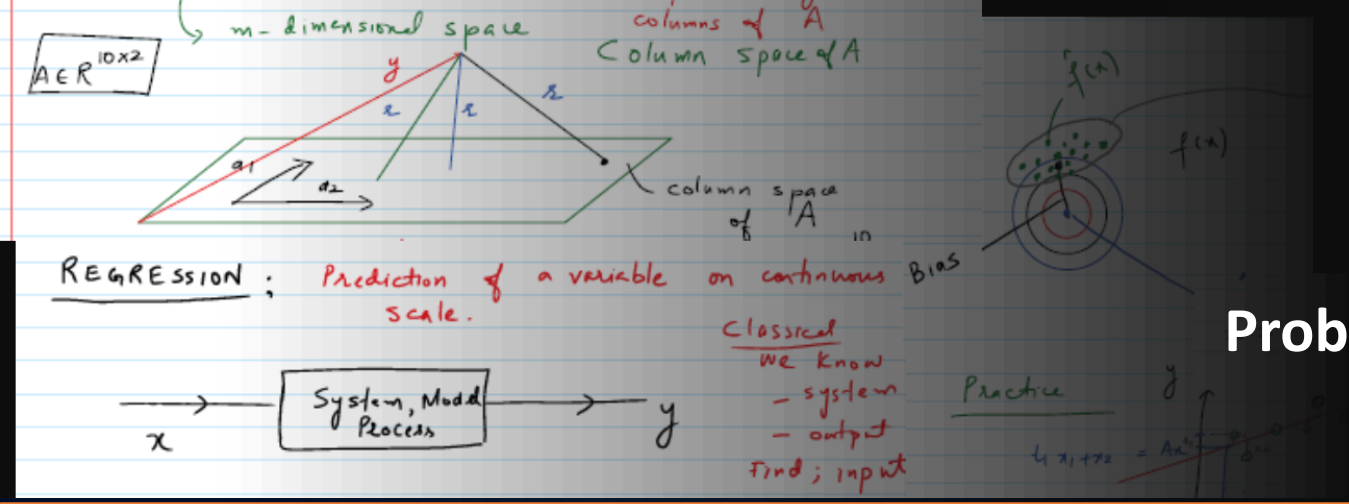


Machine Learning
EE514 – CS535

Probability Theory – Review of Basic Concepts

Zubair Khalid

School of Science and Engineering
Lahore University of Management Sciences

https://www.zubairkhalid.org/ee514_2021.html

LUMS
A Not-for-Profit University

Probability Theory Overview

Basic Definitions

Relative Frequency:

- Consider an experiment that can result in *M* possible outcomes O_1 , O_2 … O_M
- Let $N_n(O_i)$ denotes the number of times ,O_i occurred in *n* trials
- Relative frequency of outcome: $\frac{N_n(O_i)}{n}$
- When number of trials *n* becomes large, the relative frequency converge to some limiting value.
- This behaviour is known as statistical regularity.

Reference: ***Probability and Random Processes for Electrical and Computer Engineers*** by John A. Gubner

Probability Theory Overview

Basic Definitions

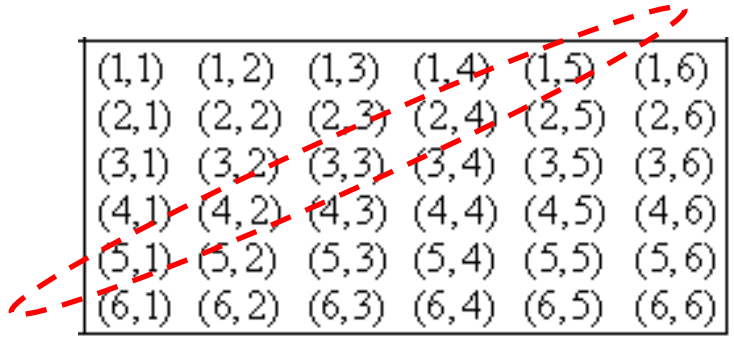
- **Sample Space:** set of all possible distinct **outcomes** of an experiment.
  - Outcome: ω
  - Sample space: Ω

- **Events:** Collection of outcomes are called events. Usually denoted by capital letters. Every event is a subset of sample space.

Examples:

1. Rolling two dies together.
   - sample space?
   - event: the sum of numbers on two dies = *6*?

2. Noise voltage can be between *0* and *5* volts.
   1. sample space?
   2. event: the noise voltage is between *2* and *3* volts?



Sample space

Probability Theory Overview

Probability Model:

- Mathematical modelling of the problem.
- Basic model if each outcome is equally likely: $P(A) = \frac{|A|}{|\Omega|}$
- Example: Fair die

- If outcomes are not equally likely;

$$P(A) = \frac{|A|}{|\Omega|} = \sum_{\omega \in A} \frac{1}{|\Omega|} = \sum_{\omega \in A} p(\omega)$$

- $p(\omega)$ probability for each outcome $\omega \in \Omega$

Probability Theory Overview

Axioms of Probability:

Given a nonempty set Ω, called the sample space, and a function P defined on the subsets of Ω, we say P is a probability measure if the following four axioms are satisfied:

1. The empty set $\emptyset$ is called the impossible event. $P(\emptyset) = 0$.
2. For any event $A \subset \Omega, P(A) \geq 0$.
3. If $A1$, $A2$, ... are events that are mutually exclusive, that is, $A_n \cap A_m = \emptyset$ for $n \neq m$, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

4. $P(\Omega) = 1$, Ω is a sure event.

Monotonicity property:

$$A \subset B \Rightarrow P(A) \leq P(B)$$

Inclusion-exclusion property:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probability Theory Overview

Conditional Probability and Bayes Theorem:

Motivation:

Conditional probability, important concept in probabilistic modeling, allows us to update probabilistic models when additional information is revealed.

Probability of event A given event B;

$$P(A|B) = \frac{\text{outcomes in A and B}}{\text{outcomes in B}} = \frac{\text{outcomes in A and B}}{\text{total}} \frac{\text{total}}{\text{outcomes in B}} = \frac{P(A \cap B)}{P(B)}$$

Similarly, we have

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Since $P(A \cap B) = P(B \cap A)$, we can write

$$P(A|B)P(B) = P(B|A)P(A)$$

$$\Rightarrow \boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Bayes Theorem

$$P(A) = P(A \cap B) + P(A \cap B^c)$$
$$= P(A|B)P(B) + P(A|B^c)P(B^c)$$

Law of Total Probability

Probability Theory Overview

Example:

Very accurate Covid Test:

Given that the 0.008 of the entire population is positive, the probability of correct positive prediction is 0.98 and the probability of correct negative prediction is 0.97.

A patient takes a covid test and the result is positive. What is the probability that the patient is not suffering from Covid?

- $P(\text{Covid } +\text{ve}) = 0.008$
- $P(\text{Covid } -\text{ve}) = 0.992$
- $P(\text{Test } +\text{ve} \,|\, \text{Covid } +\text{ve}) = 0.98$
- $P(\text{Test } -\text{ve} \,|\, \text{Covid } -\text{ve}) = 0.97$
- $P(\text{Test } +\text{ve} \,|\, \text{Covid } -\text{ve}) = 0.03$

- $$\begin{aligned} P(\text{Test } +\text{ve}) &= P(\text{Test } +\text{ve} \,|\, \text{Covid } +\text{ve}) P(\text{Covid } +\text{ve}) \\ &\quad + P(\text{Test } +\text{ve} \,|\, \text{Covid } -\text{ve}) P(\text{Covid } -\text{ve}) \\ &= (0.98)(0.008) + (0.03)(0.992) = 0.0376 \end{aligned}$$

- $$P(\text{Covid } +\text{ve} \,|\, \text{Test } +\text{ve}) = \frac{P(\text{Test } +\text{ve} \,|\, \text{Covid } +\text{ve})\, P(\text{Covid } +\text{ve})}{P(\text{Test } +\text{ve})} = \frac{(0.98)(0.008)}{0.0376}$$

$$= 0.2085$$

Probability Theory Overview

Independence:

In probability theory, if events A and B satisfy $\mathsf{P}(A|B) = \mathsf{P}(A|B^c)$, we say A does not depend on B. This condition says that

$$\Rightarrow \qquad \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(A \cap B^c)}{\mathsf{P}(B^c)}$$

Using $\mathsf{P}(A) = \mathsf{P}(A \cap B) + \mathsf{P}(A \cap B^c)$ and $\mathsf{P}(B^c) = 1 - \mathsf{P}(B)$

$$\Rightarrow \qquad \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(A) - \mathsf{P}(A \cap B)}{1 - \mathsf{P}(B)}$$

$$\Rightarrow \qquad \mathsf{P}(A \cap B)[1 - \mathsf{P}(B)] = \mathsf{P}(B)[\mathsf{P}(A) - \mathsf{P}(A \cap B)]$$

$$\Rightarrow \qquad \mathsf{P}(A \cap B) - \mathsf{P}(A \cap B)\mathsf{P}(B) = \mathsf{P}(A)\mathsf{P}(B) - \mathsf{P}(A \cap B)\mathsf{P}(B)$$

$$\Rightarrow \qquad \boxed{\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)}$$

Probability Theory Overview

Independence:

- Mutually exclusive events not to be confused with independent events. **(Different)**

- Interpretation of independence between events:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A)\,\mathrm{P}(B)}{\mathrm{P}(B)} = \mathrm{P}(A)$$

- If A and B are independent events, A^c and B, A and B^c, A^c and B^c are also independent events.

- Independence of more than 2 events:

$$\mathrm{P}(A_i \cap A_j \cap A_k) = \mathrm{P}(A_i)\,\mathrm{P}(A_j)\,\mathrm{P}(A_k)$$

- Mutually independence vs Pairwise independence

- Mutually independence implies pairwise independence but not the other way.

Probability Theory Overview

Discrete Random Variable:

- **Random Variable (Definition):**

  Random variable is a **function** which **maps** elements from the **sample space** to the **real line**.

- Random variables are denoted by upper case letters (X or Y).

- Individual outcomes for RV are denoted by lower case letters (x or y).

- Mathematically, $X(\omega)$ is a real-valued function defined for $\omega \in \Omega$.

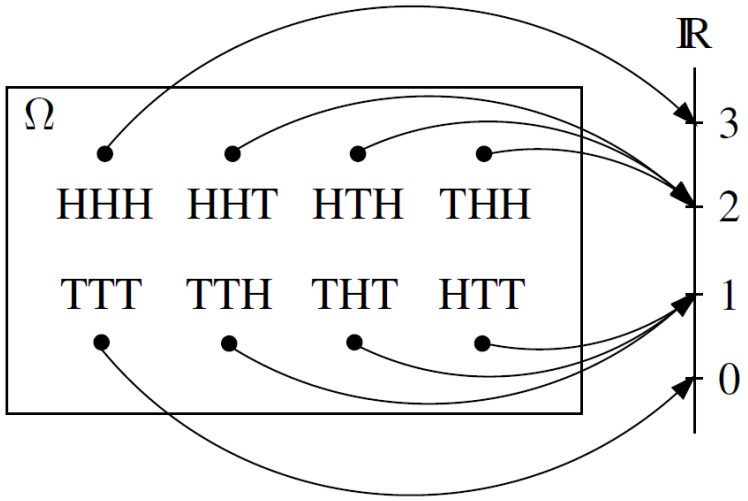
- For each element of an experiment's sample space, the random variable can take on exactly one value.
- **Discrete Random Variable: A RV that can take on only a finite or countably infinite set of outcomes.**

Probability Theory Overview

Discrete Random Variable – Example 1:

A random variable $X(\omega)$ = number of heads if three coins are tossed at the same time

Sampe space: $\Omega := \{\text{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH}\}$



$$X(\omega) := \begin{cases} 0, & \omega = \text{TTT}, \\ 1, & \omega \in \{\text{TTH, THT, HTT}\}, \\ 2, & \omega \in \{\text{THH, HTH, HHT}\}, \\ 3, & \omega = \text{HHH}. \end{cases}$$

Probability Theory Overview

Discrete Random Variable – Example 2:

A random variable $X(\omega)$ = number of girls in a family of 4 kids.

Lower case x is a particular value of $X(\omega)$.

ω	**Random Variable *X***
BBBB	*x=0*
GBBB	*x=1*
BGBB	*x=1*
BBGB	*x=1*
BBBG	*x=1*
GGBB	*x=2*
GBGB	*x=2*
GBBG	*x=2*
BGGB	*x=2*
BGBG	*x=2*
BBGG	*x=2*
BGGG	*x=3*
GBGG	*x=3*
GGBG	*x=3*
GGGB	*x=3*
GGGG	*x=4*

Probability Theory Overview

Discrete Random Variable – Example 3:

Random variable, Y = Sum of the up faces of the two die.

Die 1 →

Die 2 ↓	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

y
2
3
4
5
6
7
8
9
10
11
12

Probability Theory Overview

Probability Mass Function:

- **Probability Mass Function:** Assigns probabilities (masses) to the individual outcomes. (Also referred as probability density function.)

- For a random variable X, its pmf is given by

$$p_X(x_i) := \mathsf{P}(X = x_i)$$

- By axioms of probability;
  - pmf is between 0 and 1 $$0 \leq p_X(x_i) \leq 1$$
  - sum of all probabilities equal to 1 $$\sum_i p_X(x_i) = 1$$

Probability Theory Overview

Probability Mass Function – Example 1:

A random variable $X(\omega)$ = number of heads if three coins are tossed at the same time

$$p_X(0) = \mathrm{P}(X=0) = \mathrm{P}(\{\mathrm{TTT}\}) = \frac{|\{\mathrm{TTT}\}|}{|\Omega|} = \frac{1}{8}$$

$$p_X(1) = \mathrm{P}(X=1) = \mathrm{P}(\{\mathrm{HTT,THT,TTH}\}) = \frac{3}{8}$$

$$p_X(2) = \mathrm{P}(X=2) = \mathrm{P}(\{\mathrm{HHT,HTH,HHT}\}) = \frac{3}{8}$$

$$p_X(3) = \mathrm{P}(X=3) = \mathrm{P}(\{\mathrm{HHH}\}) = \frac{1}{8}$$

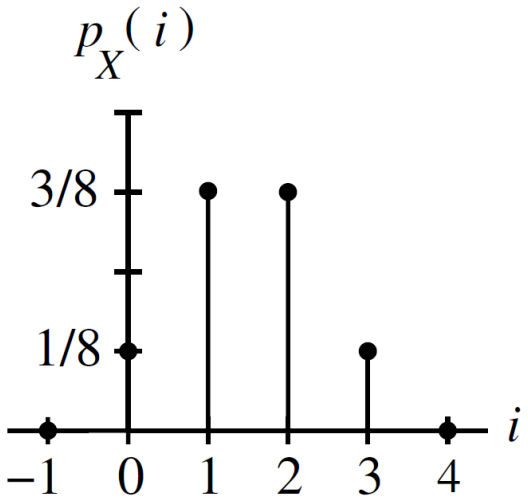


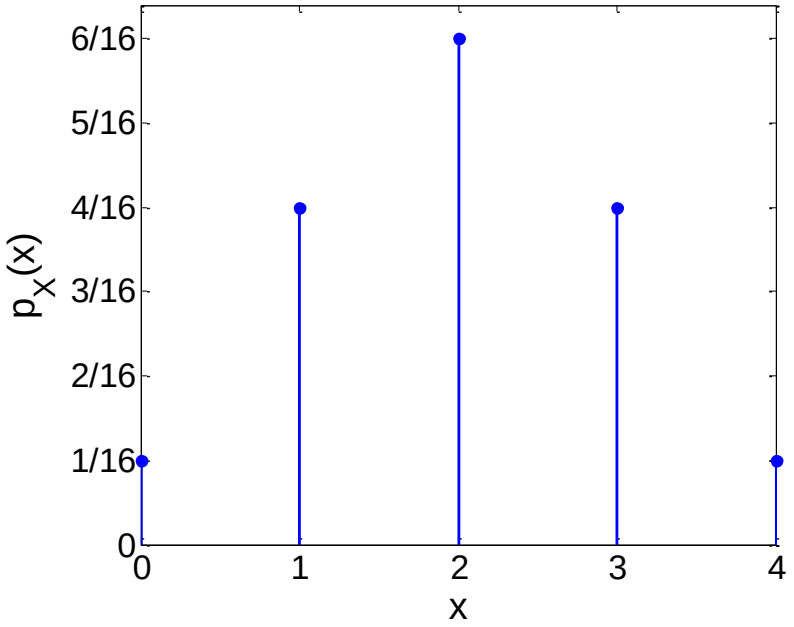
Fig 2: pmf of RV X

Probability Theory Overview

Probability Mass Function – Example 2:

A random variable X = number of girls in a family of 4 kids

Number of Girls, x	Probability, $p_X(x)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16
Total	16/16=1.00



What is the probability of exactly 3 girls in 4 kids?

What is the probability of at least 3 girls in 4 kids?

Probability Theory Overview

Important Random Variables:

1. **Bernoulli Random Variable**

   If there are only two outcomes of an experiment, the experiment is modeled with uniform random variable. For example, the tossing of coin is modeled with Bernoulli random variable.

   - It is most common to associate {0,1} to the outcomes of an experiment.
   - pmf is given by,

$$p_X(0) = \theta$$

$$p_X(1) = 1 - \theta$$

Probability Theory Overview

Important Random Variables:

2. Uniform Random Variable:

If the outcomes of an experiment are finite, and are equally likely, the experiment is modeled with uniform random variable.

- If there are *n* outcomes of an experiment, probability of each outcome = $\frac{1}{n}$.
- If outcomes are indexed, *k=1, 2,..., n*, $\mathrm{P}(X = k) = \frac{1}{n}, \quad k = 1, \ldots, n$
- pmf is given by,

$$p_X(k) = \begin{cases} 1/n, & k = 1, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

Probability Theory Overview

Important Random Variables:

3. Poisson Variable:

A random variable X is said to have a Poisson probability mass function with parameter $\lambda > 0$, denoted by $X \sim \text{Poisson}(\lambda)$, if

$$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots.$$

- Parameter λ fully characterizes the distribution.

- Used in modelling of physical phenomenon arising in different applications:
- arrival of photons at a telescope
- distribution of nodes in wireless sensor networks
- telephone calls arriving in a system
- arrival of network messages in a queue for transmission

Probability Theory Overview

Multiple Random Variables:

- When events are defined by more than one random variable.
- Let X represent one variable and Y represent another random variable, which maps elements of sample space to real line, but can be different, then the event involving both X and Y is described as

$$\{X \in B, Y \in C\} := \{\omega \in \Omega : X(\omega) \in B \text{ and } Y(\omega) \in C\}$$

- This is taken as an event that X belongs to B and Y belongs to C.
- **Very important to understand the concept:** the event above is a function of two random variable and is comprised of only those points on the real line which are common between B and C, that is,

$$\{X \in B, Y \in C\} = \{X \in B\} \cap \{Y \in C\}$$

Probability Theory Overview

Multiple Random Variables – Probability Mass Function :

- The joint probability involving two random variables is given by the probability of the joint event

$$\begin{aligned} \mathsf{P}(X \in B, Y \in C) &:= \mathsf{P}(\{X \in B, Y \in C\}) \\ &= \mathsf{P}(\{X \in B\} \cap \{Y \in C\}) \end{aligned}$$

- taking $B = \{x_i\}$ and $C = \{y_j\}$, define joint probability mass function,

$$p_{XY}(x_i, y_j) := \mathsf{P}(X = x_i, Y = y_j)$$

- **Interpretation:** $p_{XY}(x_i, y_j)$ gives the probability that the RV $X = x_i$ and RV $Y = y_i$ at the same time.

- **Marginal probability mass function:** We can obtain $p_X(x_i)$ and $p_Y(y_j)$

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j)$$

$$p_Y(y_j) = \sum_i p_{XY}(x_i, y_j)$$

Probability Theory Overview

Multiple Random Variables – Concept of Independence:

- When RVs X and Y are independent events, we can write the joint probability as

$$\mathsf{P}(X \in B, Y \in C) = \mathsf{P}(X \in B)\mathsf{P}(Y \in C)$$

$$\mathsf{P}(X = x_i, Y = y_j) = \mathsf{P}(X = x_i)\,\mathsf{P}(Y = y_j)$$

- Equivalently, we can write in terms of joint pmf and individual pms of RVs:

$$p_{XY}(x, y) = p_X(x)p_Y(y)$$

- The concepts presented for two random variables are also valid for more than two random variables.

Probability Theory Overview

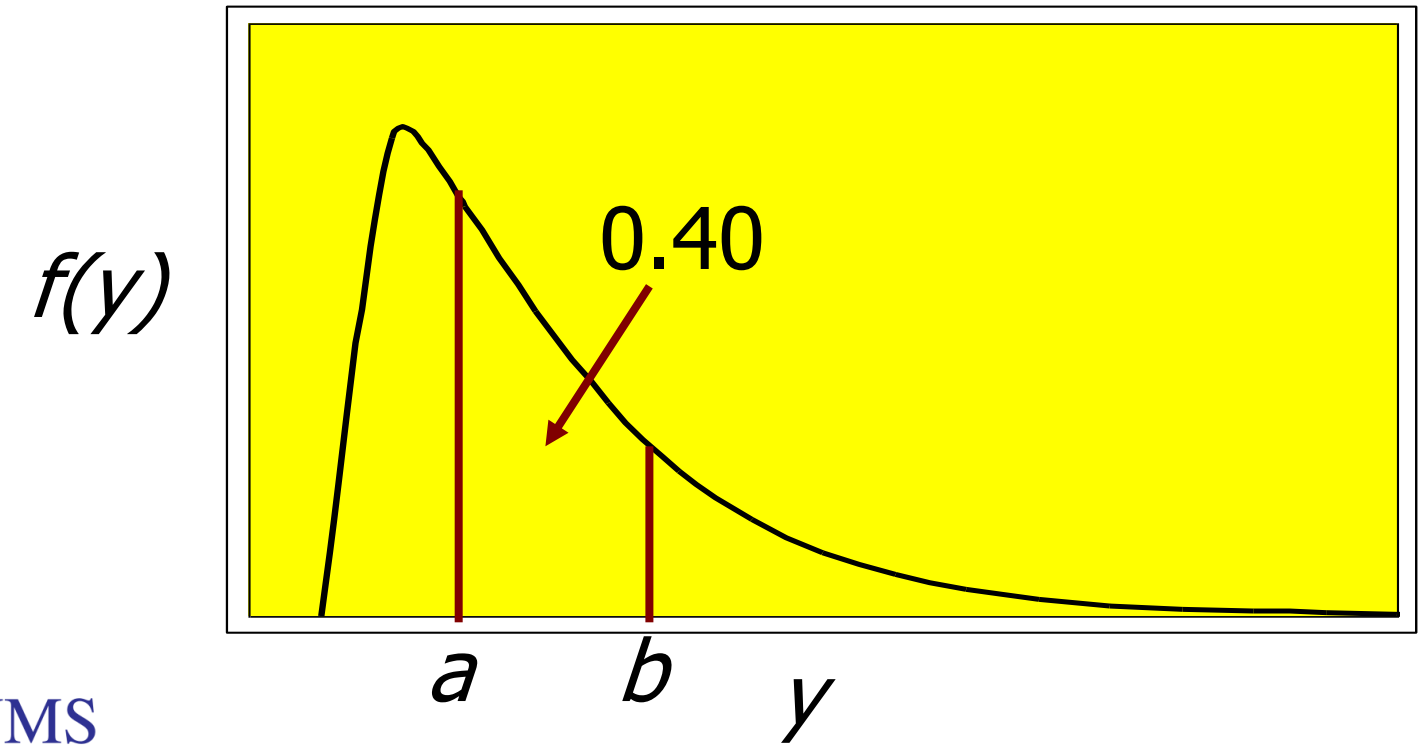
Continuous Random Variable:

- A continuous random variable is one for which the outcome can be any value in an interval of the real number line.
- There are always infinitely many sample points in the sample space.
- For **discrete** random variables, only the value listed in the **pmf** have positive probabilities, all other values have probability zero.
- For continuous random variables, the probability of every specific value is zero. Probability only exists for an interval of values for continuous RV., that is, for continuous RV Y,
  - We don't calculate $P(Y = y)$, we calculate $P(a < Y < b)$, where a and b are real numbers.
  - For a continuous random variable $P(Y = y) = 0$.

Probability Theory Overview

Continuous Random Variable - Probability density function:

- The **probability density function (pdf)** denotes a curve against the possible values of random variable and the area under an interval of the curve is equal to the probability that random variable is in that interval.
- For example if *f(y)* denotes the pdf of RV *Y*, we calculate P($a < Y < b$),



Probability Theory Overview

pmf vs pdf:

- For a discrete random variable, we have probability mass function (pmf).
- The pmf looks like a bunch of spikes, and probabilities are represented by the heights of the spikes.
- For a continuous random variable, we have a probability density function (pdf).
- The pdf looks like a curve, and probabilities are represented by areas under the curve.

Probability Theory Overview

Continuous Random Variable – Characteristics of pdf:

- Given Y is a continuous random variable with pdf is $f(x)$.
- By axioms of probability, $f(x)$ must satisfy the following conditions:

1. $f(x) \geq 0$ for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Probability Theory Overview

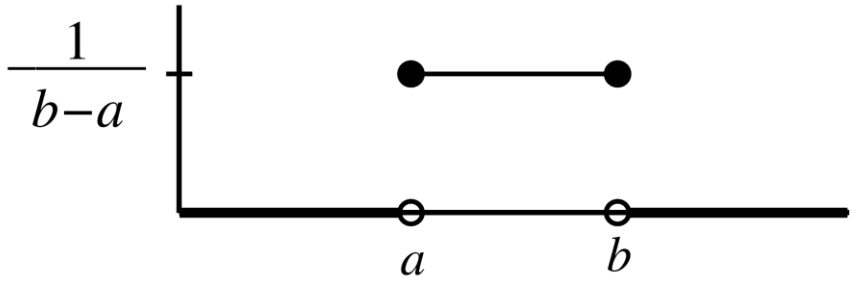
Important Continuous Random Variable:

- **Uniform random variable:** used to model the experiments in which outcome is constrained to lie in a known interval, say [a,b] and all possible outcomes are equally likely.

- Define uniform random variable $f \sim$ uniform[a,b] for $a<b$ with pdf

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

- Plot of pdf



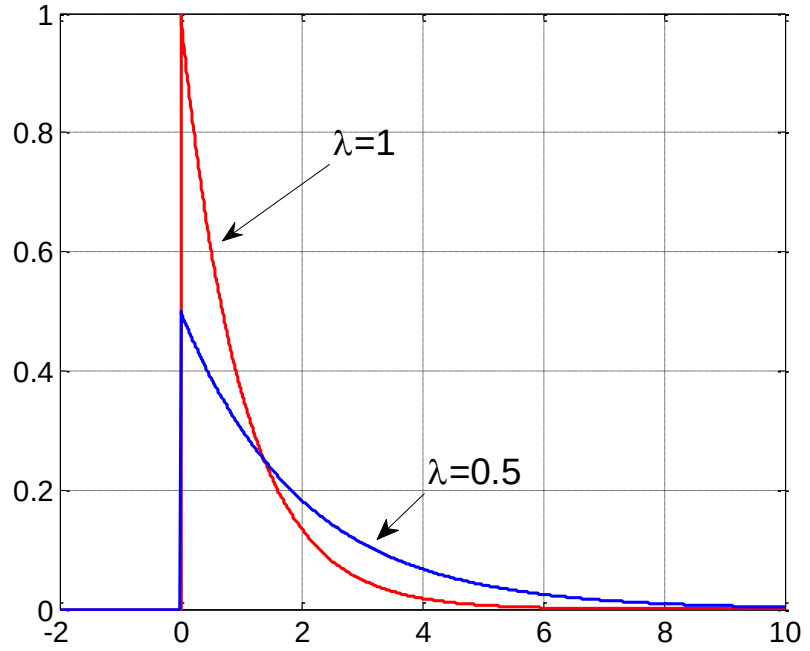
Probability Theory Overview

Important Continuous Random Variable:

- **Exponential random variable:** used to model lifetimes, such as
  - how long it takes before next phone call arrives
  - how long it takes a computer network to transmit a message
  - how long it takes a radioactive particle to decay

- Define $f \sim \exp(\lambda)$ for $\lambda > 0$ with pdf given by:

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



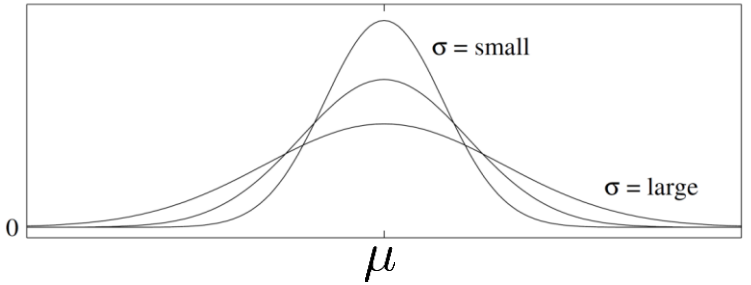
Probability Theory Overview

Important Continuous Random Variable:

- **Gaussian (Normal) random variable:**

- Define Gaussian RV $f \sim N(\mu, \sigma^2)$.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



- center $\mu \in \mathbf{R}$
- standard deviation, $\sigma^2 \in R^+$, quantifies the spread of the pdf
- $N(0,1)$ is called standard normal density