

# Probability\*

2025-07-26

---

\*This note contains parts that I learnt from the Probability and Statistics course of Georgia Tech university in [edx.org](https://edx.org).

# Contents

<b>1. Pre-requisites</b>	<b>4</b>
1.1 Bootcamp: Set . . . . .	4
1.2 Bootcamp: Derivative . . . . .	6
1.3 Bootcamp: Integration . . . . .	8
<b>2. Introduction to Probability</b>	<b>11</b>
2.1 Introduction . . . . .	11
2.1.1 Counting Techniques . . . . .	13
2.1.2 Permutation . . . . .	13
2.1.3 Combination . . . . .	14
2.1.4 Binomial Theorem . . . . .	15
2.1.5 Problems . . . . .	15
2.2 Hypergeometric Distribution . . . . .	17
2.3 Binomial Distribution . . . . .	17
2.4 Multinomial Coefficients . . . . .	17
2.5 Conditional Probability . . . . .	18
2.6 Independence . . . . .	20
2.7 Partitions and laws of probability . . . . .	22
2.8 Bayes Theorem . . . . .	23
2.9 Probability Problems . . . . .	23
<b>3. Random Variables</b>	<b>25</b>
3.1 Univariate Random Variables . . . . .	25
3.1.1 Discrete Random Variable . . . . .	26
3.1.1.1 Uniform Distribution: . . . . .	26
3.1.1.2 Binomial Distribution: . . . . .	26
3.1.1.3 Poisson Distribution: . . . . .	27
3.1.2 Continuous Random Variables . . . . .	27
3.1.2.1 Uniform Distribution . . . . .	28

3.1.2.2 Exponential Distribution . . . . .	28
3.1.3 Cumulative Probability Distribution . . . . .	29
3.1.4 Great Expectations . . . . .	32
3.1.4 Law of the unconscious statistician (LOTUS) . . . . .	34
3.1.5 Some Probability Inequalities . . . . .	40
3.1.6 Functions of Random Variable . . . . .	42
3.1.7 Inverse Transform Theorem/Probabiity Integral Transform . . . . .	45

# 1. Pre-requisites

## 1.1 Bootcamp: Set

**Set** is a collection of objects. Members of set are called elements.

**Notation:**

For sets,  $A, B, C, \dots$

For elements,  $a, b, c, \dots$

For membership,  $\in$  e.g.  $a \in A$

For non membership,  $\notin$ .

For universal set,  $\mathbb{U}$  i.e. everything.

For null set,  $\phi$ .

**Example:**

$B = \{x/0 \leq x \leq 1\}$  where  $/$  means such that.

$C = \{x/x \in \mathbf{R}, x^2 = -1\} = \phi$

**Definition:** If every element of  $A$  is an element of  $B$  then  $A$  is subset of  $B$ . i.e.  $A \subset B$ .

**Definition:**  $A = B$  iff (if and only if)  $A \subset B$  and  $B \subset A$ .

**Properties:**

- $\phi \subset A$  ;  $A \subset U$  ;  $A \subset A$
- $A \subset B, B \subset C \implies A \subset C$

**Remark:** The order in which the elements of set are listed is immaterial. E.g.  $\{a, b, c\} = \{b, c, a\}$ .

**Definition:** The complement of  $A$  with respect to  $U$  is  $A^c = \{x \mid x \in U \text{ and } x \notin A\}$ .

**Definition:** The intersection of  $A$  and  $B$  is  $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$ .

**Definition:** The union of  $A$  and  $B$  is  $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$ .

If  $A \cap B = \phi$ , then  $A$  and  $B$  are **disjoint** or **mutually exclusive**.

**Definition:**

- Minus:  $A - B = A \cap B^c$
- Symmetric difference or XOR:  $A \triangle B = (A - B) \cup (B - A) = (A \cup B) - (A \cap B)$

- The **cardinality** of  $A$ , denoted by  $|A|$  is the number of elements in  $A$ .  $A$  is finite if  $|A| < \infty$ .  
 $B = \{1, 2, 3, \dots\}$  is **countably infinite** i.e.  $|B| = \aleph_0$   
 $C = \{x | x \in [0, 1]\}$  is **uncountably infinite** i.e.  $|C| = \aleph_1$

#### Laws of Operation:

- **Complement Law:**  $A \cup A^c = U$ ,  $A \cap A^c = \phi$ ,  $(A^c)^c = A$
- **Commutative Law:**  $A \cup B = B \cup A$ ,  $A \cap B = B \cap A$
- **Associative Law:**  $A \cup (B \cup C) = (A \cup B) \cup C$ ,  $A \cap (B \cap C) = (A \cap B) \cap C$
- **Distributive Law:**  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ ,  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- **De-Morgan's Law:**  $(A \cup B)^c = A^c \cap B^c$ ,  $(A \cap B)^c = A^c \cup B^c$

## 1.2 Bootcamp: Derivative

**Definition:** The function  $f(x)$  maps values of  $X$  from a certain domain  $X$  to a certain range  $Y$  which can be denoted  $f : x \rightarrow Y$ .

If  $f(x) = x^2$  then the function takes  $x$ -values from the real line  $\mathbb{R}$  to the non-negative portion of real line  $\mathbb{R}^+$ .

**Definition:** We say that  $f(x)$  is **continuous** function if for any  $x_0$  &  $x \in X$ , we have  $\lim_{x \rightarrow 0} f(x) = f(x_0)$  where  $f(x)$  is assumed to exist for all  $x \in X$ .

The function  $f(x) = 3x^2$  is continuous for all  $x$ . The function  $f(x) = \lfloor x \rfloor$  i.e. round down to nearest integer e.g.  $\lfloor 3.4 \rfloor = 3$ . This function has discontinuity at any integer  $x$ .

**Definition:** The **inverse** of function  $f : X \rightarrow Y$  is reverse mapping of  $g : Y \rightarrow X$  such that  $f(x) = y$  iff  $g(y) = x$  for all appropriate  $x$  and  $y$ . The inverse is often written as  $f^{-1}$  and is especially useful if  $f(x)$  strictly increasing or decreasing function. Note that  $f^{-1}(f(x)) = x$ .

**Definition:** If  $f(x)$  is continuous, then it is **differentiable** if,

$$\frac{d}{dx} f(x) = f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

exists and is well defined for given  $x$ . The derivative of  $f(x)$  is slope of the function.

$$[x^k]' = kx^{k-1}$$

$$[e^x]' = e^x$$

$$[\sin(x)]' = \cos(x)$$

$$[\cos(x)]' = -\sin(x)$$

$$[\ln(x)]' = \frac{1}{x}$$

$$[\arctan(x)]' = \frac{1}{1+x^2}$$

**Theorem:** Some properties of derivatives

$$[af(x) + b]' = af'(x)$$

$$[f(x) + g(x)]' = f'(x) + g'(x)$$

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x)$$

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{g(x)f'(x) - f(x)g'(x)}{g^2(x)}$$

$$[f(g(x))]' = f'(g(x))g'(x)$$

**Remark:** The second derivative  $f''(x) = \frac{d}{dx}f'(x)$  and is the “slope of slope”. If  $f(x)$  is position, then  $f'(x)$  can be regarded as “velocity” and  $f''(x)$  as “acceleration”.

The minimum or maximum of  $f(x)$  can only occur when slope of  $f(x)$  is 0, i.e. only when  $f'(x) = 0$ , say at the critical point  $x = x_0$ . Exception: Check the endpoints of your intervals of interest as well.

If  $f''(x) < 0$ , you get maximum, if  $f''(x) > 0$ , you get a minimum. If  $f''(x) = 0$ , you get a **point of inflection**.

### 1.3 Bootcamp: Integration

**Definition:** The function  $F(x)$  having derivative  $f(x)$  is called the **anti-derivative** or **indefinite integral**. It is denoted by  $F(x) = \int f(x)dx$ .

**Fundamental Theorem of Calculus:** If  $f(x)$  is continuous, then the area under the curve for  $x \in [a, b]$  is denoted and given by the **definite integral**.

$$\int_a^b f(x)dx = F(x)|_a^b = F(b) - F(a)$$

$$\int x^k dx = \frac{x^{k+1}}{k+1} + c \quad \text{for } k \neq -1 \text{ where } c \text{ is arbitrary constant}$$

$$\int \frac{dx}{x} = \ln|x| + c$$

$$\int e^x dx = e^x + c$$

$$\int \cos(x)dx = \sin(x) + c$$

$$\int \frac{1}{1+x^2} dx = \arctan(x) + c$$

**Theorem:** Some well known properties of definite integrals

$$\int_a^a f(x)dx = 0$$

$$\int_a^b f(x)dx = - \int_b^a f(x)dx$$

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

**Theorem:** Some other properties of general integrals:

$$\int [f(x) + g(x)]dx = \int f(x)dx + \int g(x)dx$$



$$\int f(x)g'(x)dx = f(x)g(x) - \int g(x)f'(x)dx \quad \text{integration by parts}$$

$$\int f(g(x))g'(x)dx = \int f(u)du \quad \text{Substitution rule with } u = g(x)$$

**Definition:** Derivative of arbitrary order  $K$  can be written as  $f^k(x)$  or  $\frac{d^k}{dx^k}f(x)$ . By convention  $f^0(x) = f(x)$ .

The **Taylor Series Expansion** of  $f(x)$  about a point  $a$  is given by

$$f(x) = \sum_{k=0}^{\infty} \frac{f^k(a)(x-a)^k}{k!}$$

The **Maclaurin Series** is simply Taylor expanded around  $a = 0$ .

Some famous **Maclaurin Series**;

$$\sin(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$$

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Here are some miscellaneous sums:

$$\sum_{k=1}^n k = \frac{n(n+1)}{2}$$

$$\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{k=0}^{\infty} \frac{1}{1-p} \quad (\text{for } -1 < p < 1)$$

**Theorem:** Occasionally, we run into trouble when taking indeterminate ratios of form  $\frac{0}{0}$  or  $\frac{\infty}{\infty}$ . In such cases, **L' Hospital Rule** is useful. If the limits  $\lim_{x \rightarrow a} f(x)$  and  $\lim_{x \rightarrow a} g(x)$  both go to 0 or both go to  $\infty$ , then,

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}$$

Example:

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = \lim_{x \rightarrow 0} \frac{\cos(x)}{1} = 1$$

### Double Integration:

Whereas single integrals get us the area under a curve, double integrals represent the volume under a three dimensional function.

The volume under  $f(x, y) = 8xy$  over region  $0 < x < y < 1$  is given by

$$\int_0^1 \int_0^y f(x, y) dx dy = \int_0^1 \int_0^y 8xy dx dy = \int_0^1 4y^3 dy = 1$$

We can swap the order of integration to get same answer.

$$\int_0^1 \int_x^1 8xy dy dx = \int_0^1 4x(1 - x^2) dx = 1$$

## 2. Introduction to Probability

### 2.1 Introduction

Mathematical models are either

- Deterministic (no uncertainty/randomness)
- Probabilistic (have some uncertainty)

Q. A couple has two kids and at least one is boy. What is the probability that both are boys?

Possibilities: GG, BG, GB, BB. Eliminate GG since we know that there's at least one boy. Then  $P(BB) = \frac{1}{3}$ .

**Probability** is methodology that describes the random variation in systems. **Statistics** uses data (sample) to draw conclusion about population.

**Definition:** A **sample space** associated with an experiment E is the set of all possible outcome of E. It's usually denoted by  $S$  or  $\Omega$ .

Coin Toss:  $S = \{H, T\}$

Toss a coin 2 times:  $S : \{HH, HT, TH, TT\}$

**Definition:** An **event** is a set of possible outcomes. Thus, any subset of  $S$  is event.

Toss a dice,  $S = \{1, 2, \dots\}$

If  $A$  is event "odd number occurs",  $A = \{1, 3, 5\}$

The **empty set**  $\phi$  is an event of  $S$ .

$A$  is an event of  $S$ .

If  $A$  is an event, then  $A^c$  is the **complementary** event.

If  $A$  and  $B$  are events, then  $A \cup B$  and  $A \cap B$  are events.

**Definition:** The **Probability** of a generic event  $A \subset S$  is a function that adheres to following axioms:

- $0 \leq P(A) \leq 1$  (probabilities are always between 0 and 1)
- $P(S) = 1$  (probability of some outcome is 1)
- If  $A$  and  $B$  are disjoint events, i.e.  $A \cap B = \phi$  then,  $P(A \cup B) = P(A) + P(B)$ .
- Suppose  $A_1, A_2, \dots$  is a sequence of disjoint events, i.e.  $A_i \cap A_j = \phi$  for  $i \neq j$ .

$$\begin{aligned}
P(S) &= P(U_{i=1}^{\infty} A_i) \\
&= \sum_{i=1}^{\infty} P(A_i) \\
&= \sum_{i=1}^{\infty} \frac{1}{2^i}
\end{aligned}$$

**Theorem:**  $P(A^c) = 1 - P(A)$

**Proof:**

$$\begin{aligned}
1 &= P(S) \\
&= P(A \cup A^c) \\
&= P(A) + P(A^c) \quad \therefore A \cap A^c = \phi
\end{aligned}$$

**Corollary:**  $P(\phi) = 0$

**Proof:** By definition,  $\phi = S^c$ ; so the result follows the theorem and axiom 2. **Remark:** The converse is false:  $P(A) = 0$  doesn't imply  $A = \phi$ .

**Theorem:** For any two events  $A$  and  $B$ ,

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

**Proof:** Use Venn-diagram.

**Remark:** Axiom 3 is special case of this theorem with  $A \cap B = \phi$ .

**Theorem:** For any three events  $A$ ,  $B$  and  $C$ ,

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

**Theorem:** Here is the **Principle of inclusion-exclusion**:

$$\begin{aligned}
P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_{i=1}^n P(A_i) - \sum \sum_{i < j} P(A_i \cap A_j) + \sum \sum \sum_{i < j < k} P(A_i \cap A_j \cap A_k) \\
&\quad + \dots + (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n)
\end{aligned}$$

**Remark:** You “include” all of the “single” events, “exclude” the double events, include the triple events etc.

**Finite Sample Space:**

Suppose  $S$  is finite  $S = S_1, S_2, \dots, S_n$ . Finite sample space often allows us to calculate the probabilities of certain events more efficiently. To illustrate, let  $A \subset S$  be any event, then  $P(A) =$

$$\Sigma_{S_i \in A} P(S_i).$$

You have 2 red cards, a blue and a yellow card. Pick a card at random then,

$$S = \{S_1, S_2, S_3\} = \{red, blue, yellow\}$$

$$P(S_1) = \frac{1}{2} \quad P(S_2) = \frac{1}{4} \quad P(S_3) = \frac{1}{4}$$

$$P(\text{red or yellow}) = \frac{1}{2} + \frac{1}{4}$$

**Definition:** A **simple sample space (SSS)** is a finite sample space in which outcomes are equally likely.

**Remark:** In above example,  $S$  is not simple sample space since  $P(S_1) \neq P(S_2)$ .

**Example:** Toss 2 fair coins,

$S = \{HH, HT, TH, TT\}$  is a *SSS* (all probabilities are  $\frac{1}{4}$ ).

**Theorem:** For any event  $A$  in *SSS*,

$$P(A) = \frac{|A|}{|S|} = \frac{\text{no. of elements in } A}{\text{no. of elements in } S}$$

### 2.1.1 Counting Techniques

Muffin (blueberry or oatmeal) or a bagel (sesame, plain, salt, garlic) but not both. You have  $2 + 4 = 6$  choices in total.

$n_{AB} = 3$  ways to go from city A to B (walk, car, bus) and  $n_{BC} = 4$  ways to go from B to C (car, bus, train, plane). Then you can go from A to C (via B) using  $n_{AB} \cdot n_{BC} = 3 * 4 = 12$  ways.

Roll two dice. How many outcomes?

$(3, 2) \neq (2, 3)$  so, answer  $= 6 * 6 = 36$  ways.

Toss  $n$  dice. Outcome  $= 6^n$  possibilities.

Toss  $n$  coins. Outcome  $= 2^n$  possibilities.

### 2.1.2 Permutation

An arrangement of  $n$  symbols in a **definite order** is a **permutation** of  $n$  symbols.

Example: How many ways to arrange 1, 2, 3 ?

Answer: 6 ways: 123, 132, 213, 312, 321, 231

- \*\*Number of ways to arrange 1, 2, ...,  $n = n * (n - 1) * (n - 2) * \dots * 2 * 1 = n!$

**Definition:** The number of **r-tuples** we can make from  $n$  different symbols (each used at most once) is called the **number of permutations of  $n$  things taken  $r$  at a time**.

$$P_{n,r} = \frac{n!}{(n-r)!}$$

Note:  $0! = 1$  &  $P_{n,n} = n!$

**Proof:**

$$\begin{aligned} P_{n,r} &= (\text{choose first})(\text{choose second})\dots(\text{choose } r^{\text{th}}) \\ &= n(n-1)(n-2)\dots(n-r+1) \\ &= \frac{n(n-1)\dots(n-r+1)(n-r)\dots 2 * 1}{(n-r)\dots 2 * 1} \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

Example: How many license plates of 6 digits can be formed from numbers  $\{1,2,\dots,9\}$ ? + with no repetitions:  $P_{9,3} = 60480$  + with repetitions:  $9 * \dots * 9 = 9^6$  ways + containing repetitions:  $9^6 - 60480 = 470961$

### 2.1.3 Combination

How many subsets of  $\{1, 2, 3\}$  contain exactly 2 elements? (order isn't important)

Answer: 3 subsets -  $\{1, 2\}, \{1, 3\}, \{2, 3\}$

**Definition:** The number of subsets with  $r$  elements of a set with  $n$  elements is called **number of combinations of  $n$  things taken  $r$  at a time**.

**Notation:**  $C_{n,r}$  or  $\binom{n}{r}$ . These are also called **binomial coefficients**.

$$C_{n,r} = \frac{n!}{r!(n-r)!}$$

The difference between permutation and combination:

- Combination:  $(a, b, c) = (b, a, c)$  i.e. order doesn't concern,
- Permutation:  $(a, b, c) \neq (b, a, c)$  i.e. concerned with order.

Choosing a permutation is same as first choosing a combination and putting the elements in order.

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

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

Following results should be intuitive:

- $\binom{n}{r} = \binom{n}{n-r}$
- $\binom{n}{0} = \binom{n}{n} = 1$
- $\binom{n}{1} = \binom{n}{n-1} = n$

### 2.1.4 Binomial Theorem

$$(x + y)^n = \sum_{i=0}^n \binom{n}{i} x^i y^{n-i}$$

This is where **Pascal's triangle** comes from.

**Corollary:** Surprising fact

$$\sum_{i=0}^n \binom{n}{i} = 2^n$$

**Proof:** By the binomial theorem:

$$2^n = (1 + 1)^n$$

$$= \sum_{i=0}^n \binom{n}{i} 1^i 1^{n-i}$$

### 2.1.5 Problems

**Q.** Select 2 cards from a deck without replacement and care about order? Possibilities  
 $= 52 * 51 = 2652$  ways.

**Q.** Box of 10 sox - 2 red and 8 black. Pick 2 without replacement.

- Let  $A$  be event that both are red.

$$P(A) = \frac{\text{ways to pick 2 reds}}{\text{ways to pick 2 sox}} = \frac{2*1}{10*9} = \frac{1}{45}$$

- Let  $B$  be event that both are black.

$$P(B) = \frac{8*9}{10*9} = \frac{28}{45}$$

- Let  $C$  be one of each color. Since,  $A$  and  $B$  are disjoint,

$$P(C) = 1 - P(C^c) = 1 - P(A \cup B) = 1 - \frac{1}{45} - \frac{28}{45} = \frac{16}{45}$$

**Q. An NBA team has 12 players. How many ways can the coach choose the starting 5?**

$$\binom{12}{5} = \frac{12!}{5!7!} = 792$$

**Q. Smith is one of the players on the team. How many of 792 starting lineup include him?**

$$\binom{11}{4} = \frac{11!}{4!7!} = 330$$

**Q. 4 red marbles, 2 whites. Put them in random order.**

a.  $P(2 \text{ end marbles are W})$

$S = \{\text{Possible pairs of slots that W's occupy}\}$

$$|S| = \binom{6}{2} = \frac{6!}{2!(6-2)!} = 15$$

Since, W's must occupy end slots so,  $|A| = \binom{2}{2} = 1$

$$P(A) = \frac{|A|}{|S|} = \frac{1}{15}$$

b.  $P(2 \text{ end marbles aren't both W}) = 1 - P(A) = \frac{14}{15}$

c.  $P(2 \text{ W's are side by side})$

WRRRRR or RWWRRR or RRWWRR or RRRWWRR or RRRRWW

$$|B| = 5$$

$$P(B) = \frac{5}{15}$$



## 2.2 Hypergeometric Distribution

**Definition:** You have  $a$  objects of type 1 and  $b$  objects of type 2. Select  $n$  objects **without replacement** from  $a + b$  objects. Then,

$$P(k \text{ type 1's were picked}) = \frac{(\text{Number of ways to choose } k \text{ type 1's out of } a)(\text{Choose } n-k \text{ type 2's out of } b)}{(\text{Number of ways to choose } n \text{ out of } a+b)}$$

$$= \frac{\binom{a}{k} \binom{b}{n-k}}{\binom{a+b}{n}}$$

The number of type 1's chosen is said to have the **hypergeometric distribution**.

**Example:** 3 sox in box with  $a = 2$  red,  $b = 1$  blue. Pick  $n = 3$  without replacement.

$$P(\text{Exactly } k=2 \text{ reds are picked}) = \frac{\binom{a}{k} \binom{b}{n-k}}{\binom{a+b}{n}}$$

$$= \frac{\binom{2}{2} \binom{1}{1}}{\binom{3}{3}}$$

$$= 1$$

## 2.3 Binomial Distribution

**Definition:** You again have  $a$  objects of type 1 and  $b$  objects of type 2. Now, select  $n$  objects **with replacement** from  $a + b$  objects.

$$P(k \text{ type 1's were picked}) = (\text{Number of ways to choose } k \text{ 1's and } n-k \text{ 2's})$$

$$P(\text{Choose } k \text{ 1's in a row then } n-k \text{ 2's in a row})$$

$$P(k \text{ type 1's were picked}) = \binom{n}{k} \left( \frac{a}{a+b} \right)^k \left( \frac{b}{a+b} \right)^{n-k}$$

## 2.4 Multinomial Coefficients

**Example:**  $n_1$  blue sox,  $n_2$  reds. The number of assortments is  $\binom{n_1+n_2}{n_1}$ . Generalization for  $k$  types of objects:  $n = \sum_{i=1}^k n_i$  The number of arrangements is

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1! n_2! \dots n_k!}$$

This is known as **multinomial coefficient**.

**Example:** How many ways letters in “MISSISSIPPI” be arranged?

$$\frac{\text{Number of permutations of 11 letters}}{(\text{Number of M's})(\text{Number of P's})(\text{Number of I's})(\text{Number of S's})}$$

$$= \frac{11!}{1!2!4!4!}$$

## 2.5 Conditional Probability

The probability of A occurs given B occurs is

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

**Definition:** If  $P(B) > 0$ , the conditional probability of A given B is

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

**Remark:** If A and B are disjoint, then  $P(A/B) = 0$ . If B occurs, there is no chance that A can occur.

What happens if  $P(B) = 0$ ? In that case, no need to consider  $P(A/B)$ .

**Example:** Toss 2 dice and take the sum.

A: odd toss = {3, 5, 7, 9, 11}

B: {2, 3}

$$P(A) = P(3) + \dots + P(11) = \frac{2}{36} + \frac{4}{36} + \dots + \frac{2}{36} = \frac{1}{2}$$

$$P(B) = \frac{1}{36} + \frac{2}{36} = \frac{1}{12}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{36}}{\frac{1}{12}} = \frac{2}{3}$$

**Example:** A couple has two kids and at least one is boy. What's the probability that both are boys?

$S = \{GG, GB, BG, BB\}$

$C$  : Both are boys = {BB}

$D$  : At least 1 boy = {GB, BG, BB}

$$P(C/D) = \frac{P(C \cap D)}{P(D)} = \frac{P(C)}{P(D)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

**Example:** A couple has two kids and at least one is born on tuesday. What is the probability that both are boys?

$$B_x[G_x] = Boy[Girl]$$

born on day x;  $x = 1, 2, \dots, 7$

$x = 3$  is Tuesday.

$$S = \{(G_x, G_y), (G_x, B_y), (B_x, G_y), (B_x, B_y), x, y = 1, 2, \dots, 7\}$$

$$\text{So, } |S| = 4 * 49 = 196$$

i.e. 4 combination of B and G and 49 combination of x and y.

C: Both are boys (with at least one born on tuesday)

$$= \{(B_x, B_3), x = 1, 2, \dots, 7\} \cup \{(B_3, B_y), y = 1, 2, \dots, 7\}$$

Note:  $|C| = 13$  {to avoid double counting  $(B_3, B_3)$ }

D: There is at least one boy born on Tuesday.

$$= C \cup \{(G_x, B_3), (B_3, G_y), x, y = 1, 2, \dots, 7\}$$

$$|D| = 27$$

$$P(C/D) = \frac{P(C \cap D)}{P(D)} = \frac{P(C)}{P(D)} = \frac{\frac{13}{196}}{\frac{27}{197}} = \frac{13}{27}$$

**Properties:** Analogous to axioms of probability

- $0 \leq P(A/B) \leq 1$
- $P(S/B) = 1$
- $A_1 \cap A_2 = \phi \rightarrow P(A_1 \cap A_2/B) = P(A_1/B) + P(A_2/B)$
- If  $A_1, A_2, \dots$  are all disjoint then

$$P(\bigcup_{i=1}^{\infty} A_i/B) = \sum_{i=1}^{\infty} P(A_i/B)$$

## 2.6 Independence

Any unrelated events are independent.

### Example:

A: It rains on Mars tomorrow.

B: Coin lands on H.

**Definition:** A & B are independent iff  $P(A \cap B) = P(A).P(B)$

**Remark:** If  $P(A) = 0$ , then  $A$  is independent of any other event.

**Remark:** Events don't have to be physically unrelated to be independent.

**Theorem:** Suppose  $P(B) > 0$ . Then  $A$  and  $B$  are independent  $\leftrightarrow P(A/B) = P(A)$ .

**Proof:** A & B independent  $\leftrightarrow P(A \cap B) = P(A).P(B) \leftrightarrow \frac{P(A \cap B)}{P(B)} = P(A)$

**Remark:** So, if  $A$  and  $B$  are independent, the probability of  $A$  doesn't depend on whether or not  $B$  occurs.

**Bayes Theorem:**  $A$  and  $B$  are independent  $\leftrightarrow A'$  and  $B'$  are also independent.

**Proof:** Only need to prove in  $\rightarrow$  direction (then  $\leftarrow$  follows trivially).

$$P(A) = P(A \cap B') + P(A \cap B)$$

So,

$$\begin{aligned} P(A \cap B') &= P(A) - P(A \cap B) \\ &= P(A) - P(A).P(B) \quad \{A, B \text{ are independent}\} \\ &= P(A)\{1 - P(B)\} \\ &= P(A).P(B') \end{aligned}$$

### Don't confuse independence with disjointness!

**Theorem:** If  $P(A) > 0$  and  $P(B) > 0$ ,  $A$  and  $B$  can't be independent and disjoint at the same time.

**Proof:** Suppose  $A$  and  $B$  are disjoint,  $A \cap B = \phi$ . Then,  $P(A \cap B) = 0 < P(A).P(B)$ . Thus,  $A$  and  $B$  aren't independent. Similarly, independent doesn't imply disjoint.

**Remark:** In fact, independence and disjointness are almost opposite. If  $A$  and  $B$  are disjoint and  $A$  occurs, then you have information that  $B$  cannot occur. So,  $A$  and  $B$  can't be independent.

Extension to more than two events:

**Definition:** A, B, C are independent iff

- $P(A \cap B \cap C) = P(A).P(B).P(C)$

- All pairs are independent:

$$P(A \cap B) = P(A).P(B)$$

$$P(A \cap C) = P(A).P(C)$$

$$P(B \cap C) = P(B).P(C)$$

**General Definition:**  $A_1, \dots, A_k$  are independent iff  $P(A_1 \cap \dots \cap A_k) = P(A_k)$  and all subsets of  $\{A_1, \dots, A_k\}$  are independent.

**Independent Trials:** Perform  $n$  trials of an experiment such that the outcome of one trial is independent of outcomes of other trials. Eg. Flip 3 coins independently.

**Remark:** For independent trials, you just multiply the individual probabilities.

Eg. Flip a coin infinitely many times (each flip is independent of others).

$$\begin{aligned}
 P_n &= P(\text{First H on } n\text{th trial}) \\
 &= P(\underbrace{TT\dots T}_{n-1} H) \\
 &= \underbrace{P(T).P(T)\dots P(T)}_{n-1}.P(H) \\
 &= \left(\frac{1}{2}\right)^{n-1} \cdot \frac{1}{2} = \frac{1}{2^n} \\
 &= \frac{1}{2^n} \quad \{\text{Each has probability } 1/2\} \\
 P(H \text{ eventually}) &= \sum_{n=1}^{\infty} P_n \\
 &= \sum_{n=1}^{\infty} 2^{-n} \\
 &= 1
 \end{aligned}$$

## 2.7 Partitions and laws of probability

**Partition of Sample Space** split the sample space into disjoint, yet all encompassing subsets.

**Definition:** The events  $A_1, A_2, \dots, A_n$  form a partition of sample space  $S$  if

- $A_1, A_2, \dots, A_n$  are disjoint.
- $\bigcup_{i=1}^n A_i = S$
- $P(A_i) > 0$  for all  $i$ .

**Remark:** When an experiment is performed, exactly one  $A_i$ 's occur.

**Example:**  $A$  and  $A'$  form partition.

Suppose  $A_1, A_2, \dots, A_n$  form partition of  $S$  and  $B$  is arbitrary event. Then,

$$B = \bigcup_{i=1}^n (A_i \cap B)$$

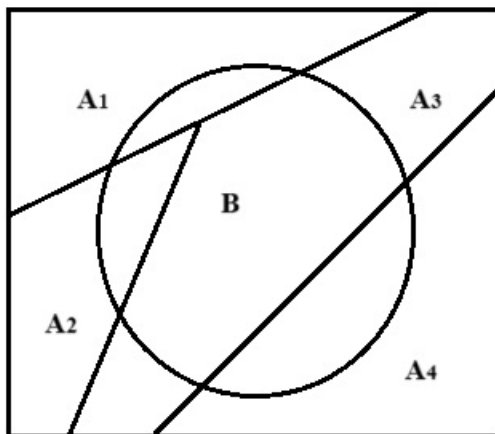


Figure 1: Partitions

$$\begin{aligned} P(B) &= P[\bigcup_{i=1}^n (A_i \cap B)] \\ &= \sum_{i=1}^n P(A_i \cap B) \quad (\text{Since } A_1, A_2, \dots, A_n \text{ are disjoint}) \\ &= \sum_{i=1}^n P(A_i)P(B/A_i) \quad (\text{Definition of conditional Probability}) \end{aligned}$$

This is **law of probability**.

**Example:** Suppose we have 10 Georgia Tech students and 20 University of Georgia students taking a test. GT students have 95% chance of passing but UGA have 50%. Determine probability that he/she passes.

$$P(\text{passes}) = P(GT)P(\text{passes}/GT) + P(UGA)P(\text{passes}/UGA)$$

## 2.8 Bayes Theorem

Immediate consequence of law of total probability.

**Bayes Theorem:** If  $A_1, A_2, \dots, A_n$  form partition of  $S$  and  $B$  is any event then,

$$\begin{aligned} P(A_j/B) &= \frac{P(A_j \cap B)}{P(B)} \\ &= \frac{P(A_j)P(B/A_j)}{\sum_{i=1}^n P(B/A_i)} \end{aligned}$$

The  $P(A_j)$ 's are prior probabilities ("before B").

The  $P(A_j/B)$ 's are posterior probabilities ("after B").

The  $P(A_j/B)$ 's add up to 1.

## 2.9 Probability Problems

### Birthday Problem

**Q. There are  $n$  people in room. Find the probability that at least two have the same birthday. (Ignore Feb 29 and assume that all 365 days have equal probability.**

The (simple) sample size is  $S = \{(x_1, \dots, x_n) : x_i \in \{1, 2, \dots, 365\}, V_i\}$

( $x_i$  is person  $i$ 's birthday) and note that  $|S| = (365)^n$ .

Let A: All birthdays are different then,

$$\begin{aligned} P(A) &= \frac{(365)(364)\dots(365 - n + 1)}{365^n} \\ &= 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \dots \frac{365 - n + 1}{365} \end{aligned}$$

$$P(A') = 1 - P(A)$$

When,  $n = 366$ ,  $P(A') = 1$

For,  $P(A') > \frac{1}{2}$ ,  $n$  must be  $\geq 23$ .

When,  $n = 50$ ,  $P(A') = 0.97$ ,  $P(A')$  is probability of at least one birthday match (not unique).

### The Envelope Problem

**Q. A group of  $n$  people receives  $n$  envelopes with their name on them but someone has completely mixed up the envelopes. Find the probability that at least one person will receive the proper envelope.**

Let  $A_i$ : Person  $i$  receive correct envelope.

We want  $P(A_1 \cup A_2 \dots \cup A_n)$

By principle of Inclusion-Exclusion,

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i) - \sum \sum_{i < j} P(A_i \cap A_j) + \sum \sum \sum_{i < j < k} P(A_i \cap A_j \cap A_k) \\ + \dots + (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n)$$

Since all  $P(A_i)$ 's are same, all of  $P(A_i \cap A_j)$ 's are the same.

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = nP(A_1) - \binom{n}{2}P(A_1 \cap A_2) + \binom{n}{3}P(A_1 \cap A_2 \cap A_3) + \dots + (-1)^{n-1}P(A_1 \cap A_2 \dots \cap A_n)$$

$$\begin{aligned} P(A_1) &= \frac{1}{n} \\ P(A_2) &= \frac{1}{n-1} \\ P(A_1 \cap A_2) &= \frac{1}{n(n-1)} \\ P(A_1 \cup A_2 \cup \dots \cup A_n) &= \frac{n}{n} - \binom{n}{2} \frac{1}{n} \cdot \frac{1}{n-1} + \binom{n}{3} \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} + \dots + (-1)^{n-1} \frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} + \dots + (-1)^{n-1} \frac{1}{n!} \\ &= 1 - \frac{1}{e} \quad \{\text{Very similar to Mclaurin Series}\} \\ &= 0.6321 \end{aligned}$$

If  $n = 4$  envelopes:

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup A_4) &= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} \\ &= 0.625 \end{aligned}$$



### 3. Random Variables

#### 3.1 Univariate Random Variables

**Definition:** A **random variable (RV)** is a function from the sample space to the real line.  $X : S \rightarrow \mathbb{R}$ .

**Example:** Flip 2 coins:  $S = \{HH, HT, TH, TT\}$

Suppose  $X$  is RV corresponding to the number of  $H$ 's,

$$X(TT) = 0, X(HT) = 1, X(HH) = 2$$

$$P(X = 0) = \frac{1}{4}, P(X = 1) = \frac{1}{2}, P(X = 2) = \frac{1}{4}$$

**Notation:** Capital letters like  $X, Y, Z$  usually represent RV's. Small letters like  $x, y, z$  represent particular values of RV's.

**Example:** Flip a coin

$$X = \begin{cases} 1 & \text{if T} \\ 0 & \text{if H} \end{cases}$$

Roll a die

$$Y = \begin{cases} 0 & \text{if } \{1,2,3\} \\ 1 & \text{if } \{4,5,6\} \end{cases}$$

For our purpose,  $X$  and  $Y$  are same, since  $P(X = 0) = P(Y = 0) = \frac{1}{2}$  and  $P(X = 1) = P(Y = 1) = \frac{1}{2}$ .

**Example:** Select a real number at random between 0 and 1. There are infinite number of "equally likely" outcome.

Conclusion:  $P(\text{we choose the individual point } x) = P(X = x) = 0$ .

But  $P(X \leq 0.65) = 0.65$  and  $P(X \in [0.3, 0.7]) = 0.4$ .

If  $A$  is an interval in  $[0, 1]$  then  $P(X \in A)$  is the length of  $A$ .

**Definition:** If a number of possible values of a RV  $X$  is finite or countably infinite then  $X$  is **discrete** RV otherwise,

A **continuous** RV is one with probability 0 at every point.

**Example:**

- Flip a coin - get H or T. Discrete
- Pick a point at random in  $[0, 1]$ . Continuous
- The amount of time you wait in line is either 0 (with positive probability) or some positive real number - a combined discrete - continuous RV.

### 3.1.1 Discrete Random Variable

**Definition:** If  $X$  is discrete RV, its **probability mass function (pmf)** is

$$f(x) = P(X = x)$$

Note that  $0 \leq f(x) \leq 1$ ,  $\sum_x f(x) = 1$

**Example:** Flip 2 coins. Let  $X$  be number of heads.

$$f(x) = \begin{cases} \frac{1}{4} & \text{if } x = 0 \text{ or } 2 \\ \frac{1}{2} & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

**3.1.1.2 Uniform Distribution:** Uniform distribution of integers  $1, 2, \dots, n$ .  $X$  can equal  $1, 2, \dots, n$  each with probability  $\frac{1}{n}$ .

$$f(i) = \frac{1}{n} \quad i = 1, 2, \dots, n$$

**3.1.1.2 Binomial Distribution:** Let  $X$  denote number of “successes” from  $n$  independent trials such that  $P$  (success) at each trial is  $p$  ( $0 \leq p \leq 1$ ). Then  $X$  has the binomial distribution with parameters  $n$  and  $p$ . The trials are referred to as **Bernoulli Trials**.

**Notation:**  $X \sim \text{Bern}(n, p)$

**Example:** Roll a die 3 independent times. Find  $P$ (Get exactly two 6’s)

“success (6)” and “failure” (1,2,3,4,5)

All trials are independent,  $P(\text{success}) = \frac{1}{6}$  doesn’t change from trial to trial.

Let  $X$  = number of 6’s. Then  $X \sim \text{Bern}(3, \frac{1}{6})$ .

**Theorem:** If  $X \sim \text{Bern}(n, p)$  then probability of  $k$  successes in  $n$  trials is

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

where,

$$K = 0, 1, \dots, n \quad \text{and} \quad q = 1 - p$$

**Proof:** Consider the particular sequence of success and failures.

$$\begin{array}{c} \underline{SS\dots S} \quad \underline{FF\dots F} \\ k \text{ success} \quad n-k \text{ failure} \\ \text{prob} = p^k q^{n-k} \end{array}$$

The number of ways to arrange the sequence is  $\binom{n}{k}$ .

**Example:** Roll 2 dice and get the sum. Repeat 12 times. Find P(Sum will be 7 or 11 exactly 3 times).

$$\begin{aligned} P(7 \text{ or } 11) &= P(7) + P(11) \\ &= \frac{7}{36} + \frac{2}{36} \\ &= \frac{2}{9} \end{aligned}$$

So,  $X \sim \text{Bin}(12, \frac{2}{9})$  then,

$$P(X = 3) = \binom{12}{3} \left(\frac{2}{9}\right)^3 \left(\frac{7}{9}\right)^9$$

**3.1.1.3 Poisson Distribution:** If  $P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$ ,  $k = 0, 1, 2, \dots, \lambda$ ,  $\lambda > 0$ , we say that X has the **Poisson distribution** with parameter  $\lambda$ .

**Notation:**  $X \sim \text{Pois}(\lambda)$

**Example:** Suppose the number of raisins in a cup of cookie dough is  $\text{Pois}(10)$ . Find the probability that cup of dough has at least 4 raisins.

$$\begin{aligned} P(X \geq 4) &= 1 - P(X = 0, 1, 2, 3) \\ &= 1 - e^{-10} \left( \frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!} \right) \\ &= 0.9897 \end{aligned}$$

### 3.1.2 Continuous Random Variables

**Example:** Pick a point X randomly between 0 and 1 and define the continuous function.

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

Intuitively, if  $0 \leq a \leq b \leq 1$  then,

$$\begin{aligned} P(a \leq X \leq b) &= \text{the "area" under } f(x) \text{ from } a \text{ to } b \\ &= b - a \end{aligned}$$

**Definition:** Suppose  $X$  is a continuous RV, the magic function  $f(x)$  is **probability density function (PDF)** if

- $\int_{\mathbb{R}} f(x)dx = 1$  (area under the  $f(x)$  is 1)
- $f(x) \geq 0$
- If  $A \subseteq \mathbb{R}$ , then  $P(X \in A) = \int_A f(x)dx$  (probability that  $X$  is in a certain region of  $A$ )

**Remark:** If  $X$  is continuous RV then,

$$P(a < X < b) = \int_a^b f(x)dx$$

An individual point has probability 0 i.e.  $P(x = x) = 0$ .

If  $X$  is discrete then  $f(x) = P(X = x)$  and must have  $0 \leq f(x) \leq 1$ .

If  $X$  is continuous,

- $f(x)$  is continuous,
- Instead think of  $f(x)dx \approx P(x < X < x + dx)$ .
- Must have  $f(x) \geq 0$  and possibly  $> 1$ .

**3.1.2.1 Uniform Distribution** If  $X$  is “equally likely” to be anywhere between  $a$  and  $b$  then  $X$  has the uniform distribution on  $(a, b)$ .

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

**Notation:**  $X \sim Unif(a, b)$

**Remark:**  $\int_{\mathbb{R}} f(x)dx = \int_a^b \frac{1}{b-a}dx = 1$

**Example:** If  $X \sim Unif(-2, 8)$  then,

$$P(-1 < X < 6) = \int_{-1}^6 \frac{1}{8 - (-2)}dx = 0.7$$

**3.1.2.2 Exponential Distribution**  $X$  has the exponential distribution with parameter  $\lambda > 0$  if it has PDF  $f(x) = \lambda e^{-\lambda x}$ , for  $x \geq 0$ .

**Notation:**  $X \sim Exp(\lambda)$

**Remark:**  $\int_{\mathbb{R}} f(x)dx = \int_0^{\infty} \lambda e^{-\lambda x}dx = 1$

**Example:** Suppose  $X$  is a continuous RV with PDF

$$f(x) = \begin{cases} cx^2 & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

First of all, let's find c. Noting that PDF must integrate to 1 we have,

$$1 = \int_{\mathbb{R}} f(x) dx$$

$$1 = \int_0^2 cx^2 dx$$

$$1 = \left. \frac{cx^3}{3} \right|_0^2$$

$$1 = \frac{8c}{3}$$

$$c = \frac{3}{8}$$

this means

$$f(x) = \frac{3x^2}{8}$$

$$\begin{aligned} P(0 < X < 1 \mid \tfrac{1}{2} < X < \tfrac{3}{2}) &= \frac{P(0 < X < 1 \cap \tfrac{1}{2} < X < \tfrac{3}{2})}{P(\tfrac{1}{2} < X < \tfrac{3}{2})} \\ &= \frac{P(\tfrac{1}{2} < X < 1)}{P(\tfrac{1}{2} < X < \tfrac{3}{2})} \\ &= \frac{\int_{\frac{1}{2}}^1 \frac{3}{8} x^2 dx}{\int_{\frac{1}{2}}^{\frac{3}{2}} \frac{3}{8} x^2 dx} \\ &= \frac{7}{26} \end{aligned}$$

X has the **standard normal distribution** if its PDF is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{for all } x \in \mathbb{R}$$

### 3.1.3 Cumulative Probability Distribution

**Definition:** For any RV (discrete or continuous), the cumulative distribution function (cdf) is defined for all x by,

$$F(x) = P(X \leq x)$$

For  $X$  discrete,

$$F(x) = \sum_{\{y/y \leq x\}} f(y) = \sum_{\{y/y \leq x\}} P(X = y)$$

For  $X$  continuous,

$$F(x) = \int_{-\infty}^x f(y) dy$$

**Example:** Flip a coin twice. let  $X$  = number of H's.

$$X = \begin{cases} 0 \text{ or } 2 \text{ with prob } \frac{1}{4} \\ 1 \text{ with prob } \frac{1}{2} \end{cases}$$

The CDF is a step function

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{4} & \text{if } 0 \leq x < 1 \\ \frac{3}{4} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

#### Explanation

$X$  defined as number of heads from two independent flips of a fair coin.  $X = 0$  (no heads),  $X = 1$  (one head) and  $X = 2$  (two heads). The probability distribution of  $X$  follows binomial distribution i.e.  $X \sim \text{Binom}(2, \frac{1}{2})$ .

$$P(X = 0) = \binom{2}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

$$P(X = 1) = \binom{2}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^1 = \frac{1}{2}$$

$$P(X = 2) = \binom{2}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^0 = \frac{1}{4}$$

For  $x < 0$ ,  $F(x) = 0$ .

For  $0 \leq x < 1$ ,  $F(x) = P(X = 0) = \frac{1}{4}$ .

For  $1 \leq x < 2$ ,  $F(x) = P(X = 0) + P(X = 1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$

For  $x \geq 2$ ,  $F(x) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$

**Warning:** For discrete RVs, be careful about  $\leq$  vs  $<$  as the endpoints of the range (where step function jumps).

**Theorem (Continuous CDF):** If  $X$  is continuous RV, then  $f(x) = F'(x)$  (assuming the derivative exists.)

**Proof:**

$$F'(x) = \frac{d}{dx} \int_{-\infty}^x f(t)dt = f(x), \quad \text{by the fundamental theorem of calculus}$$

**Example:**  $X \sim Unif(0, 1)$ . The PDF and cdf are

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

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Explanation

$$f(x) = \frac{1}{1-0} = 1 \quad \text{for } 0 < x < 1$$

So,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

For  $x \leq 0$ ,  $F(x) = 0$  since all values are between 0 and 1.

For  $0 < x < 1$ ,  $F(x) = \int_{-\infty}^x 1dt = x$ .

For  $x \geq 1$ ,  $F(x) = 1$  since  $x \geq 1$  includes all the probability.

**Example:**  $X \sim Exp(\lambda)$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$F(x) = \int_{-\infty}^x f(t)dt = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

We can use CDF to find **median** of X that is the point m such that,

$$0.5 = F(m) = 1 - e^{-\lambda m} \implies m = \left(\frac{1}{\lambda}\right) \ln(2)$$

#### Explanation

For,  $x \leq 0$ ,  $f(x) = 0$ . So,  $F(x) = 0$ .

For,  $x > 0$ ,  $F(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$

### Properties of CDF

$F(x)$  is non-decreasing in  $x$  i.e.  $a < b$  implies  $F(a) \leq F(b)$ .

$$\lim_{x \rightarrow \infty} F(x) = 1 \quad \lim_{x \rightarrow -\infty} F(x) = 0$$

$F(x)$  is right continuous at every point  $x$ .

**Theorem:**  $P(X > x) = 1 - F(x)$

**Proof:**

By complements,

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x)$$

**Theorem:**  $a < b \implies P(a < X \leq b) = F(b) - F(a)$ .

**Proof:** Since  $a < b$ , we have,

$$\begin{aligned} P(a < X \leq b) &= P(X > a, X \leq b) \\ &= P(X > a) + P(X \leq b) - P(X > a \cup X \leq b) \\ &= 1 - F(a) + F(b) - 1 \\ &= F(b) - F(a) \end{aligned}$$

where,

$$P(X > a) = 1 - F(a) \quad P(X \leq b) = F(b) \quad P(X > a \cup X \leq b) = 1$$

### 3.1.4 Great Expectations

**Definition:** The **mean** or **expected value** or **average** of random variable  $X$  is

$$\mu = E(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

The mean gives an indication of RV's central tendency. Think of it as a weighted average of the possible  $x$ 's where the weights are given by  $f(x)$ .



**Example:** Suppose  $X$  has the Bernoulli distribution with parameter  $p$  i.e. ( $P(X = 1) = p$  and  $P(X = 0) = q = 1 - p$ ). Then,

$$E(x) = \sum_x x f(x) = 1.p + 0.q = p$$

**Example:** Die toss.  $X = 1, 2, \dots, 6$  each with probability  $\frac{1}{6}$ . Then,

$$E(x) = \sum_x x f(x) = 1.\frac{1}{6} + \dots + 6.\frac{1}{6} = 3.5$$

Suppose  $X$  has the **geometric distribution** with the parameter  $p$  i.e.  $x$  is the number of Bern( $p$ ) trials until you obtain your first success (e.g.  $FFFS$  gives  $x = 4$ ).

$$f(x) = (1 - p)^{x-1}p, \quad x = 1, 2, \dots$$

**Notation:**  $X \sim \text{Geom}(p)$

Suppose I take independent foul shots but the chance of making any particular shot is only 0.4. What's the probability that it will take me at least 3 tries to make successful shot?

The number of tries until my first success is  $X \sim \text{Geom}(0.4)$ . Thus,

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - P(X = 1) - P(X = 2) \\ &= 1 - 0.4 - 0.6 * 0.4 \\ &= 0.36 \end{aligned}$$

Now, let's find the **expected value** of  $X \sim \text{Geom}(p)$

$$\begin{aligned}
E(X) &= \sum_x x f(x) \\
&= \sum_{x=1}^{\infty} x q^{x-1} p \quad (\text{where } q = 1 - p) \\
&= p \sum_{x=1}^{\infty} \frac{d}{dq} q^x \\
&= p \frac{d}{dq} \sum_{x=1}^{\infty} q^x \quad (\text{swap derivative and sum}) \\
&= p \frac{d}{dq} \frac{q}{1-q} \quad (\text{geometric sum}) \\
&= p \left\{ \frac{(1-q) - q(-1)}{(1-q)^2} \right\} \\
&= \frac{1}{p}
\end{aligned}$$

**Example:**  $X \sim \text{Exp}(\lambda)$ .  $f(x) = \lambda e^{-\lambda x}$ ,  $x \geq 0$ . Then,

$$\begin{aligned}
E(X) &= \int_{\mathbb{R}} x f(x) dx \\
&= \int_0^{\infty} x \lambda e^{-\lambda x} dx \\
&= -x e^{-\lambda x} \Big|_0^{\infty} - \int_0^{\infty} (-e^{-\lambda x}) dx \quad (\text{by parts}) \\
&= \int_0^{\infty} e^{-\lambda x} dx \quad (\text{L' H\^opital rule}) \\
&= \frac{1}{\lambda}
\end{aligned}$$

### 3.1.4 Law of the unconscious statistician (LOTUS)

**Theorem:** The expected value of a function of  $X$ , say  $h(x)$  is,

$$E[h(X)] = \begin{cases} \sum_x h(x) f(x) & \text{if } X \text{ is discrete,} \\ \int_{\mathbb{R}} h(x) f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

$E[h(x)]$  is weighted function of  $h(x)$  where the weights are  $f(x)$ 's.

**Remark:** It looks like a definition, but it's really a theorem - that's why they call it LOTUS.

**Example:**  $E[\sin x] = \int_{\mathbb{R}} \sin x f(x) dx$

**Definition:** The  $k^{th}$  moment is

$$E[X^k] = \begin{cases} \sum_x x^k f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x^k f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

**Example:** Suppose  $X \sim \text{Bern}(p)$  so that  $f(1) = p$  and  $f(0) = q$ .

$$E[X^k] = \sum_x x^k f(x) = 0^k q + 1^k p = p \quad \text{for all } k!$$

**Example:** Suppose  $X \sim \text{Exp}(\lambda)$ .  $f(x) = \lambda e^{-\lambda x}$ ,  $x \geq 0$  then,

$$\begin{aligned} E[X^k] &= \int_{\mathbb{R}} x^k f(x) dx \\ &= \int_0^{\infty} x^k \lambda e^{-\lambda x} dx \\ &= \int_0^{\infty} \left(\frac{y}{\lambda}\right)^k \lambda e^{-\lambda \frac{y}{\lambda}} \frac{1}{\lambda} dy \quad (\text{substitute } y = \lambda x) \\ &= \frac{1}{\lambda^k} \int_0^{\infty} y^{k+1-1} e^{-y} dy \\ &= \frac{\Gamma(k+1)}{\lambda^k} \quad (\text{by definition of gamma function}) \\ &= \frac{k!}{\lambda^k} \end{aligned}$$

**Definition:** The  $k^{th}$  **central moment** of  $X$  is

$$E[(X - \mu)^k] = \begin{cases} \sum_x (x - \mu)^k f(x) & X \text{ is discrete} \\ \int_{\mathbb{R}} (x - \mu)^k f(x) dx & X \text{ is continuous} \end{cases}$$

**Definition:** The **variance** of  $X$  is the second central moment i.e. the expected squared deviation of  $X$  from its mean.

$$\text{Var}(X) = E[(X - \mu)^2]$$

**Notation:**  $\sigma^2 = \text{Var}(X)$

**Definition:** The **standard deviation** of  $X$  is  $\sigma = +\sqrt{\text{var}(x)}$

**Example:**  $X \sim \text{Bern}(p)$  so that  $f(1) = p$ ,  $f(0) = q = 1 - p$

$$\mu = E[X] = p \quad \text{then,}$$

$$\begin{aligned}
Var[X] &= E[(X - \mu)^2] \\
&= \sum_x (x - p)^2 P(X = x) \\
&= (0 - p)^2 \cdot q + (1 - p)^2 \cdot p \\
&= p^2 q + q^2 p \\
&= pq(p + q) \\
&= pq \quad (\text{since } p + q = 1)
\end{aligned}$$

**Theorem:** For any  $h(x)$  and constants  $a$  and  $b$  - “shift happens”,

$$E[ah(X) + b] = aE[h(X)] + b$$

**Proof:**

$$\begin{aligned}
E[ah(X) + b] &= \int_{\mathbb{R}} (ah(x) + b)f(x)dx \\
&= a \int_{\mathbb{R}} h(x)f(x)dx + b \int_{\mathbb{R}} f(x)dx \\
&= aE[h(x)] + b
\end{aligned}$$

**Corollary:** In particular,

$$E[aX + b] = aE[X] + b$$

**Theorem (Easier way to calculate variance):**

$$Var(X) = E[X^2] - (E[X])^2$$

**Proof:**

$$\begin{aligned}
Var(X) &= E[(X - \mu)^2] \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - \mu^2 \quad \text{where } E[X] = \mu
\end{aligned}$$

**Example:** Suppose  $X \sim Bern(p)$ . Recall that  $E[X^k] = p$  for all  $k = 1, 2, \dots$ . Then,

$$\begin{aligned}
Var[X] &= E[X^2] - (E[X])^2 \\
&= p - p^2 \\
&= p \cdot q
\end{aligned}$$

**Example:**  $X \sim Unif(a, b)$ .  $f(x) = \frac{1}{b-a}$ ,  $a < x < b$  then,

$$E[X] = \int_{\mathbb{R}} f(x)dx = \int_a^b \frac{x}{b-a}dx = \frac{a+b}{2}$$

$$E[X^2] = \int_{\mathbb{R}} x f(x) dx = \int_a^b \frac{x^2}{b-a} dx = \frac{a^2 + ab + b^2}{3}$$

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

**Theorem:** Variance doesn't put up with shift b.

$$Var(aX + b) = a^2 \cdot Var(X)$$

**Proof:**

$$\begin{aligned} Var(aX + b) &= E[(aX + b)^2] - (E[aX + b])^2 \\ &= E[a^2 X^2 + b^2 + 2abX] - (aE[X] + b)^2 \\ &= a^2 E[X^2] + b^2 + 2abE[X] - a^2(E[X])^2 - 2abE[X] - b^2 \\ &= a^2 \{E[X^2] - (E[X])^2\} \\ &= a^2 Var(X) \end{aligned}$$

**Example:**  $X \sim Bern(0.3)$

$$E[X] = p = 0.3$$

$$Var[X] = pq = 0.3 * 0.7 = 0.21$$

Let

$$Y = h(x) = 4x + 5 \text{ then,}$$

$$E[Y] = E[4X + 5] = 4E[X] + 5 = 6.2$$

$$Var[Y] = Var[4X + 5] = 16Var[X] = 3.36$$

**Approximations to  $E[h(x)]$  and  $Var[h(x)]$**

Sometimes  $Y = h(x)$  is messy and we may have to approximate  $E[h(x)]$  and  $Var[h(x)]$  via a Taylor series approach. Let  $\mu = E[X]$  and  $\sigma^2 = Var(X)$  and note that

$$Y = h(\mu) + (X - \mu) \cdot h'(\mu) + \frac{(X - \mu)^2}{2} \cdot h''(\mu) + R$$

where, R is remainder term that we will ignore. Then,

$$E[Y] = h(\mu) + E[X - \mu] \cdot h'(\mu) + \frac{E[(X - \mu)^2]}{2} \cdot h''(\mu) = h(\mu) + \frac{h''(\mu)\sigma^2}{2}$$

and (now an even-crude approximation)

$$Var(Y) = Var[h(\mu) + (X - \mu) \cdot h'(\mu)] = [h'(\mu)]^2 \sigma^2$$

**Example:** Suppose  $X$  has pdf  $f(x) = 3x^2$ ,  $0 \leq x \leq 1$  and we want to test out our approximations

on the “complicated” random variable  $Y = h(X) = X^{\frac{3}{4}}$ .

$$E[Y] = \int_{\mathbb{R}} x^{\frac{3}{4}} \cdot f(x) dx = \int_0^1 3x^{\frac{11}{4}} dx = \frac{4}{5}$$

$$E[Y^2] = \int_{\mathbb{R}} x^{\frac{6}{4}} f(x) dx = \int_0^1 3x^{\frac{7}{2}} dx = \frac{2}{3}$$

$$Var(Y) = E[Y^2] - (E[Y])^2 = 0.0267$$

Before approximation, note that,

$$\mu = E[X] = \int_{\mathbb{R}} x f(x) dx = \int_0^1 3x^3 dx = \frac{3}{4}$$

$$E[X^2] = \int_{\mathbb{R}} x^2 f(x) dx = \int_0^1 3x^4 dx = \frac{3}{5}$$

$$\sigma^2 = Var[X] = E[X^2] - (E[X])^2 = 0.0375$$

$$h(\mu) = \mu^{\frac{3}{4}} = \left(\frac{3}{4}\right)^{\frac{3}{4}} = 0.8059$$

$$h'(\mu) = \frac{3}{4} \cdot \mu^{-\frac{1}{4}} = 0.8059$$

$$h''(\mu) = -\left(\frac{3}{16}\right) \cdot \mu^{-\frac{5}{4}} = -0.2686$$

$$E[Y] = h(\mu) + \frac{h''(\mu)\sigma^2}{2} = 0.8009$$

$$Var(Y) = [h'(\mu)]^2 \cdot \sigma^2 = 0.0243$$

## Moment Generating Functions

**Definition:** The **moment generating function** (mgf) of RV of X is

$$M_X(t) = E[e^{tX}]$$

**Remark:**  $M_X(t)$  is a function of t and not of X.

**Example:**  $X \sim Bern(p)$  so that  $X = 1$  with probability  $p$  and 0 with probability  $q$ .

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= \sum_x e^{tx} f(x) \\ &= e^{t \cdot 1} \cdot p + e^{t \cdot 0} q \\ M_X(t) &= p \cdot e^t + q \end{aligned}$$

**Example:** If  $X \sim \text{Exp}(\lambda)$ , then,

$$\begin{aligned}
 M_x(t) &= E[e^{tX}] \\
 &= \int_{\mathbb{R}} e^{tX} f(x) dx \quad (\text{LOTUS}) \\
 &= \int_0^{\infty} e^{tX} \lambda e^{-\lambda x} dx \\
 &= \lambda \int_0^{\infty} e^{(t-\lambda)x} dx \\
 &= \frac{\lambda}{\lambda - t} \quad \lambda > t
 \end{aligned}$$

**Big theorem:** Under certain conditions (e.g.  $M_X(t)$ ) must exist for all  $t \in (-\epsilon, \epsilon)$  for some  $\epsilon > 0$ , we have

$$E[X^k] = \frac{d^k}{dt^k} M_X(t) \Big|_{t=0}, \quad k = 1, 2, \dots$$

Thus, we can generate the moments of  $X$  from mgf. Sometimes it's easier to get moments this way directly.

**Proof:**

$$\begin{aligned}
 M_X(t) &= E[e^{tX}] \\
 &= E\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right] \quad (\text{McLaurin Series}) \\
 &= \sum_{k=0}^{\infty} E\left[\frac{(tX)^k}{k!}\right] \\
 &= 1 + tE(X) + \frac{t^2 E[X^2]}{2} + \dots
 \end{aligned}$$

This implies,

$$\frac{d}{dt} M_X(t) = E[X] + tE[X^2] + \dots$$

and so,

$$\frac{d}{dt} M_X(t) \Big|_{t=0} = E[X]$$

**Example:**  $X \sim \text{Bern}(p)$ . Then  $M_x(t) = pe^t + q$  and

$$\begin{aligned}
 E[X] &= \frac{d}{dt} M_x(t) \Big|_{t=0} \\
 &= \frac{d}{dt} (pe^t + q) \Big|_{t=0} \\
 &= pe^t \Big|_{t=0} \\
 &= p
 \end{aligned}$$

In fact, it's easy to see that  $E[X^k] = \frac{d^k}{dt^k} M_x(t) \Big|_{t=0} = p$  for all  $k$ .

**Example:**  $X \sim \text{Exp}(\lambda)$ . Then  $M_x(t) = \frac{\lambda}{\lambda-t}$  for  $\lambda > t$ . So,

$$E[X] = \frac{d}{dt} M_x(t) \Big|_{t=0} = \frac{\lambda}{(\lambda-t)^2} \Big|_{t=0} = \frac{1}{\lambda}$$

$$E[X^2] = \frac{d^2}{dt^2} M_x(t) \Big|_{t=0} = \frac{2\lambda}{(\lambda-t)^3} \Big|_{t=0} = \frac{2}{\lambda^2}$$

$$\text{Var}[X] = E[X^2] - (E[X])^2$$

**Theorem (mgf of linear function of X):** Suppose X has mgf  $M_x(t)$  and let  $Y = aX + b$ .

$$M_Y(t) = e^{tb} M_X(at)$$

**Proof:**

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t(ax+b)}] \\ &= e^{tb} E[e^{t(ax)}] \\ &= e^{tb} M_X(at) \end{aligned}$$

**Example:** Let  $X \sim \text{Exp}(\lambda)$  and  $Y = 3X + 2$ . Then,

$$\begin{aligned} M_Y(t) &= e^{2t} M_X(3t) \\ &= e^{2t} \frac{\lambda}{\lambda - 3t} \quad \text{if } \lambda > 3t \end{aligned}$$

**Theorem (identifying distribution):** In this class, each distribution has a unique mgf.

**Proof:** Not here!

**Example:** Suppose that Y has mgf,

$$\begin{aligned} M_Y(t) &= e^{2t} M_X(3t) \\ &= e^{2t} \frac{\lambda}{\lambda - 3t} \quad \text{if } \lambda > 3t \end{aligned}$$

Then by previous example and uniqueness of Mgf's, it must be the case that  $Y \sim 3X + 2$ , where  $X \sim \text{Exp}(\lambda)$ .

### 3.1.5 Some Probability Inequalities

**Theorem:** Markov's Inequality

If X is non-negative random variable and  $c > 0$  then  $P(X \geq c) \leq E[X]/C$ .



**Proof:** Because  $X$  is non-negative, we have,

$$\begin{aligned}
 E[X] &= \int_{\mathbb{R}} x f(x) dx \\
 &= \int_0^{\infty} x f(x) dx \\
 &\geq \int_c^{\infty} x f(x) dx \\
 &\geq c \int_c^{\infty} f(x) dx \\
 &= c \cdot P(X \geq c)
 \end{aligned}$$

**Theorem: Chebychev's Inequality**

Suppose  $E[X] = \mu$  and  $Var[X] = \sigma^2$ . Then, for any  $c > 0$ ,  $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

**Proof:** By Markov with  $|X - \mu|^2$  in place of  $X$  and  $c^2$  in place of  $c$ , we have,

$$\begin{aligned}
 P(|X - \mu| \geq c) &= P((X - \mu)^2 \geq c^2) \\
 &\leq \frac{E[(X - \mu)^2]}{c^2} \\
 &\quad \{ |X - \mu| \geq c \text{ if and only if } (X - \mu)^2 \geq c^2 \} \\
 &= \frac{\sigma^2}{c^2}
 \end{aligned}$$

**Remark:** Can also write  $P(|X - \mu| < c) \geq 1 - \frac{\sigma^2}{c^2}$ . If  $c = k \cdot \sigma$  then  $P(|X - \mu| \geq k \cdot \sigma) \leq \frac{1}{k^2}$ .

It means if you move further out in terms of standard deviation, the smaller the probability will be. Chebychev gives a bound on the probability that  $X$  deviates from the mean by more than a constant in terms of constant and the variance. You can always use Chebychev but it's crude.

**Example:**  $X \sim Unif(0, 1)$   $f(x) = 1$  for  $0 < x < 1$ .

Recall that  $E[X] = \frac{1}{2}$ ,  $Var(X) = \frac{1}{12}$ .

Chebychev implies that

$$P(|X - \frac{1}{2}| \geq c) \leq \frac{1}{12c^2}$$

In particular, for  $c = \frac{1}{3}$ ,

$$P(|X - \frac{1}{2}| \geq \frac{1}{3}) \leq \frac{3}{4} \text{ (upper bound)}$$

Let's compare the upper bound to exact answer,

$$\begin{aligned}
P(|x - \frac{1}{2}| \geq \frac{1}{3}) &= 1 - P(|x - \frac{1}{2}| < \frac{1}{3}) \\
&= 1 - P(-\frac{1}{3} < x - \frac{1}{2} < \frac{1}{3}) \\
&= 1 - P(\frac{1}{6} < x < \frac{5}{6}) \\
&= 1 - \int_{1/6}^{5/6} f(x) dx \\
&= 1 - \frac{2}{3} \\
&= \frac{1}{3}
\end{aligned}$$

So, Chebychev bound of  $\frac{3}{4}$  is pretty high in comparison.

**Theorem (Chernoff's inequality):** For any  $c$ ,

$$P(X \geq C) \leq e^{-ct} M_X(t)$$

**Proof:** By Markov with  $e^{tx}$  in place of  $X$  and  $e^{tc}$  in place of  $c$ , we have,

$$\begin{aligned}
P(X \geq C) &= P(e^{tx} \geq e^{tc}) \\
&= e^{-tc} E[e^{tX}] \quad \text{from Markov's inequality, } P(Y \geq a) \leq \frac{E[Y]}{a} \\
&= e^{-ct} M_X(t)
\end{aligned}$$

**Example:** Suppose  $X$  has the standard normal distribution with pdf  $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$  for  $x \in \mathbb{R}$ . It is easy to show that mgf of standard normal is

$$\begin{aligned}
M_x(t) &= E[e^{tX}] \\
&= \int_{\mathbb{R}} e^{tx} \phi(x) dx \\
&= e^{\frac{t^2}{2}}
\end{aligned}$$

Then using Chernoff with  $t = c$  immediately yields the tail probability.

$$\begin{aligned}
P(X \geq C) &\leq e^{-C^2} M_X(c) \\
&= e^{-\frac{c^2}{2}}
\end{aligned}$$

### 3.1.6 Functions of Random Variable

**Problem:** You have RV  $X$  and you know its pmf/pdf  $f(x)$ .

Define  $Y = h(x)$  (Some function of  $X$ ). Find  $g(y)$  the pmf/pdf of  $X$ .

**Remark:** Recall that LOTUS gave us results for  $E[h(x)]$ . But this is much more general than LOTUS, because we are going to get entire distribution of  $h(X)$ .

**Discrete Case:**  $X$  discrete implies  $Y$  discrete.

$$\begin{aligned}
 g(y) &= P(Y = y) \\
 &= P(h(X) = y) \\
 &= P(x|h(x) = y) \quad (\text{Probability of } x\text{'s such that } h(x) = y) \\
 &= \sum_{x|h(x)=y} f(x)
 \end{aligned}$$

**Example:**  $X$  is the number of H's in 2 coin tosses. We want the pmf of  $Y = h(x) = x^3 - x$ .

{TT, TH, HT, HH}

	$x$	0	1	2
$f(x) = P(X = x)$	$f(x) = P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$
	$y = x^3 - x$	0	0	6

3 values of  $x$  map into 2 values of  $y$  i.e. 0 & 6.

$$\begin{aligned}
 g(0) &= P(Y = 0) = P(X = 0 \text{ or } 1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \\
 g(6) &= P(Y = 6) = P(X = 2) = \frac{1}{4}
 \end{aligned}$$

$$g(y) = \begin{cases} \frac{3}{4} & \text{if } y = 0 \\ \frac{1}{4} & \text{if } y = 6 \end{cases}$$

**Example:**  $X$  is discrete with

$$f(x) = \begin{cases} \frac{1}{8} & \text{if } x = -1 \\ \frac{3}{8} & \text{if } x = 0 \\ \frac{1}{3} & \text{if } x = 1 \\ \frac{1}{6} & \text{if } x = 2 \end{cases}$$

Let  $Y = X^2$  (so  $Y$  can equal to 0, 1 or 4).

$$g(y) = \begin{cases} P(Y = 0) = f(0) = \frac{3}{8} \\ P(Y = 1) = f(1) + f(-1) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4} \\ P(Y = 4) = f(2) = \frac{1}{6} \end{cases}$$

**Continuous Case:**  $X$  is continuous implies  $Y$  can be continuous/discrete.

Example:  $Y = X^2$  (clearly continuous)

Example:

$$Y = \begin{cases} 0 & \text{if } X < 0 \\ 1 & \text{if } X \geq 0 \end{cases} \quad \text{is not continuous}$$

Method: Compute  $G(y)$ , the cdf of  $Y$ ,

$$\begin{aligned} G(y) &= P(Y \leq y) = P(h(x) \leq y) \\ &= \int_{\{x|h(x) \leq y\}} f(x) dx \end{aligned}$$

If  $G(y)$  is continuous, construct the pdf  $g(y)$  by differentiating.

**Example:**  $f(x) = |x|$ ,  $-1 \leq x \leq 1$ . Find the pdf of RV  $Y = h(X) = x^2$ .

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(X^2 \leq y) = \begin{cases} 0 & \text{if } y \leq 0 \\ 1 & \text{if } y \geq 1 \\ (*) & \text{if } 0 < y < 1 \end{cases} \quad x^2 \text{ must be between 0 and 1} \end{aligned}$$

where,

$$\begin{aligned} * &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} |x| dx \\ &= y \end{aligned}$$

Thus,

$$G(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ 1 & \text{if } y \geq 1 \\ y & \text{if } 0 < y < 1 \end{cases}$$

This implies,

$$g(y) = G'(y) = \begin{cases} 0 & \text{if } y < 0 \text{ and } y \geq 1 \\ 1 & \text{if } 0 < y < 1 \end{cases}$$

This means  $Y$  has the  $Unif(0, 1)$  distribution.

### Explanation

Since,  $Y = X^2$  so,  $0 \leq Y \leq 1$ . The cdf is determined as

\* If  $y \leq 0$  then  $P(X^2 \leq y) = 0$  because  $X^2$  is non-negative.

\* If  $y \geq 1$  then  $P(X^2 \leq y) = 1$  because  $X^2$  is at most 1.

\* If  $0 < y < 1$ , then  $P(-\sqrt{y} \leq X \leq \sqrt{y})$

**Example:** Suppose  $U \sim Unif(0, 1)$  Find the pdf of  $Y = -\ln(1 - U)$ .

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(-\ln(1 - U) \leq y) \\ &= P(1 - U \leq e^{-y}) \\ &= \int_0^{1-e^{-y}} f(u) du \\ &= 1 - e^{-y} \quad (\text{since } f(u) = 1) \\ g(y) &= G'(y) = e^{-y}, y > 0 \end{aligned}$$

This implies  $Y \sim Exp(\lambda = 1)$ .

$$G(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - e^{-y} & \text{if } y \geq 0 \end{cases}$$

For  $y < 0$ ,  $P(Y \leq y) = 0$  because  $Y = -\ln(1 - U)$  is non-negative.

### 3.1.7 Inverse Transform Theorem/Probability Integral Transform

Suppose  $X$  is a continuous random variable having cdf  $F(x)$ . Then the random variable  $F(X) \sim Unif(0, 1)$ .

*Proof:* Let  $Y = F(X)$ . The cdf of  $Y$  is

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(F(X) \leq y) \\ &= P(X \leq F^{-1}(y)) \quad \{\text{cdf is monotonically increasing}\} \\ &= F(F^{-1}(y)) \\ &= y \end{aligned}$$

**Monotonically increasing** means as input increases, output never decreases. Example:  $f(x) = x^2, x \geq 0$ .

**Remark:** This is a great theorem since it applies to all continuous RVs  $X$ .

**Corollary:**  $X = F^{-1}(U)$  so that you can plug  $Unif(0, 1)$  RV into the inverse cdf to generate a realization of RV having X's distribution.

**Method:** Set  $F(X) = U$  and solve  $X = F^{-1}(U)$  to generate X.

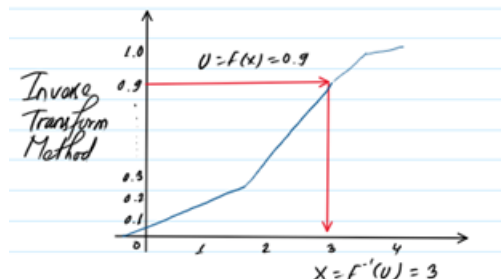


Figure 2: Inverse function

**Example:** Suppose  $X$  is  $Exp(\lambda)$  so that it has cdf  $F(x) = 1 - e^{-\lambda x}$ . Similar to previous example, set  $F(x) = 1 - e^{-\lambda x} = U$  and generate an  $Exp(\lambda)$  RV by solving for,

**Remark:** If you'd like to generate a nice, beautiful  $Exp(\lambda)$  pdf on a computer then

- Generate 10000  $Unif(0, 1)$ . Use rand function in excel or unifrnd in Matlab).
- Plug those 10000 into equation for X above and
- Plot the histogram of X.

### Another way to find pdf of a function of a continuous RV

Suppose that  $Y = h(x)$  is a monotonic function of a continuous RV X having pdf  $f(x)$  and cdf  $F(x)$ . Let's get the pdf  $g(y)$  of Y directly.

$$\begin{aligned}
 g(y) &= \frac{d}{dy} G(y) \\
 &= \frac{d}{dy} P(Y \leq y) \\
 &= \frac{d}{dy} P(X \leq h^{-1}(y)) \quad (h(x) \text{ is monotonic}) \\
 &= \frac{d}{dy} F(h^{-1}(y)) \\
 &= f(h^{-1}(y)) \left| \frac{dh^{-1}(y)}{dy} \right| \quad (\text{chain rule})
 \end{aligned}$$

**Example:** Suppose that  $f(x) = 3x^2$ ,  $0 < x < 1$ . Let  $Y = h(x) = x^{1/2}$  which is monotone

increasing.

$$\begin{aligned}
 g(y) &= & &= f(h^{-1}(y)) \left| \frac{dh^{-1}(y)}{dy} \right| \\
 &= f(y^2) \left| \frac{d(y^2)}{dy} \right| \\
 &= 3y^4 2y \\
 &= 6y^5, \quad 0 < y < 1
 \end{aligned}$$

Explanation

$$Y = X^{1/2}$$

$$X = Y^2$$

$h^{-1}(y) = X = Y^2$   $h^{-1}(y)$  is inverse function of  $h(x)$  expressed in terms of  $y$

$$f(h^{-1}(y)) = f(y^2) = 3y^4$$

**Theorem (why LOTUS works):** Let us assume  $h(\cdot)$  is monotonically increasing. Then

$$\begin{aligned}
 E[h(x)] &= E[Y] \\
 &= \int_R y f(h^{-1}(y)) \left| \frac{dh^{-1}(y)}{dy} \right| dy \\
 &= \int_R h(x) f(x) \left| \frac{dx}{dy} \right| dy \\
 &= \int_R h(x) f(x) dx
 \end{aligned}$$