b) = \int_{a}^{b} f(x)dx$

Expectation

- E(X): The expected value (the mean) of X
  - Discrete $$E(x)=\sum_{\text{all } i} x_i p(x_i)$$
  - Continuous $$E(x)=\int_{-\infty}^{\infty} xf(x)dx$$
- Var(X) : The variance of X
  - Discrete $$V(X)=\sum_{i=1}^{n}(x_i-\mu_x)^2\cdot p(x_i)=\left(\sum_{i=1}^{n}(x_i)^2\cdot p(x_i)\right)-\left(\sum_{i=1}^{n}x_i\cdot p(x_i)\right)^2$$
  - Continuous $$V(X)=\int_{-\infty}^{+\infty}(x-\mu_x)^2\cdot f(x)dx=\left(\int_{-\infty}^{+\infty}x^2\cdot f(x)dx\right)-\mu_x^2$$

Discrete Probability Distributions

- Bernoulli Trials
- Binomial Distribution
- Geometric Distribution
- Poisson Distribution
- Poisson Process

Discrete Probability Distributions

- Bernoulli Trials
- Binomial Distribution

Modeling of Random Events with Two-States

Discrete Probability Distributions

- Bernoulli Trials

- **Context:** Random events with two possible values
  - Two events: Yes/No, True/False, Success/Failure
  - Two possible values: 1 for success, 0 for failure.
  - Example: Tossing a coin, Packet Transmission Status…
- Probability Mass Function (PMF): Probability in one trial

$$\text{PMF: } p\left(\text{one trial}\right) = p(x_j) = \begin{cases} p, & x_j = 1,\ j = 1,2,\dots,n \\ 1-p = q, & x_j = 0,\ j = 1,2,\dots,n \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Expected Value: } E\left[X_j\right] = p$$

$$\text{Variance : } V\left[X_j\right] = \sigma^2 = p \cdot (1-p)$$

$$p\left(X_1, X_1, \dots, X_n,\right) = p\left(X_1\right) \cdot p\left(X_2\right) \cdot \dots p\left(X_n\right)$$

Discrete Probability Distributions

Binomial Distribution

- **Context:** Number of successes in a series of **n** trials.
  - Ex: the number of 'heads' occurring when a coin is tossed 50 times.
  - The number of successful transmissions of 100 packets
- Probability Mass Function (PMF)
  - Binomial distribution with parameters *n* and *p*, *X* ~ *B*(*n*,*p*)

$$P(X = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k} \quad \text{with} \quad \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

$$\text{Variance}: V[X] = \sigma^2 = n \cdot p \cdot (1-p)$$

$$\text{Expected Value: } E[X] = n \cdot p$$

Where;

- $k = 0, 1, 2, \ldots, n$ — **k: number of success**
- $n = 1, 2, 3, \ldots$ — n: number of trials
- $0 < p < 1$ — p = success probability

Discrete Probability Distributions

Binomial Distribution

Probability of k=2 successes in n=3 trials?

	E1: 1 1 0	P(E1)= p(1 and 1 and 0) = p p (1-p) = p^2 (1-p)
or	E2: 1 0 1	P(E2)= p(1 and 0 and 1) = p (1-p) p = p^2 (1-p)
or	E3: 0 1 1	P(E3)= p(1 and 1 and 0) = p p (1-p) = p^2 (1-p)

Probability of k=2 successes in n=3 trials?

$$p(X=2) = p(E1) + p(E2) + p(E3) = 3 \cdot p^2 \cdot (1-p)$$

$$= \frac{3!}{2!\,(3-2)!} \cdot p^2 \cdot (1-p) = \frac{6}{2} p^2 \cdot (1-p)$$

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

Discrete Probability Distributions

- Poisson Distribution
- Poisson Process

Modeling of Random Number of Arrivals/Events

Discrete Probability Distributions

Poisson Distribution

- **Context:** number of events occurring in a fixed period of time
  - Events occur with a known average rate and are independent

- Possion distribution is characterized by the average rate λ
  - The average number of arrival in the fixed time period.

Discrete Probability Distributions

Poisson Distribution

Examples:

- The number of cars passing a fixed point in a 5 minute interval. Average rate: λ = 3 cars/5 minutes
- The number of calls received by a switchboard during a given period of time. Average rate: λ =3 call/minutes
- The number of message coming to a router per second
- The number of travelers arriving to the airport for flight registration

Discrete Probability Distributions

Poisson Distribution

- The Poisson distribution with the average rate parameter λ

PMF: $$p(k) = P(X = k) = \begin{cases} \dfrac{\lambda^k}{k!}\exp(-\lambda) & \text{for } k = 0,1,2, \ldots. \\ 0, & \text{otherwise} \end{cases}$$

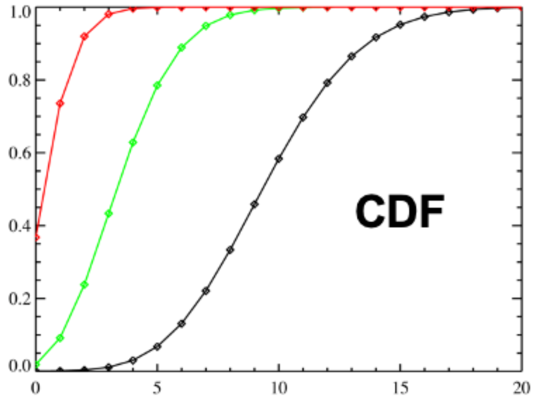
the probability that there are **exactly** *k* arrivals in a certain period of time.

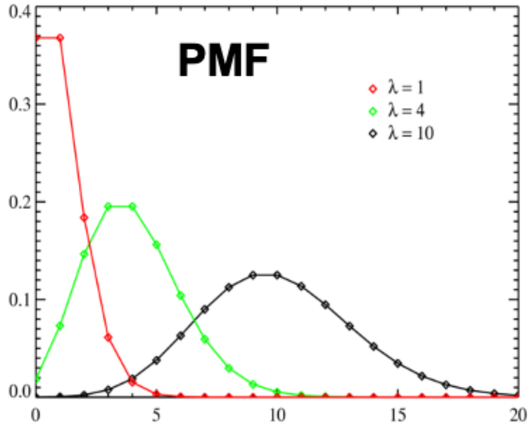
CDF: $$F(k) = p(X \le k) = \sum_{i=0}^{k} \frac{\lambda^i}{i!} \cdot \exp(-\lambda)$$

the probability that there are **at least** *k* arrivals in a certain period of time.

Expected value: $E[X] = \lambda$

Variance: $V[X] = \lambda$





Discrete Probability Distributions

Poisson Distribution

Example:

- The number of cars that enter the parking follows a Poisson distribution with a mean rate equal to λ = 20 **cars/hour**

  - The probability of having exactly **15 cars** entering the parking in **one hour**:

$$p(15) = P(X = 15) = \frac{20^{15}}{15!} \cdot \exp(-20) = 0.051649$$

or

$$p(15) = F(15) - F(14) = 0.156513 - 0.104864 = 0.051649$$

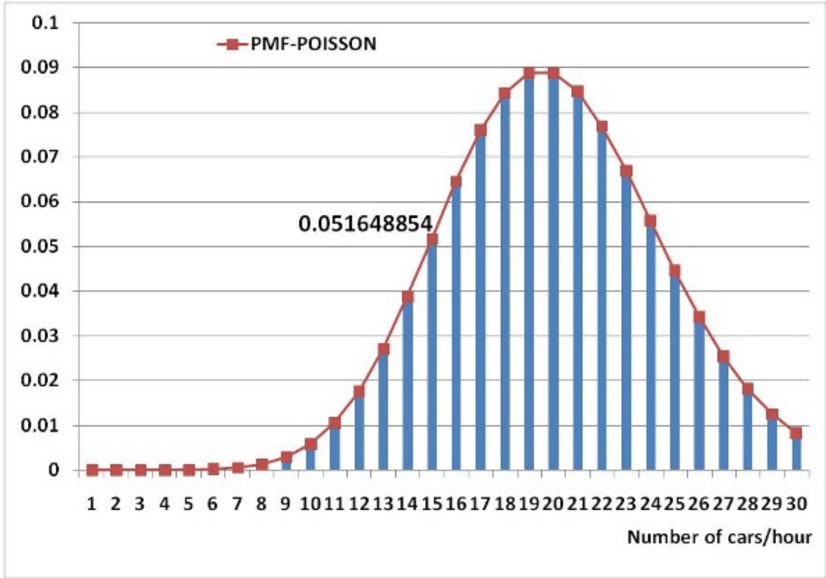
  - The probability of having more than **3 cars** entering the parking in **one hour**:

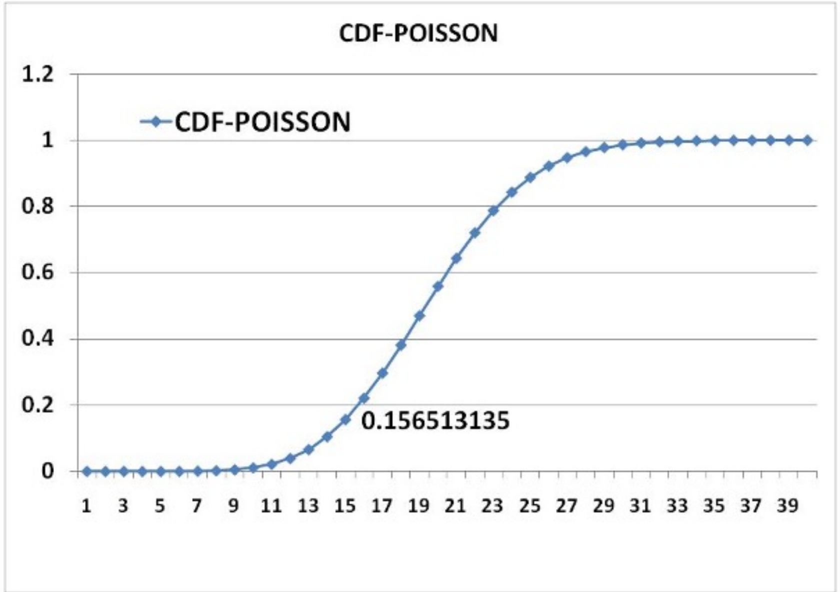
$$p(X > 3) = 1 - p(X \leq 3) = 1 - F(3)$$
$$= 1 - \left[p(0) + p(1) + p(2) + p(3)\right]$$
$$= 0.9999967$$

Discrete Probability Distributions

Poisson Distribution

Example:





Probability Mass Function
Poisson (λ = 20 cars/hour)

$$p(X = k) = \frac{20^k}{k!} \cdot \exp(-20)$$

Cumulative Distribution Function
Poisson (λ = 20 cars/hour)

$$F(k) = p(X \leq k) = \sum_{i=0}^{k} \frac{20^i}{i!} \cdot \exp(-20)$$

Discrete Probability Distributions

Poisson Process

- Process N(t)
  - random variable N that depends on time t

- Poisson Process: a Process that has a Poisson Distribution
  - $N(t)$ is called a **counting** Poisson process

There are two types:

- *Homogenous Poisson Process*: constant average rate
- *Non-Homogenous Poisson Process*: variable average rate

Discrete Probability Distributions

Poisson Process

- The Poisson random variable arises frequently in computer and communication system modelling.
- It is typically used as a counting variable, recording the number of events that occur *in a given period of time*;

  for example; the number of messages transmitted in a communication network node in one minute or the number of users arriving at a terminal in one hour.

Discrete Probability Distributions

Poisson Process

Examples:

- The number of web page requests arriving at a server may be characterized by a Poisson process except for unusual circumstances such as coordinated denial of service attacks.
- The number of telephone calls arriving at a switchboard, or at an automatic phone-switching system, may be characterized by a Poisson process.
- The arrival of "customers" is commonly modelled as a Poisson process in the study of simple queueing systems.

Discrete Probability Distributions

Homogenous Poisson Process

Examples:

- The number of web page requests arriving at a server may be characterized by a Poisson process except for unusual circumstances such as coordinated denial of service attacks.
- The number of telephone calls arriving at a switchboard, or at an automatic phone-switching system, may be characterized by a Poisson process.
- The arrival of "customers" is commonly modelled as a Poisson process in the study of simple queueing systems.

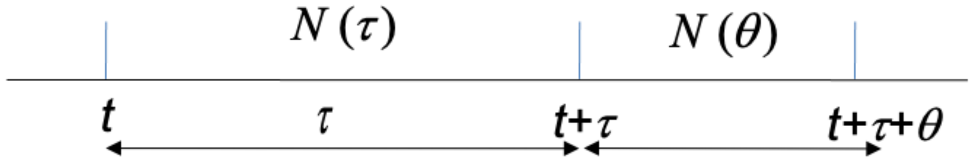
Discrete Probability Distributions

Homogenous Poisson Process

- Formally, a counting process $\{N(\tau), \tau \geq 0\}$ is a (homogenous) **Poisson process** with constant average rate λ if:

for $t \geq 0$ and $n = 0, 1, 2, \ldots$

$$\text{PMF:} \quad p\left[N(t+\tau) - N(t) = n\right] = p\left[N(\tau) = n\right] = \frac{(\lambda \cdot \tau)^n}{n!} \exp(-\lambda \cdot \tau)$$



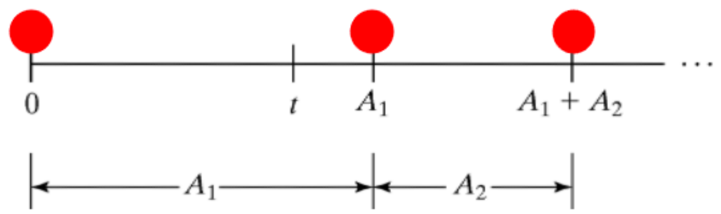
- $N(t+\tau) - N(t)$ describes the number of events in time interval $(t, t+\tau]$
- The mean and the variance are equal

$$E\left[N(\tau)\right] = V\left[N(\tau)\right] = \lambda \cdot \tau$$

Discrete Probability Distributions

Inter-arrival Times of a Poisson Process

Consider the **inter-arrival times** of a Poisson process (A_1, A_2, …), where A_i is the elapsed time between arrival *i* and arrival *i+1*



The first arrival occurs after time *t* **MEANS** that there are no arrivals in the interval [0,t], As a consequence:

$$p\left(A_1 > t\right) = p\left(N\left(t\right) = 0\right) = \exp\left(-\lambda \cdot t\right)$$

$$p\left(A_1 \leq t\right) = 1 - p\left(A_1 > t\right) = 1 - \exp\left(-\lambda \cdot t\right)$$

CDF of Exponential distribution

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

The Inter-arrival times of a Poisson process are exponentially distributed and independent with mean 1/ λ

Discrete Probability Distributions

Relationship between Poisson & Exponential Distributions

- If the occurrence of events is governed by a Poisson distribution, then the inter-event times are governed by an exponential distribution with the same parameter, and vice versa.
- Therefore, if we know that "the delay until an event" or "the time btw two events" is exponentially distributed then the number of that event occurring in an interval of time has Poisson distribution.
- The exponential distribution is used extensively in performance modelling! Think about it?

Discrete Probability Distributions

Non-Homogenous Poisson Process

- *The* non-homogeneous (non-stationary) Poisson process (NSPP) is characterized by a **VARIABLE rate parameter λ(t),** the arrival rate at time *t*. In general, the rate parameter may change over time.

- The stationary increments, property is not satisfied

$$\exists s, t : \; N(t) - N(s) \neq N(t-s)$$

- The **expected number of events** (e.g. arrival) between time *s* and time *t* is

$$\lambda_{s,t} = \int_s^t \lambda(u) \cdot du$$

Continuous Probability Distributions

- Uniform Distribution
- Exponential Distribution
- Normal (Gaussian) Distribution

Continuous Probability Distributions

Uniform Distribution

The **continuous uniform distribution** is a family of probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are **equally probable**

A random variable *X* is uniformly distributed on the interval [*a*,*b*], *U*(*a*,*b*), if its PDF and CDF are:

$$\text{PDF:} \quad f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

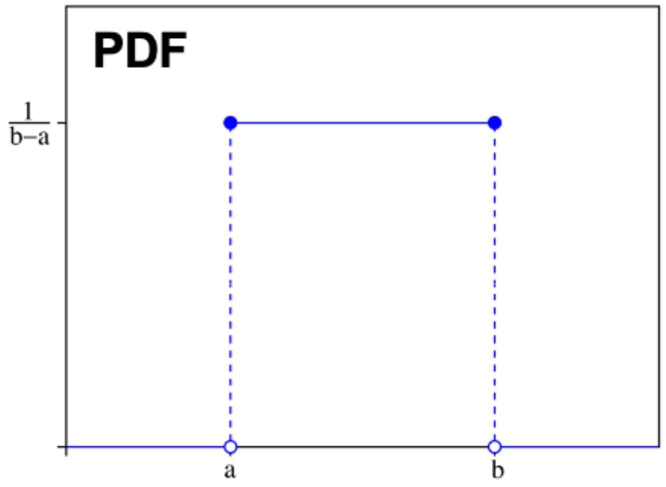
$$\text{CDF:} \quad F(x) = \begin{cases} 0, & x \prec a \\ \dfrac{x-a}{b-a}, & a \le x \prec b \\ 1, & x \ge b \end{cases}$$

$$\text{Expected value:} \quad E[X] = \frac{a+b}{2}$$

$$\text{Variance:} \quad V[X] = \frac{(a+b)^2}{12}$$

Continuous Probability Distributions

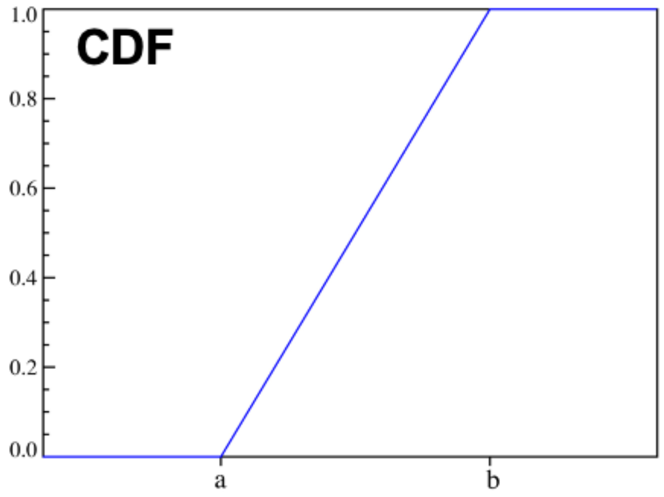
Uniform Distribution



Properties

$p(x1 \le X \le x2)$ is proportional to the length of the interval

$$F(X_2) - F(X_1) = \frac{X_2 - X_1}{b - a}$$



Special case: a **standard uniform** distribution U(0,1).

- Very useful for random number generators in simulators

Continuous Probability Distributions

Exponential Distribution

- The **exponential distribution** describes the times between events in a Poisson process, in which events occur continuously and independently at a constant average rate.
- A random variable X is exponentially distributed with parameter $\mu = 1/\lambda > 0$ if its PDF and CDF are:

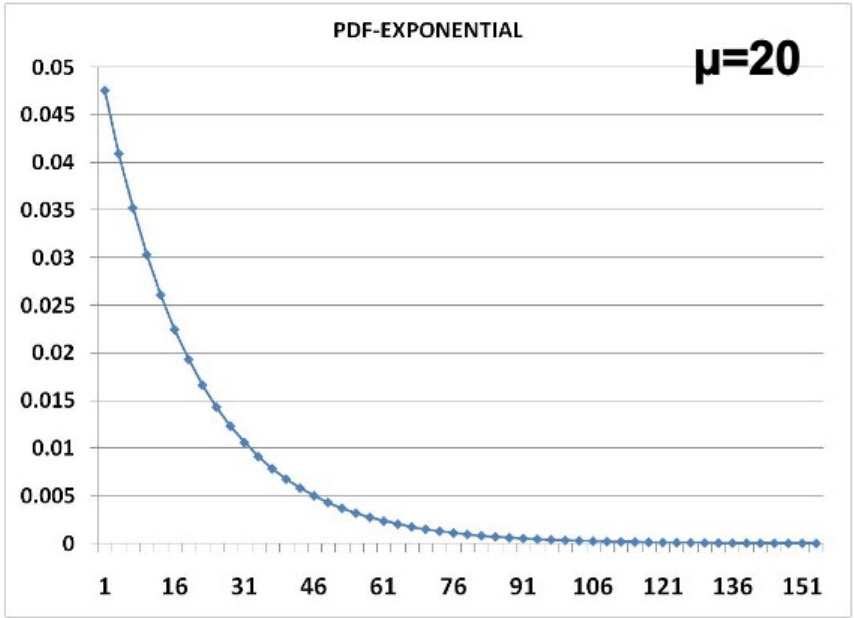
$$\text{PDF:} \quad f(x) = \begin{cases} \lambda \cdot \exp(-\lambda \cdot x), & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \Longrightarrow \quad f(x) = \begin{cases} \dfrac{1}{\mu} \cdot \exp\left(-\dfrac{x}{\mu}\right), & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

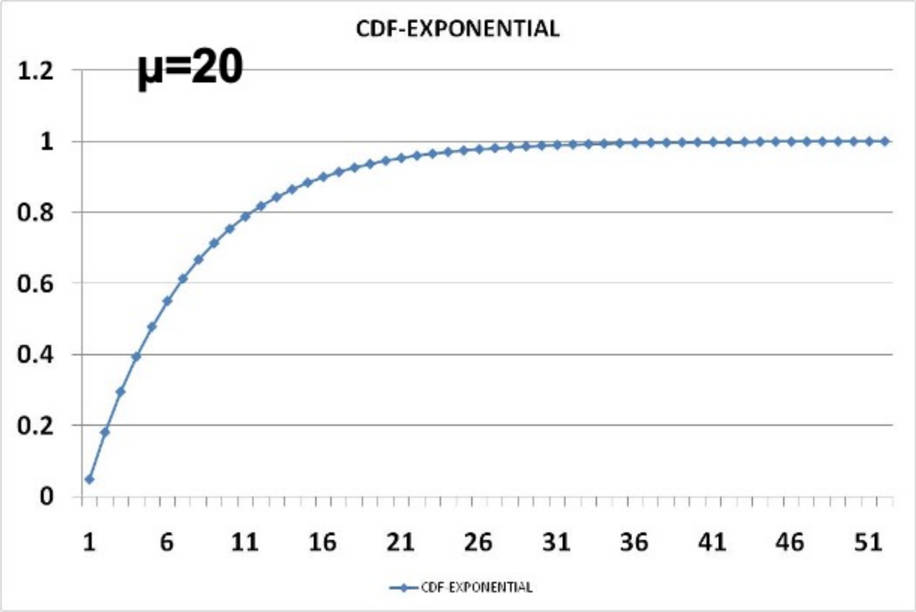
$$\text{CDF:} \quad F(x) = \begin{cases} 0, & x < 0 \\ \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}, & x \geq 0 \end{cases} \quad \Longrightarrow \quad F(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp\left(-\dfrac{x}{\mu}\right), & x \geq 0 \end{cases}$$

$$\text{Expected value:} \quad E[X] = \frac{1}{\lambda} = \mu \qquad \text{Variance:} \quad V[X] = \frac{1}{\lambda^2} = \mu^2$$

Continuous Probability Distributions

Exponential Distribution





$$f(x) = \begin{cases} \frac{1}{20} \cdot \exp\left(-\frac{x}{20}\right), & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp\left(-\frac{x}{20}\right), & x \geq 0 \end{cases}$$

Continuous Probability Distributions

Exponential Distribution

- **The memoryless property:** In probability theory, **memoryless** is a property of certain probability distributions: the exponential distributions and the geometric distributions, wherein any derived probability from a set of random samples is distinct and has no information (i.e. "memory") of earlier samples.
- Formally, the memoryless property is:
  For all *s* and *t* greater or equal to 0:

$$p\left(X > s + t \mid X > s\right) = p\left(X > t\right)$$

- This means that the future event do not depend on the past event, but only on the present event
  - The fact that $Pr(X > 40 \mid X > 30) = Pr(X > 10)$ does not mean that the events $X > 40$ and $X > 30$ are **independent**;
    i.e. it does not mean that $Pr(X > 40 \mid X > 30) = Pr(X > 40)$

Continuous Probability Distributions

Exponential Distribution

- **The memoryless property:** can be read as "the probability that you **will** wait more than *s+t* minutes given that you **have already** been waiting *s* minutes is equal to the probability that you will wait *t* minutes."

- In other words "The probability that you will wait s more minutes given that you have already been waiting t minutes is the same as the probability that you had wait for more than s minutes from the beginning."

$$p\left(X > s + t \mid X > s\right) = p\left(X > t\right)$$

- The fact that Pr(X > 40 | X > 30) = Pr(X > 10) does not mean that the events X > 40 and X > 30 are **independent**;
i.e. it does not mean that Pr(X > 40 | X > 30) = Pr(X > 40).

Continuous Probability Distributions

Normal (Gaussian) Distribution

- The **Normal distribution**, also called the **Gaussian distribution**, is an important family of continuous probability distributions, applicable in many fields.

- Each member of the family may be defined by two parameters, **location** and **scale**: **the mean** ("average", μ) and **variance** (standard deviation squared, σ^2) respectively.

- The importance of the normal distribution as a model of quantitative phenomena in the **natural** and **behavioral** sciences is due in part to the **Central Limit Theorem**.

- It is usually used to model system error (e.g. channel error), the distribution of natural phenomena, height, weight, etc.

Continuous Probability Distributions

Normal (Gaussian) Distribution

A **continuous random variable** *X*, taking all real values in the range (-∞,+∞) is said to follow a **Normal distribution** with parameters **μ** and **σ** if it has the following PDF and CDF:

$$\text{PDF: } f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

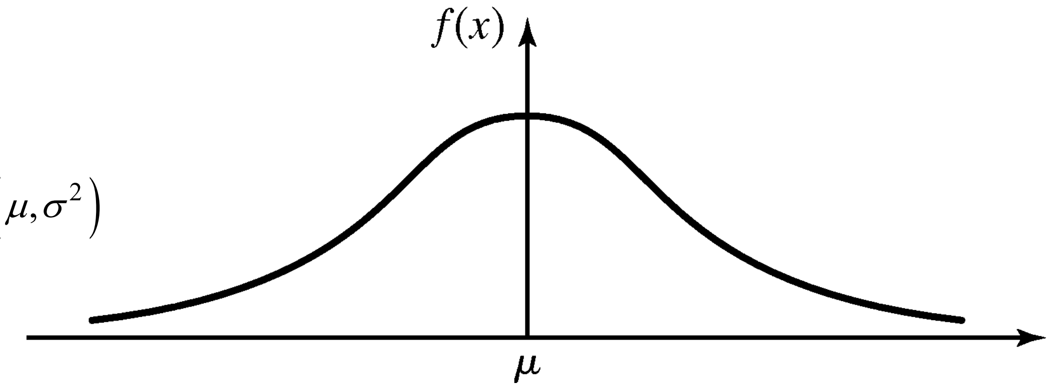
$$\text{CDF: } F(x) = \frac{1}{2} \cdot \left(1 + erf\left(\frac{x-\mu}{\sigma \cdot \sqrt{2}}\right)\right)$$

Where;

The Normal distribution is denoted as $X \sim N(\mu, \sigma^2)$

This probability density function (PDF) is

- a symmetrical, bell-shaped curve,
- centered at its expected value μ.
- The variance is s².

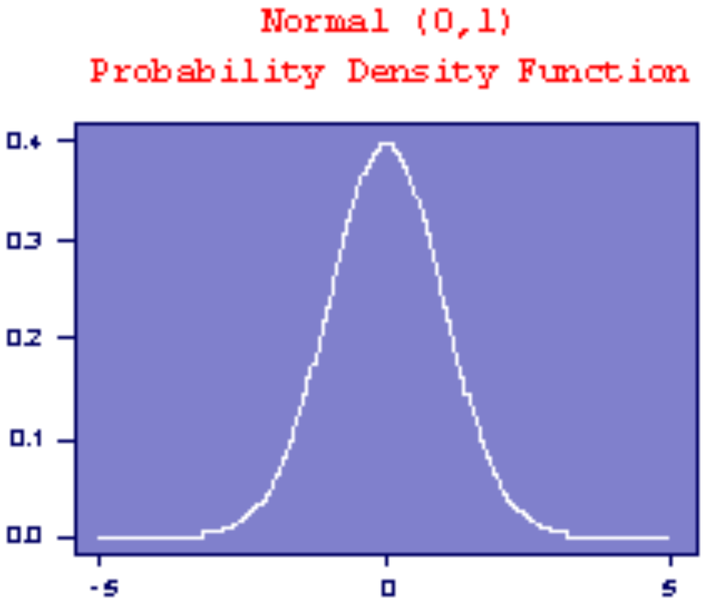


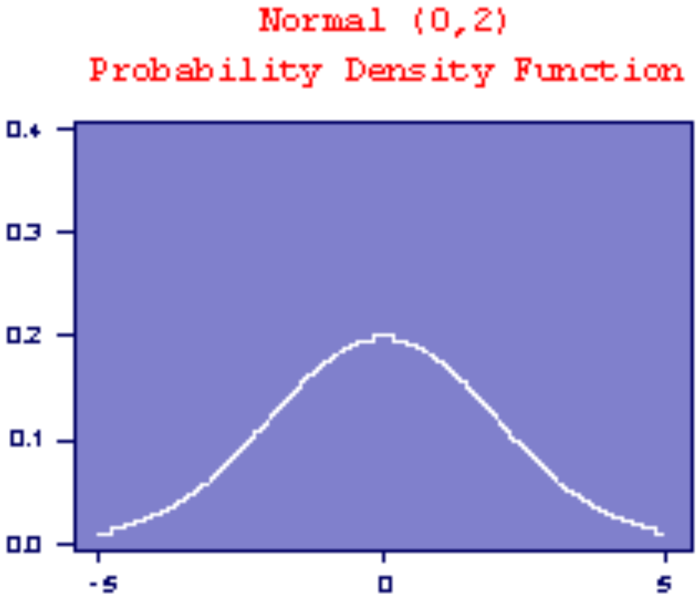
Continuous Probability Distributions

Normal (Gaussian) Distribution

Example:

The simplest case of the normal distribution, known as the Standard Normal Distribution, has expected value zero and variance one. This is written as $N(0,1)$.





Statistical Analysis Tools

Statistical Analysis Tools

- Comparing Distributions: Tests for Goodness-of-Fit
  - Chi-Square Distribution (for discrete models: PMF)
  - Kolmogorov-Smirnov Test (for continuous models: CDF)

- Evaluating the relationship
  - Linear Regression
  - Correlation

Goodness-of-fit

- Statistical Tests enables to compare between two distributions, also known as **Goodness-of-Fit**.
- The **goodness-of-fit** of a statistical model describes how well it fits a set of observations.
- Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question
- Goodness-of-fit means how well a statistical model fits a set of observations

Goodness-of-fit

- Chi-Square (χ^2) Test
- Kolmogorov-Smirnov (K-S) Test

Goodness-of-fit

Chi-Square Test

- The Pearson's chi-square χ^2 test enables to compare two probability mass functions of two distribution.

- If the difference value (Error) is greater than **the critical value,** the two distribution are said to be different or *the first distribution does not fit (well) the second distribution.*

- If the difference if smaller that the critical value, *the first distribution fits well the second distribution*

Goodness-of-fit

Chi-Square Test

- Pearson's chi-square is used to assess two types of comparison:
  - tests of goodness of fit: it establishes whether or not an observed frequency distribution differs from a theoretical distribution.
  - tests of independence: it assesses whether paired observations on two variables are independent of each other.

Goodness-of-fit

Chi-Square Test

- If the calculated **chi-square value is less or equal than the critical (tabulated) chi-square value,** then we say that
  - both distributions **are equal** (goodness-of-fit) or
  - the row variable is unrelated (that is, only randomly related) to the column variable (test of independence).