# PROBABILITY THEORY

ESSENCE OF THE SUBJECT.

BY

SHAURYA PRATAP SINGH

*Department of Computer Science and Engineering*
*DIT University*
*Dehradun*

# Contents

# Chapter 1

# Introduction

## 1.1 Important Assumptions

- We will not be concerning ourselves with modes of inductive reasoning but with what is known as *statistical probability,* meaning our probabilities do not refer to judgments but to the possible outcomes of a conceptual experiment.
- We will use *idealized models* of our conceptual experiment. Example: When tossing a coin, we will take "landing heads" or "landing tails" as the only two possible outcomes, disregarding the possibility of the coin landing on its edge.
- At the outset, we must agree on all possible outcomes of the experiment (*sample space*) and the likelihood of landing an outcome (*probability*) associated with it.
- Our idealized model should run along the lines, *"Out of infinitely many worlds, one is selected at random."*
- Probabilities are numbers of the same nature as distances in geometry or masses in mechanics. The theory assumes they are given but need assume nothing about their actual numerical values or how they are measured in practice.

## 1.2 The Sample Space

- The *sample space* of an experiment is the set of all possible outcomes or results of that experiment. A sample space is usually depicted by *set notation* and the possible ordered outcomes are listed as *elements* in the set, known as the *sample points.*
- *Events* can be defined as *aggregations* of sample points or a *subset* of the sample space. We can further categorize events into:
  - *simple events / decomposible events*: Events containing a single sample point.
  - *complex events / indecomposible events*: Events containing more than one sample points.

  Example: Landing a "sum six" using two dice is a complex event that decomposes into six simple events: $(1,5),(2,4),(3,3),(4,2),(5,1)$.
- The events "sum six" and "two odd faces" have the sample point $(3,3)$ in common, making the two events **not** *mutually exclusive.*

## 1.3 Relations Among Events

**Note:** Capital letters denote an event *i.e.* set of sample points.

- Let us assume an arbitrary, but fixed sample space $\Omega$ with every sample point $x \in \Omega$ and any event $A \subset \Omega$.
- $A = B$ if and only if $A, B \in \Omega$ and $A \oplus B = \emptyset$
- $A = 0$ will be used to denote $A$ contains no sample points. This is impossible because whenever $A$ does not occur, another event occurs that contains all the sample points not contained in $A$.
- The event consisting of all the points not contained in $A$ is known as it's *complementary event* (or *negation*) of $A$, denoted by $A'$. Example: $\Omega' = 0$.

- To every collection $A, B, C, \ldots$ of events we define two new events as follows. The aggregate of sample points that belong to all the given sets will be denoted by $ABC \cdots$ called the *intersection* (or *simultaneous realization*) of $A, B, C \ldots$ The aggregate of sample points which belong to atleast one of $A, B, C, \ldots$ will be denoted by $A \cup B \cup C \cup \ldots$ and called the *union* (or *atleast one*) of the given events. The events $A, B, C \ldots$ are *mutually exclusive* if no two of them have a common sample point, that is, if $AB = 0$, $AC = 0$, $BC = 0, \ldots$

- The symbols $A \subset B$ and $B \supset A$ are equivalent and signify that every point of $A$ is contained in $B$, they are read respectively as "$A$ implies $B$" and "$B$ is implied by $A$". If this is the case, we shall write $B - A$ instead of $BA'$ to denote "the event $B$ occurs but not $A$."

- In the case of mutually exclusive events $AB = 0$ since ones occurrence means the others non-occurrence.

## 1.4 Discrete Sample Spaces

- A sample space is called *discrete* if it contains only finitely many points or infinitely many points which can be arranged into a simple sequence $E_1, E_2, \ldots$

- Probabilities in discrete spaces can be obtained by the means of simple addition, whereas integration is required in continuous sample spaces.

- All the sample points are *equally probable.* Given $n$ sample points, the probability of getting any point as an outcome must be $\frac{1}{n}$. Even if this is not directly involved in our experiment, it can be a **crude first approximation** as it can serve as a simple model for general orientation.

  **Note:** In the example of $r = n = 3$ balls and containers, with the balls being indistinguishable, we can argue that it might not be true since one could still physically differentiate the balls and thus assign different probabilities to the 10 new outcomes based on the previous 27 outcomes. (*Maxwell-Boltzman Statistics* vs *Bose-Einstein Statistics*) This example shows us the intricate interrelation between theory and experience. It teaches us not to rely on a priori arguments and be prepared to accept new and unforeseen schemes.

## 1.5 Definitions And Rules

- **Fundamental Convention:** Given a discrete sample space $\Omega$ with sample points $E_1, E_2, \ldots$, we shall assume that with each point $E_j$ there is associated a number, called the *probability* of $E_j$ and denoted by $\mathbf{P}\{E_j\}$. It is to be non-negative such that

$$\mathbf{P}\{E_1\} + \mathbf{P}\{E_2\} + \cdots = 1$$

- The probability $\mathbf{P}\{A\}$ of any event $A$ is the *sum* of probabilities of the sample points in it.

- By the virtue of this

$$\mathbf{P}\{\Omega\} = 1$$

- For any event $A$ we can say that

$$0 \leq \mathbf{P}\{A\} \leq 1$$

- We entertain the case of $\mathbf{P}\{E_j\} = 0$ which can be thought of as an *impossibility.*

- For any two arbitrary events $A_1$ and $A_2$, for either one of them or both to occur, we add their probabilities

$$\mathbf{P}\{A_1 \cup A_2\} \leq \mathbf{P}\{A_1\} + \mathbf{P}\{A_2\}$$

  For any point $E$ contained in both, it occurs only once in the $L.H.S$ and twice in the $R.H.S$, which gives rise to the inequality. This enables us to define a few theorems:

$$\mathbf{P}\{A_1 \cup A_2\} \leq \mathbf{P}\{A_1\} + \mathbf{P}\{A_2\} - \mathbf{P}\{A_1 A_2\}$$

  In case $A_1$ and $A_2$ are mutually exclusive, $A_1 A_2 = 0$, reducing the above equation to

$$\mathbf{P}\{A_1 \cup A_2\} = \mathbf{P}\{A_1\} + \mathbf{P}\{A_2\}$$

- For an arbitrary number of events, we have **Boole's Inequality:**

  $$\mathbf{P}\{A_1 \cup A_2 \cup \cdots\} \leq \mathbf{P}\{A_1\} + \mathbf{P}\{A_2\} + \cdots$$

  and in case of mutually exclusive events we have

  $$\mathbf{P}\{A_1 \cup A_2 \cup \cdots\} = \mathbf{P}\{A_1\} + \mathbf{P}\{A_2\} + \cdots$$

# Chapter 2

# Conditional Probability. Stochastic Independence

## 2.1 Conditional Probability

- Denoted by $\mathbf{P}\{A|H\}$, read as "probability of event $A$, assuming the event $H$." Let $A$ be any arbitrary event and $H$ be an event with positive probability, then

  $$\mathbf{P}\{A|H\} = \frac{\mathbf{P}\{AH\}}{\mathbf{P}\{H\}}$$

  The quantity so defined will be called the *conditional probability* of $A$ on the *hypothesis H*. When all sample points have equal probabilities, $\mathbf{P}\{A|H\}$ is equal to the ratio $\frac{N_{AH}}{N_H}$ of the sample points common to $A$ and $H$, to the number of points in $H$.

- Conditional probabilities remain undefined when hypothesis has 0 probability.

- Probabilities in the original sample space are sometimes called *absolute probabilities.*

- Taking conditional probability $w.r.t$ a particular hypothesis $H$, amounts to choosing $H$ as the new sample space with probabilities proportional to the original ones. The *probability factor* $\mathbf{P}\{H\}$ is necessary to reduce the total probability of the new sample space to unity.

- All general theorems on probabilities are valid also for conditional probabilities with respect to a particular hypothesis $H$.

  $$\mathbf{P}\{A \cup B|H\} = \mathbf{P}\{A|H\} + \mathbf{P}\{B|H\} - \mathbf{P}\{AB|H\}$$

  $$\mathbf{P}\{AH\} = \mathbf{P}\{A|H\} \cdot \mathbf{P}\{H\}$$

- The previous relation is called the *theorem on Compound Probabilities*. We can generalize it to three events by taking $H = BC$ as our hypothesis

  $$\mathbf{P}\{ABC\} = \mathbf{P}\{A|BC\} \cdot \mathbf{P}\{B|C\} \cdot \mathbf{P}\{C\}$$

- Let $H_1, \ldots, H_n$ be a set of mutually exclusive events out of which one necessary occurs

  $$H_1 \cup H_2 \cup \cdots H_n = \Omega$$

  Then we can say any event $A$ can occur in conjunction with some event $H_j$

  $$A = AH_1 \cup AH_2 \cup \cdots AH_n$$

  Since $AH_j$ are mutually exclusive, their probabilities can be added. Using the theorem on compound probabilities we get

  $$\mathbf{P}\{A\} = \sum \mathbf{P}\{A|H_j\} \cdot \mathbf{P}\{H_j\}$$

  More often the calculation of conditional probability is much easier than the calculation of $\mathbf{P}\{A\}$ directly.

- *Random Sampling (Without Replacement):* Whatever the first $r$ choices, at the $(r+1)st$ step, each of the remaining $n-r$ elements has a probability $1/(n-r)$ to be chosen. This definition arises due to conditional probability. If we pick a sample, we are left with $n-1$ samples, which leads to $\mathbf{P}\{A|H\} = 1/(n-1)$.

- *Stratified Sampling:* Suppose the human population consists of sub-populations or *strata*

$H_1, H_2, \ldots$ (mutually exclusive). These may be races, age groups, etc. Let $p_j$ be the probability that an individual chosen at random belongs to $H_j$ and $q_j$ be the conditional probability of the event $L$ (left-handedness) on the hypothesis that an individual belongs to $H_j$. Thus the probability that an individual chosen at random is left handed is

$$\mathbf{P}\{L\} = \sum_{i=1}^{n} \mathbf{P}\{L|H_i\} \cdot \mathbf{P}\{H_i\} = \sum_{i=1}^{n} p_i q_i$$

Given that an individual is left handed, the conditional probability of his belonging to stratum $H_j$ is

**Bayes's Rule:**

$$\mathbf{P}\{H_j|L\} = \frac{\mathbf{P}\{L|H_j\} \cdot \mathbf{P}\{H_j\}}{\sum_{i=1}^{n} \mathbf{P}\{L|H_i\} \cdot \mathbf{P}\{H_i\}} = \frac{p_j q_j}{\sum_{i=1}^{n} p_i q_i}$$

## 2.2 Probabilities Defined By Conditional Probabilities. Urn Models

- Probabilities of the sample space are usually not known in real world problems. But they can be theoretically derived from the given conditional probabilities.

- *Families:* Let $p_k$ be the probability of a family having $k$ children. We can thus have strata for number of kids in a family $H_1, H_2, \ldots$ with probabilities $p_1, p_2, \ldots$. Given that, "For any family size, sex distributions have equal probabilities." We can say for a family with $n$ kids, each of the $2^n$ distributions has the probability $2^{-n}$. Thus the absolute probability of any order of girls and boys in a family is given by $p_n.2^{-n}$. Let $B$ depict the event, "Family has boys but no girls."

$$\mathbf{P}\{B\} = \sum_{j=1}^{n} \mathbf{P}\{B|H_j\} \cdot \mathbf{P}\{H_j\} = \sum_{j=1}^{n} p_j.2^{-j}$$

Now if we as the question, "When it is known that a family has no boys (A becomes hypothesis), what is the (conditional) probability that it has only one child?"

$$\mathbf{P}\{H_1|B\} = \frac{\mathbf{P}\{B|H_1\} \cdot \mathbf{P}\{H_1\}}{\mathbf{P}\{B\}} = \frac{p_1 2^{-1}}{\sum_{j=1}^{n} p_j.2^{-j}}$$

- *Urn Models:* An urn consists of $b$ black and $r$ red balls. A ball is drawn at random and replaced. Moreover, $c$ balls of the color drawn and $d$ balls of the opposite colour are placed in the urn. A new random drawing is made from the urn now containing $r+b+c+d$ balls, and this procedure is repeated. Here, $c, d$ can be any arbitrary integers. Example: If $c = -1$ and $d = 0$, we have a model for *random drawings without replacement* which terminates after $r+b$ steps.

probability to draw a black ball first is given by

$$\frac{b}{b+r}$$

We also know that the conditional probability for drawing a second black ball is

$$\frac{b+c}{r+b+c+d}$$

This gives us the absolute probability for drawing a two black balls is succession as

$$\frac{b}{b+r} \cdot \frac{b+c}{r+b+c+d}$$

These probabilities can be easily calculated and add up to one, which is verifiable through induction.

- *Polya's Urn Scheme:* It is characterized by $d = 0$ and $c > 0$. Helps model phenomena where each occurrence increases the probability of further occurrences (*contagious diseases*). The probability that of $n = n_1 + n_2$ drawings, first $n_1$ result in black balls and remaining $n_2$ result in red balls is given by

$$\frac{b(b+c)\cdots(b+c(n_1-1)) \cdot r(r+c)\cdots(r+c(n_2-1))}{(b+r)(b+r+c)\cdots(b+r+c(n-1))}$$

For an arbitrary ordering where $n$ drawings result in $n_1$ black and $n_2$ red

$$p_{n_1,n} = \frac{\binom{n_1-1+b/c}{n_1}\binom{n_2-1+r/c}{n_2}}{\binom{n-1+(b+r)/c}{n}} = \frac{\binom{-b/c}{n_1}\binom{-r/c}{n_2}}{\binom{-(b+r)/c}{n}}$$

- *Urn models for stratification. Spurious contagion:* We are given two groups in population with the ratio 1 : 5 and with each of these groups is associated an *urn*. *urn*$_1$ contains $r_1$ red balls and $b_1$ black balls, whereas *urn*$_2$ contains $r_2$ red balls and $b_2$ black balls. We randomly draw a ball from either of the urns and then put it back in the same urn. If it is a red ball, a random person from their respective population group faces an accident. "What is the (conditional) probability of a person facing an accident provided he has faced one already?" It is different from the (absolute) probability of a person facing two accidents.

  The Probability of drawing a red ball for the first time is given by

  $$\mathbf{P}\{R\} = \frac{1}{6} \cdot \frac{r_1}{r_1 + b_1} + \frac{5}{6} \cdot \frac{r_2}{r_2 + b_2}$$

  The Probability of drawing two red balls in a row is given by

  $$\mathbf{P}\{RR\} = \frac{1}{6} \cdot \left(\frac{r_1}{r_1 + b_1}\right)^2 + \frac{5}{6} \cdot \left(\frac{r_2}{r_2 + b_2}\right)^2$$

  since in our case $c = d = 0$. The Probability that a person who has suffered an accident (Hypothesis), suffers it again

  $$\mathbf{P}\{R|R\} = \frac{\mathbf{P}\{RR\}}{\mathbf{P}\{R\}}$$

  If we assume $r_1/(r_1 + b_1) = 0.6$ and $r_2/(r_2 + b_2) = 0.6$, we see that

  $$\mathbf{P}\{R\} = 0.15 \quad \mathbf{P}\{RR\} = 0.063 \quad \mathbf{P}\{R|R\} = 0.42$$

  The important point to note is that there is no *after effect* or the so called *contagion* in our model. The difference in probability is because there is more information available about the person having an accident the second time around.

- *Example 2:* We have a collection of $N + 1$ urns, each containing a total of $N$ red and white balls. The urn number $k$ contains $k$ red and $N - k$ white balls with $k = 0, 1, 2, \ldots, N$. An urn is chosen at random and $n$ drawings are made from it, the ball being replaced each time. We describe event $A$ as all balls turn out to be red. We seek the (conditional) probability that the next drawing (event $B$) will also be red.

  The probability to select a box is given by $\frac{1}{N+1}$ and for any box $k$, the probability to pick a red ball is given by $\frac{k}{N}$. Since we are not altering the number of red and white balls, for drawing $n$ red balls, the probability becomes $\left(\frac{k}{N}\right)^n$. We can thus define

  $$\mathbf{P}\{A\} = \sum_{j=1}^{N} \mathbf{P}\{A|H_j\} \cdot \mathbf{P}\{H\} = \frac{1^n + 2^n + \cdots + N^n}{N^n(N + 1)}$$

  $\mathbf{P}\{AB\}$ means $n + 1$ draws, so we can say

  $$\mathbf{P}\{B\} = \mathbf{P}\{AB\} = \frac{1^{n+1} + 2^{n+1} + \cdots + N^{n+1}}{N^{n+1}(N + 1)}$$

  The required probability is $\mathbf{P}\{B|A\} = \mathbf{P}\{B\}/\mathbf{P}\{A\}$. When $N$ is very large, the numerator can be approximated as the area between $x-$axis and the curve $xn$, changing our equation to

  $$\mathbf{P}\{A\} \approx \frac{1}{N^n(N + 1)} \int_0^N x^n dx = \frac{N}{N + 1} \cdot \frac{1}{n + 1} \approx \frac{1}{n + 1}$$

  Similarly,

  $$\mathbf{P}\{B|A\} \approx \frac{n + 1}{n + 2}$$

  This is called *the law of succession of Laplace.*

## 2.3 Stochastic Independence

- When $\mathbf{P}\{A|H\} = \mathbf{P}\{A\}$, we can say that $A$ is *stochastically independent* of $H$. Since

  $$\mathbf{P}\{AH\} = \mathbf{P}\{A|H\}.\mathbf{P}\{H\}$$

  We can now say that

  $$\mathbf{P}\{AH\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{H\}$$

- The equation is symmetric in $A$ and $H$ and shows that both are independent of one another. The equation also holds when $\mathbf{P}\{H\} = 0$ since $\mathbf{P}\{A|H\}$ is not defined in that case. The term *statistically independent* is also used in the same context.

- *Example 1:* From a deck of cards, a random drawing is made. The probabilities of drawing an "ace" and that of drawing a "spade" are $\frac{1}{4}$ and $\frac{1}{13}$ respectively. These two events are described as stochastically independent and the probability of their simultaneous realization is $\frac{1}{52}$.

- *Example 2:* Two true dice are thrown. The events "ace with first" and "even with second" are independent since the probability of their simultaneous realization, $\frac{3}{36} = \frac{1}{12}$, is the product of their probabilities, $\frac{1}{6}$ and $\frac{1}{2}$.

- Stochastic independence implies that no inference can be drawn from the occurrence of $H$ to that of $A$; therefore stochastic independence of $A$ and $H$ should mean the same as the independence of ($A$ and $H'$), ($A'$ and $H$) and ($A'$ and $H'$).

- Given

  $$\mathbf{P}\{AB\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{B\}$$

  $$\mathbf{P}\{AC\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{C\}$$

  $$\mathbf{P}\{BC\} = \mathbf{P}\{B\} \cdot \mathbf{P}\{C\}$$

  We can say that

  $$\mathbf{P}\{ABC\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{B\} \cdot \mathbf{P}\{C\}$$

  also holds in the case of stochastic independence. Some other cases can allow for $\mathbf{P}\{ABC\} = 0$.

- The events $A_1, A_2, \ldots, A_n$ are called independent if for all combinations $1 \leq i < j < k < \cdots \leq n$ the multiplication rules

  $$\mathbf{P}\{A_i A_j\} = \mathbf{P}\{A_i\}\mathbf{P}\{A_j\}$$

  $$\mathbf{P}\{A_i A_j A_k\} = \mathbf{P}\{A_i\}\mathbf{P}\{A_j\}\mathbf{P}\{A_k\}$$

  $$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

  $$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

  $$\mathbf{P}\{A_1 A_2 \cdots A_n\} = \mathbf{P}\{A_1\}\mathbf{P}\{A_2\} \cdots \mathbf{P}\{A_n\}$$

  hold true, *i.e.* $2^n - n - 1$ conditions.

## 2.4 Product Spaces and Independent Trials

# Chapter 3

# Probability: Introductory

## 3.1 Lecture 1: Basics

**Probabilistic Model:** It is a quantitative description of a situation, a phenomenon or an experiment whose outcome is uncertain. Putting together such a model involves two keys steps.

- Specifying a sample space.
- Assign probability laws: Axioms and properties derived from them.
- Identify an event of interest. (Solving)
- Calculate. (Solving)

**Specifying a sample space:** Sample space is given as a set of all possible outcomes, denoted by $\Omega$. All the sample points within the set must be:

- Mutually exclusive.
- Collectively exhaustive.
- At the "right" granularity or the level of detail.

**Discrete Sample Spaces:**

- It can be finite or countable infinite.
- It can be described sequentially in terms of a tree.
- We assign each point its own probability.

**Continuous Sample Spaces:**

- Ex: Coordinates $(x, y)$ inside a square such that $0 \leq x, y \leq 1$ and $x, y \in \mathbb{R}$.
- When assigning probabilities to continuous spaces, we take *subsets of sample points.*

**Event** A subset of the sample space.

**Axioms:**

- *Non negativity:* $\mathbf{P}\{A\} \geq 0$
- *Normalization:* $\mathbf{P}\{\Omega\} = 1$
- *(Finite) Additivity:* If $A \cap B = \emptyset$, then $\mathbf{P}\{A \cup B\} = \mathbf{P}\{A\} + \mathbf{P}\{B\}$

**Consequences I:**

- $\mathbf{P}\{A\} \leq 1$
- $\mathbf{P}\{\emptyset\} = 0$
- $\mathbf{P}\{A\} + \mathbf{P}\{A^c\} = 1$
- If $A, B, \ldots$ are disjoint then $\mathbf{P}\{A_1 \cup A_2 \cup \cdots\} = \sum \mathbf{P}\{A_i\}$ *(countable additivity axiom)* only applies in case of a sequence.

**Consequences II:**

- If $A \subset B$, then $\mathbf{P}\{A\} \leq \mathbf{P}\{B\}$
- $\mathbf{P}\{A \cup B\} = \mathbf{P}\{A\} + \mathbf{P}\{B\} - \mathbf{P}\{A \cap B\}$
- $\mathbf{P}\{A \cup B\} \leq \mathbf{P}\{A\} + \mathbf{P}\{B\}$ *(Boole's Inequality)*
- $\mathbf{P}\{A \cup B \cup C\} = \mathbf{P}\{A\} + \mathbf{P}\{A^c \cap B\} + \mathbf{P}\{A^c \cap B^c \cap C\}$

**Discrete Uniform Law:** If $n(\Omega) = N$, then the probability of each of them is $\frac{1}{N}$ and for an event $A$ consisting of $k$ sample points, the probability is

$$\mathbf{P}\{A\} = \frac{k}{N}$$

**Uniform probability Law (Continuous):** Probability = Area within the sample space.

**DeMorgan's Laws:**

- $(\cap S_n)^c = \cup S_n^c$
- $(\cup S_n)^c = \cap S_n^c$

**Sequences and their limits:** Given an element of a sequence $\{a_i\}$ where $i \in \mathbb{N}$ and $a_i \in \mathbb{R}^n$. We can define as sequence as $f : \mathbb{N} \to \mathbb{R}$ such that $f(i) = a_i$

- Convergence can be defined as

  $$\lim_{i \to \infty} a_i = a$$

  For any $\epsilon > 0$, there exists an $i_0$, such that if $i \geq i_0$, then $|a_i - a| < \epsilon$

- *Some properties:* If $\lim_{i\to\infty} a_i = a$ and $\lim_{i\to\infty} b_i = b$ then

  $$a_i + b_i \to a + b; \quad a_i b_i \to ab$$

  If a function $g$ is continuous, then

  $$g(a_i) \to g(a)$$

- *Conditions for convergence:*
  - If $a_i \leq a_{i+1}$ the sequence converges to either $\infty$ or some real number $a$.
  - If $|a_i - a| \leq b_i \quad \forall i$, and $b_i \to 0$ then $a_i \to a$.

**Infinite Series:** Provided the limit exists, we can define an infinite series as

$$\lim_{n\to\infty} \sum_{i=1}^{n} a_i$$

Here are some conditions regarding existence of limits.

- If $a_i \geq 0$, the limit exists.

- If $a_i$ all do not have same signs:
  - Limit may not exist.
  - Limit may exist but differ when the order is changed.
  - Limit exists and is independent of order of summation if

    $$\sum_{i=1}^{\infty} |a_i| < \infty$$

**Geometric Series:** Given by

$$\sum_{i=0}^{\infty} \alpha_i = 1 + \alpha + \alpha^2 + \cdots = \frac{1}{1-\alpha} \qquad |\alpha| < 1$$

**Series with multiple indices** The different terms of our series can be represented on a grid where we add them one after another. It can be done in any order as long as the series converges (sum is finite). More formally

$$\sum_{i\geq 1, j \geq 1} a_{ij} \quad ; \quad \boxed{\sum |a_{ij}| < \infty}$$

One way of adding is, by fixing the index $i$ and adding all the associated $j$ then doing so for another $i$ or vice versa

$$\sum_{i=1}^{\infty}\left(\sum_{j=1}^{\infty} a_{ij}\right) \quad = \quad \sum_{j=1}^{\infty}\left(\sum_{i=1}^{\infty} a_{ij}\right)$$

**Over a Limited Range:** This is done by a condition over $i, j$

$$\sum_{(i,j):j\leq i}^{\infty} a_{ij} = \sum_{i=1}^{\infty}\left(\sum_{j=1}^{i} a_{ij}\right) = \sum_{j=1}^{\infty}\left(\sum_{i=j}^{\infty} a_{ij}\right)$$

**Countable sets:** A set will be called countable if its elements can be put in a $1-1$ correspondence with positive integers. Examples: Set of all integers, all rational numbers such that $0 \leq q \leq 1$.

**Uncountable sets:** Can be finite but cannot be discretely sequenced. Examples: $[0,1]$, all real numbers.

**Bonferroni's Inequality:**

$$\mathbf{P}\{A_1 \cap \cdots \cap A_n\} \geq \mathbf{P}\{A_1\} + \cdots + \mathbf{P}\{A_n\} - (n-1)$$

## 3.2 Lecture 2: Fundamental Theorems

**The Idea of Conditioning:** Given that an event $A$ (a region on the sample space) occurs, what is the probability that another event $B$ (the portion of $B$ within the region of $A$ or $A \cap B$) will occur. Given by the formula

$$\mathbf{P}\{A|B\} = \frac{\mathbf{P}\{A \cap B\}}{\mathbf{P}\{B\}} \quad ; \quad \mathbf{P}\{B\} > 0$$

**Properties:**

- $\mathbf{P}\{\Omega|B\} = \mathbf{P}\{B|B\} = 1$
- If $A \cap C = \emptyset$ then $\mathbf{P}\{(A \cup C)|B\} = \mathbf{P}\{A|B\} + \mathbf{P}\{C|B\}$

**Multiplication Rule:**

$$\mathbf{P}\{A \cap B\} = \mathbf{P}\{B\} \cdot \mathbf{P}\{A|B\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{B|A\}$$

Generalizing for $n$ unions

$$\boxed{\mathbf{P}\{A_1 \cap \cdots \cap A_n\} = \mathbf{P}\{A_1\} \prod_{i=2}^{n} \mathbf{P}\{A_i | A_1 \cap \cdots \cap A_{i-1}\}}$$

**Total Probability Theorem:** Given a sample space $\Omega$, we make $n$ partitions $A_1, \ldots, A_n$, each having a probability of

$$\mathbf{P}\{A_i\} \quad \forall i \in \mathbb{N}$$

Given some event $B$ in the sample space, we can then say that

$$\mathbf{P}\{B\} = \mathbf{P}\{B \cap A_1\} + \cdots + \mathbf{P}\{B \cap A_n\}$$

Using the multiplication rule, we can change our equation to

$$\mathbf{P}\{B\} = \mathbf{P}\{A_1\}\mathbf{P}\{B|A_1\} + \cdots + \mathbf{P}\{A_2\}\mathbf{P}\{B|A_n\}$$

Finally giving us the formula

$$\boxed{\mathbf{P}\{B\} = \sum_{i=1}^{n} \mathbf{P}\{A_i\}\mathbf{P}\{B|A_i\}}$$

We notice that this is the *weighted average* of the $\mathbf{P}\{B|A_i\}$ with $\mathbf{P}\{A_i\}$ being the weights.

**Bayes' Theorem:** We partition our sample space into $A_1, \ldots, A_n$, each having a probability $\mathbf{P}\{A_i\}$ known as the *"initial beliefs"*. Capturing how likely we believe each scenario to be. Under these beliefs we also have the probability that an event $B$ of interest, given by $\mathbf{P}\{B|A_i\}$. We then carry out a probabilistic experiment and find out that the event $B$ did occur, we want to use it to revise our beliefs about the likelihood of different scenarios, given by

$$\boxed{\mathbf{P}\{A_i|B\} = \frac{\mathbf{P}\{A_i \cap B\}}{\mathbf{P}\{B\}} = \frac{\mathbf{P}\{B|A_i\} \cdot \mathbf{P}\{A_i\}}{\sum_{j=1}^{n} \mathbf{P}\{A_j\} \cdot \mathbf{P}\{B|A_j\}}}$$

This is a systematic way of *learning from experience*, and is the foundation for the branch of mathematics known as *Bayesian Inference*.

**Bayesian Inference:**

- Beliefs $\mathbf{P}\{A_i\}$ on possible causes of event $B$.
- Model of the world under each $A_i$: $\mathbf{P}\{A_i\}$

$$\textit{Model:} \quad A_i \xrightarrow{\mathbf{P}\{B|A_i\}} B$$

- Draw conclusions about causes

$$\textit{Inference:} \quad B \xrightarrow{\mathbf{P}\{A_i|B\}} A_i$$

## 3.3 Lecture 3: Independence

**Independence:** If two events occur and the knowledge of the second event does not affect our "beliefs" for the first since the second event fails to provide us with any additional information about the first event, or more formally the conditional probability remains the same as the absolute probability, then the two events are *stochastic-ally independent.* Given by

$$\mathbf{P}\{B|A\} = \mathbf{P}\{B\}$$

The probability that both $A$ and $B$ occur now becomes

$$\boxed{\mathbf{P}\{A \cap B\} = \mathbf{P}\{A\}.\mathbf{P}\{B\}}$$

which is a much cleaner way of defining the notion of *independence.* This relation is *symmetric* to both $A$ and $B$, meaning it will also apply if $\mathbf{P}\{A|B\} = \mathbf{P}\{A\}$. Lastly it also holds when $\mathbf{P}\{A\} = 0$ **Note:** Being independent does not mean being disjoint.

**Consequences:** $A$ and $B^c$ are also independent. Intuitive argument being if $A$ tells us nothing about $B$ it shouldn't tell us anything about it's complement either.

$$A = (A \cap B) \cup (A \cap B^c)$$

Since the two are disjoint, we can say through additivity axiom

$$\mathbf{P}\{A\} = \mathbf{P}\{A \cap B\} + \mathbf{P}\{A \cap B^c\}$$

Using independence and then switching sides

$$\mathbf{P}\{A\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{B\} + \mathbf{P}\{A \cap B^c\}$$

$$\mathbf{P}\{A \cap B^c\} = \mathbf{P}\{A\} - \mathbf{P}\{A\} \cdot \mathbf{P}\{B\}$$

Taking $\mathbf{P}\{A\}$ common, we get

$$\mathbf{P}\{A \cap B^c\} = \mathbf{P}\{A\}(1 - \mathbf{P}\{B\})$$

$$\boxed{\mathbf{P}\{A \cap B^c\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{B^c\}}$$

Similar arguments for ($A^c$ and $B$) and ($A^c$ and $B^c$) are also valid.

**Conditional Independence:** Given a condition $C$, *conditional independence* is defined as independence under the probability law $\mathbf{P}\{\cdot \mid C\}$

$$\boxed{\mathbf{P}\{(A \cup B)|C\} = \mathbf{P}\{A|C\} \cdot \mathbf{P}\{B|C\}}$$

- Independence does not imply conditional independence.
- Conditional independence may affect the independence.

**Independence of a collection of events:** Information on some events of the events does not change the probabilities related to the remaining events. Events $A_1, \ldots, A_n$ are called independent events if

$$\boxed{\mathbf{P}\{A_i \cap A_j \cap \cdots \cap A_m\} = \mathbf{P}\{A_i\} \cdot \mathbf{P}\{A_j\} \cdots \mathbf{P}\{A_m\}}$$

For all distinct indices $i, j, \ldots, m \leq n$

**Independence of a Collection vs Pairwise Independence:** We take two independent fair coins tosses and define the following events on them:

- $H_1$ : First coin toss is heads. $(HH, HT)$
- $H_2$ : Second coin toss is heads. $(HH, TH)$

Since the above two events are independent, we can say that:

$$\mathbf{P}\{H_1\} = \mathbf{P}\{H_2\} = \frac{1}{2}$$

Now we define a new event $C$ where both tosses have the same result $(HT, TT)$.

$$\mathbf{P}\{C\} = \frac{1}{2}$$

First we wish to find out whether all the three events are pairwise independent

$$\mathbf{P}\{H_1 \cap H_2\} = \mathbf{P}\{H_1\} \cdot \mathbf{P}\{H_2\} = \frac{1}{4} = \mathbf{P}\{HH\}$$

$$\mathbf{P}\{H_1 \cap C\} = \mathbf{P}\{H_1\} \cdot \mathbf{P}\{C\} = \frac{1}{4} = \mathbf{P}\{HH\}$$

$$\mathbf{P}\{H_2 \cap C\} = \mathbf{P}\{H_2\} \cdot \mathbf{P}\{C\} = \frac{1}{4} = \mathbf{P}\{HH\}$$

$$\mathbf{P}\{H_1 \cap H_2 \cap C\} = \frac{1}{4} = \mathbf{P}\{HH\}$$

But what is interesting is

$$\mathbf{P}\{H_1\} \cdot \mathbf{P}\{H_2\} \cdot \mathbf{P}\{C\} = \frac{1}{8}$$

Since the above two are not equal this makes the three events not independent as a collection, even though they are pairwise independent. The intuitive explanation lies in the fact that

$$\mathbf{P}\{C|H_1\} = \mathbf{P}\{C \cap H_1\} \cdot \mathbf{P}\{H_1\} = \frac{1}{2} = \mathbf{P}\{C\}$$

But we also know that

$$\mathbf{P}\{C|(H_1 \cap H_2)\} = \frac{\mathbf{P}\{C \cap H_1 \cap H_2\}}{\mathbf{P}\{H_1 \cap H_2\}} = \frac{1/4}{1/4} = 1$$

and since

$$\mathbf{P}\{C\} \neq \mathbf{P}\{C|(H_1 \cap H_2)\}$$

The three events are not collectively independent.

**Reliability:** Independence helps us break down complex situations into simpler models that we can work on separately, and later combine the results. This finds application in the analysis of the *reliability* of a system that consists of independent units.

Example: Suppose we have $n$ units working. The probability that a unit is up is $p_i$. Let the event that $i^{th}$ unit is u be given by $U_i$ and the event when it fails be given by $F_i$. Since we know that the failure or working of one unit does not affect the others, it is safe to intuitively say that the units are independent. Therefore when the units are in **series**

$$\mathbf{P}\{\text{System is up}\} = \mathbf{P}\{U_1 \cap U_2 \cap \cdots \cap U_n\}$$

$$\mathbf{P}\{\text{System is up}\} = \mathbf{P}\{U_1\} \cdot \mathbf{P}\{U_2\} \cdots \mathbf{P}\{U_n\}$$

$$\mathbf{P}\{\text{System is up}\} = p_1 p_2 \ldots p_n$$

And the probability of the system being up in **parallel** is

$$\mathbf{P}\{\text{System is up}\} = \mathbf{P}\{U_1 \cup U_2 \cup \cdots \cup U_n\}$$

$$\mathbf{P}\{\text{System is up}\} = 1 - \mathbf{P}\{F_1 \cap F_2 \cap \cdots \cap F_n\}$$

$$\mathbf{P}\{\text{System is up}\} = 1 - \mathbf{P}\{F_1\} \cdot \mathbf{P}\{F_2\} \cdots \mathbf{P}\{F_n\}$$

$$\mathbf{P}\{\text{System is up}\} = 1 - (1 - p_1)(1 - p_2)\ldots(1 - p_n)$$

## 3.4 Lecture 4: Counting

**Note:** Need a discrete and uniform sample space for counting to take place.

**Basic Counting Principle:** Construction of an object through $r$ different stages with choice at $i^{th}$ stage being $n_i$.

$$\boxed{n(\text{ch}) = \prod_{i=1}^{r} n_i}$$

**Subsets:** How many subsets are there? *i.e.* cardinality of the power set of a set $A$, given by

$$\boxed{n(\,P(A)\,) = 2^n}$$

**Combinations:** Number of ways to select $k$ elements from $n$ without ordering.

$$\boxed{\binom{n}{k} = {}^nC_k = \frac{n!}{k!(n-k)!}}$$

These are known as *binomial coefficients*. Counting Subsets can also be given by

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

**Permutations:** Number of ways of ordering the $k$ selected elements.

$$\boxed{{}^nP_k = \frac{n!}{(n-k)!} = k!\binom{n}{k}}$$

**Binomial Probabilities:** In this example, we are given $n \geq 1$ coin tosses where the probability of heads $\mathbf{P}\{H\} = p$. This implies that $\mathbf{P}\{T\} = 1 - p$. The assumption we will make in our probabilistic model is that coin tosses are independent events. Thus we can generalize

$$\mathbf{P}\{k - \text{head sequence}\} = p^k(1-p)^{n-k}$$

To find the probability of getting $k$ heads, we just need to multiply the probability of getting a $k-$ head sequence by the number of ways we can select $k$ heads in an $n$ toss sequence.

$$\boxed{\mathbf{P}\{k - \text{heads}\} = p^k(1-p)^{n-k} \cdot \binom{n}{k}}$$

**Partitions:** Our set consists of $n \geq 1$ distinct items and $r \geq 1$ persons. We want to give $n_i$ items to the person $i$. For this to be possible

$$\sum_{i=1}^{r} n_i = n$$

Now if each person was to order their $n_i$ slots in a list, they would have $n_i!$ ways to do it. Assuming we have $C$ partition choices, we get

$$Cn_1!n_2!\ldots n_r! = n!$$

In other words the *multinomial coefficient is*

$$\boxed{C = \frac{n!}{\prod_{i=1}^{r} n_i!}}$$

*Example:* Deal a deck of cards (fairly) to four players and find the probability that each player gets an ace. By "fairly" we mean to say that all partitions are equally likely.

The number of possible ways to do this is given by

$$C = \frac{52!}{13!13!13!13!}$$

We now construct the outcome where each person as an ace. This can be done by distributing the aces which can be done in 4! ways. Next we need to partition the remaining 48 cards into four subsets, which can be done in $\frac{48!}{12!12!12!12!}$ ways. Thus our desired probability becomes

$$\mathbf{P}\{\text{All have an ace}\} = \frac{24 \cdot \frac{48!}{(12!)^4}}{\frac{52!}{(13!)^4}} = 0.1055$$

**Multinomial Probabilities:**

- We have balls of different colors $i = 1, \ldots, r$.
- Probability of picking a ball of color $i$ is $p_i$.
- Draw $n$ balls independently.
- We fix numbers $n_1 + \ldots + n_r = n$
- Find the probability that for every color $i$, we have $n_i$ balls.

- Taking the case $r = 3$ and $n = 7$. We can have a possible outcome as follows

  $$(1\ 1\ 3\ 1\ 2\ 2\ 1)$$

  $$1's: \quad 4; \qquad 2's: \quad 2; \qquad 3's: \quad 1$$

  The probability of such an outcome can be given by

  $$\mathbf{P}\{\text{sequence}(n_1, n_2, \dots, n_r)\} = p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$$

  What we are interested in is obtaining the total probability of some sequence of the above type, which just involves multiplying by the number of sequences of such type.

  If we look at the above case carefully, it just involves partitioning the sequence into subsets of cardinalities as per $n_i$. We can thus generally state that the count of a sequence of type $(n_1, n_2, \dots, n_r)$ is equivalent to the number of partitions of $\{1, \dots, n\}$ into subsets of cardinalities $n_1, n_2, \dots, n_r$. Which finally gives us

  $$\boxed{\mathbf{P}\{\text{seq}(n_1, n_2, \dots, n_r)\} = \frac{n!}{\prod_{i=0}^{r} n_i!} \cdot p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}}$$

## 3.5 Lecture 5: Random Variables - I

Loosely speaking, a random variable is a numeric quantity that takes random values. It is a function that maps outcomes to a real value. Since it is a function, given two random variables, one can create a new random variable from the previous two. Determine the outcome of an experiment based on some *probability function*, plug it into the random variable to get a weighted real value. **Example:** Body Mass Index from random sampling.

**Formalism:** A random variable can be defined as a function $X : \Omega \to \mathbb{R}$.

**Probability Mass Function (P.M.F):** It is also called the *probability law* or the *probability distribution* of $X$. **Example:** Sum of rolls on two tetrahedral dice.

$$\boxed{p_X(x) \ = \ P(X = x) \ = \ P(\{\omega \in \Omega \mid X(\omega) = x\})}$$

**Properties:** $p_X(x) \geq 0$ and $\sum_x p_X(x) = 1$

**Bernoulli Random Variable:** It can be defined with parameter $p \in [0, 1]$

$$\boxed{X = \begin{cases} 1, & \text{w.p. } \ p \\ 0, & \text{w.p. } \ 1 - p \end{cases}}$$

**Uses of Bernoulli random variables**

- Bernoulli random variables are specially useful in situations where a trial results in outcomes of twos, such as, success / failure, heads / tails, etc.
- Connection between events and random variables made through *indicators.* Meaning, $I_A = 1$ if and only if event $A$ occurs and $I_A = 0$ otherwise. We can also say that $P_{I_A}(1) = P(I_A = 1) = P(A)$.

**Discrete Uniform Random Variable:** In this kind of distribution, all values are equally likely. It takes parameters of form $[a, b] \in \mathbb{Z}^+$, in other words, $\Omega = \{a, a+1, a+2, \dots, b\}$. Then we can say that

$$\boxed{p_X(x) = \frac{1}{b - a + 1}, \quad \forall x \in [a, b]}$$

- This probability distribution is loosely a model of *complete ignorance*.
- In case $a = b$ we get a single outcome which makes our experiment *deterministic.*

**Binomial Random Variable:** It gives us the probability of obtaining $k$ successful outcomes from $n$ independent trials when the number of total possible outcomes is two. It takes parameters of the form $p \in [0, 1]$ and $n \in \mathbb{Z}^+$.

$$\boxed{p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \forall k = 0, 1, \dots, n}$$

**Independence (infinitely many events):** Any finite subset of elements is independent of others.

**Geometric Random Variable: (with example)** It takes parameter of the form $p \in (0,1]$

$$\boxed{p_X(x) = P(T\dots TH) = (1-p)^{k-1}p, \quad k = 0,1,2m\dots}$$

- **Experiment:** Infinitely many independent tosses of a coin where $P(Heads) = p$.
- **Sample Space:** Set of infinite sequences of H and T.
- **Random Variable:** Number of tosses until the first heads.
- **Models:** Waiting times; number of trials until success.
- **Note:** We know that $P(TTT\dots) \Rightarrow P(T\dots T)$, thus we can say $P(TTT\dots) \leq P(T\dots T) = (1-p)^k$
  For $k \to \infty, P(TT\dots T_k) = 0$

**Expectation / mean of a random variable:** One can think about expectation as the *average* in a large number of independent repetitions of the experiment, when probabilities are considered as frequencies.

$$\boxed{\mathbf{E}[X] = \sum_x x p_X(x)}$$

**Caution:** If we have an infinite sum, it needs to be well defined (needs to be *absolute convergent*).

$$\boxed{\sum_x |x| p_X(x) < \infty}$$

**Expectation of a Bernoulli random variable:**

$$X = \begin{cases} 1, & \text{w.p.} \quad p \\ 0, & \text{w.p.} \quad 1-p \end{cases}$$

$\mathbf{E}[X] = 1.p + 0.(1-p) = p$

**Expectation of a Uniform random variable:** Assuming the distribution is uniform on $0,1,\dots,n$ we can say that $p_X(x) = \frac{1}{n+1}, \quad \forall x$ this in turn yields the following expectation. (Notice how it tends to the center.)

$$\mathbf{E}(X) = 0 \cdot \frac{1}{n+1} + 1 \cdot \frac{1}{n+1} + \cdots + n \cdot \frac{1}{n+1}$$

$$\mathbf{E}(X) = \frac{1}{n+1}(0+1+\cdots+n) = \frac{1}{n+1} \cdot \frac{n(n+1)}{2} = \frac{n}{2}$$

**Elementary properties of Expectation**

- If $X \geq 0$, then $\mathbf{E}[X] \geq 0$
- If $a \leq X \leq b$, then $a \leq \mathbf{E}[X] \leq b$
- If $c$ is a constant $\mathbf{E}[c] = c$ (deterministic case.)
- **Linearity:** $\boxed{\mathbf{E}[aX+b] = a\mathbf{E}[X] + b}$

**Expected Value Rule:** We calculate expected value of a function of a random variable, $\mathbf{E}[g(X)]$.

- Let $X$ be a r.v. with a known P.M.F and let $Y = g(X)$
- Averaging over $x$:
  $$\boxed{E[Y] = \mathbf{E}[g(X)] = \sum_x g(x) p_X(x)}$$
- **Caution:** $\mathbf{E}[g(X)] = g(\mathbf{E}[X])$ only for *linear functions*, not for *non-linear* ones.

## 3.6 Lecture 6: Random Variables - II

**Variance:** It is the quantity that measures the amount of spread, or the dispersion of probability mass functions. In some ways it quantifies the amount of randomness that is present. Together with the expected value, it summarizes some of the main qualities of a probability mass function.

- Let there be two P.M.F.s with a different spread but same mean or $\mu = \mathbf{E}[X]$. In such a case, if we take just the distance of a random variable from the mean as our dispersion and find its average, we get $\mathbf{E}[X-\mu] = \mathbf{E}[X] - \mu = 0$. Thus we need a better mathematical object to define this dispersion.
- This can be better given by variance, which is defined as
  $$\boxed{\text{var}(X) = \mathbf{E}[(X-\mu)^2]}$$
- If we look at the properties of expectations, we see that $\text{var}(X) \geq 0$
- Calculation using expected value when $g(x) = (X-\mu)^2$
  $$\boxed{\text{var}(X) = \mathbf{E}[g(x)] = \sum_x (x-\mu)^2 p_X(x)}$$

- The squaring makes the variance a bit harder to interpret. Thus we find the **standard deviation**, which is given by $\boxed{\sigma_x = \sqrt{\text{var}(X)}}$

### Properties of variance

- For a linear function of a random variable,

$$\boxed{\text{var}(aX+b) = a^2\text{var}(X)}$$

  Adding a constant does not change the variance. Intuitively, addition of a constant only shifts the probability function.

- Alternative way for calculating variance

$$\boxed{\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2}$$

### Variance of Bernoulli

$$X = \begin{cases} 1, & \text{w.p.} \quad p \\ 0, & \text{w.p.} \quad 1-p \end{cases}$$

$$\mathbf{E}[X] = 1.p + 0.(1-p) = p$$

$$\text{var}(X) = \sum_x (x - \mathbf{E}[X])^2 p_X(x)$$

$$= (1-p)^2 p + (0-p)^2(1-p)$$

$$\boxed{\text{var}(X) = p(1-p)}$$

We can see that when $p = \frac{1}{2}$, the variance, $\text{var}(X) = \frac{1}{4}$ *i.e.* the peak of the parabola. In other words, a coin toss would be most random when a coin is fair.

**Variance of Uniform** Where $n = 0,1,2,\ldots$

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \frac{1}{n+1}\cdot(0^2+1^2+2^2+\cdots+n^2) - \left(\frac{n}{2}\right)^2$$

$$\boxed{\text{var}(X) = \frac{1}{12}n(n+2)}$$

In case, we have a uniform distribution in $[a,b]$ we can say that $n = b-a$ which is nothing but just a shifted distribution. Shifting involves nothing but an addition by a constant and thus, it has no effect on the variance.

### Conditional P.M.F. and Expectation, given an event

Conditional P.M.F can be written as

$$p_{X|A}(x) = P(X = x|A)$$

Conditional expectation is given by

$$\mathbf{E}[X|A] = \sum_x x p_{X|A}(x)$$

All the properties of P.M.Fs and expectations also apply to their conditional counterparts.

**Total Expectation Theorem:** We are given the definition of the total probability theorem. Given we partition our sample space into $n$ disjoint parts, $A_1, A_2, \ldots, A_n$ and another event $B$. We can say that

$$P(B) = P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)$$

We now say that let $B$ be the event that a random variable $X = x$.

$$p_X(x) = P(A_1)p_{X|A_1}(x) + \cdots + P(A_n)p_{X|A_n}(x)$$

We can now multiply both sides by $x$ and sum over all possible values of $x$.

$$\sum_x x p_X(x) = P(A_1)\sum_x x p_{X|A_1}(x) + \cdots + P(A_n)\sum_x x p_{X|A_n}(x)$$

Since each term on the R.H.S is a conditional expectation, we can finally conclude

$$\mathbf{E}[X] = P(A_1)\mathbf{E}[X|A_1] + \cdots + P(A_n)\mathbf{E}[X|A_n]$$

$$\boxed{\mathbf{E}[X] = \sum_i^n P(A_i)\mathbf{E}[X|A_i]}$$

### Conditioning a Geometric Random Variable and Memorylessness

- X: Number of independent coin tosses until first heads, when, $P(H) = p$.
- $p_X(k) = (1-p)^{k-1}, \quad k = 0,1,2,\ldots$
- **Memorylessness:** Independent coin tosses do not have any memory. Or more formally, conditioned on $X > n$, $X - n$ is geometric with the parameter $p$.

$$\boxed{p_{X-n|X>n}(k) = p_x(k)}$$

- Expectation of the geometric distribution can be calculated by

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} kp_X(k) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

- This calculation is hard to calculate and can be done in a simpler manner. We can divide our distribution into two events, $A_1, A_2$, one where our first toss is heads and one where it isn't. This makes their respective probabilities $p, 1-p$. Through the properties of expectation,

$$\mathbf{E}[X] = 1 + \mathbf{E}[X-1]$$

Through total expectation theorem, we can say

$$\mathbf{E}[X-1] = P(A_1)\mathbf{E}[X-1|A_1] + P(A_2)\mathbf{E}[X-1|A_2]$$

Substituting,

$$\mathbf{E}[X] = 1 + p \cdot \mathbf{E}[X-1|x=1] + (1-p)\mathbf{E}[X-1|x>1]$$

By the property of memorylessness we can say that

$$\mathbf{E}[X] = 1 + 0 + (1-p)\mathbf{E}[X]$$

Which finally gives us

$$\boxed{\mathbf{E}[X] = \frac{1}{p}}$$

**Multiple Random Variables and joint P.M.F**

- Suppose our probabilistic model has two random variables $X : p_X$ and $Y : p_Y$ along with their individual P.M.F.s
- Joint P.M.F.s help us capture these two separate kinds of information together and are given by

$$\boxed{p_{XY}(x,y) = P(X = x \text{ and } Y = y)}$$

- If we add over all possible pairs we get

$$\boxed{\sum_x \sum_y p_{XY}(x,y) = 1}$$

- Given a joint P.M.F. we can find the P.M.F.s of individual variables which are also known as the **marginal P.M.F.s**
- Marginal of $x$ is given by

$$\boxed{p_X(x) = \sum_y p_{XY}(x,y)}$$

- Marginal of $y$ is given by

$$\boxed{p_Y(y) = \sum_x p_{XY}(x,y)}$$

- Considering an example of the case of more than two random variables

$$\sum_x \sum_y \sum_z p_{XYZ}(x,y,z) = 1$$

$$p_{XY}(x,y) = \sum_z p_{XYZ}(x,y,z)$$

- **Functions of multiple random variables:**

$$\boxed{p_Z(z) = P(g(X,Y) = z) = \sum_{(x,y):g(x,y)=z} p_{XY}(x,y)}$$

- **Expected Value Rule:**

$$\boxed{\mathbf{E}[g(x,y)] = \sum_x \sum_y g(x,y)p_{XY}(x,y)}$$

**Linearity for expectation in multiple random variables**

- Definition in single variable

$$\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$$

- Additional property in case of multiple variables

$$\boxed{\mathbf{E}[X_1 + \cdots + X_n] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}$$

- A simple proof can be done as follows

$$\mathbf{E}[X+Y] = \mathbf{E}[g(X,Y)]$$

$$= \sum_x \sum_y g(x,y)p_{XY}(x,y) = \sum_x \sum_x (x+y)p_{XY}(x,y)$$

$$= \sum_x \sum_y xp_{XY}(x,y) + \sum_x \sum_y yp_{XY}(x,y)$$

$$= \sum_x x \sum_y p_{XY}(x,y) + \sum_y y \sum_x p_{XY}(x,y)$$

$$= \sum_x xp_X(x) + \sum_y yp_Y(y)$$

$$= \mathbf{E}[X] + \mathbf{E}[Y]$$

This can be followed by induction to generalize for $n$.

**The mean of binomial random variable**

- $X$ : binomial with parameters $n, p$. This can be interpreted as successes in $n$ independent trails with each having a probability $p$ for success.
- Since we know the P.M.F of a binomial, we can obtain an expression for expectation

  $$\mathbf{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

  Since this is a formidable sum to calculate so we use a trick with *indicator variables*,

  $X_i = 1$ if $i^{th}$ trial is a success and
  $X_i = 0$ if it is a failure.

- Counting the number of $X_i = 1$ would give us the number of successes.

  $$X = X_1 + \cdots + X_n$$

  Thus we can say that

  $$\mathbf{E}[X] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]$$

  Since the expectation for a Bernoulli random variable is $p$ we see

  $$\boxed{\mathbf{E}[X] = np}$$

## 3.7 Lecture 7: Random Variables - III

**Conditional P.M.F.s**

- Conditioning on another random variable

  $$p_{X|Y}(x|y) = P(X = x \mid Y = y) = \frac{P(X = x,\ Y = y)}{P(Y = y)}$$

  $$\boxed{p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}}$$

- **Note:** We have a conditional P.M.F. for every possible value of $y$.
- $\sum_x p_{X|Y}(x,y) = 1$
- All the properties of conditional probabilities apply to random variables also. For example, the multiplication rule

  $$\boxed{p_{XYZ} = p_X(x)\ p_{Y|X}(y|x)\ p_{Z|XY}(z|x,y)}$$

**Conditional Expectation in terms of random variables**

- $\boxed{\mathbf{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y)}$
- **Expected value rule:**

  $$\boxed{\mathbf{E}[g(x)|Y = y] = \sum_x g(x) p_{X|Y}(x|y)}$$

- **Total probability Theorem:**

  $$\boxed{p_X(x) = \sum_y p_Y(y)\ p_{X|Y}(x|y)}$$

- **Total Expectation Theorem:**

  $$\boxed{\mathbf{E}[X] = \sum_y p_Y(y)\ \mathbf{E}[X|Y = y]}$$

- These facts are also true for $n \to \infty$, provided that $\mathbf{E}[|X|] < \infty$ i.e. this value is well defined.

**Independence of a random variable:** Useful where probabilistic experiments have no common source of uncertainty.

$$\boxed{p_{XY}(x,y) = p_X(x) \cdot p_Y(y), \quad \forall x, y}$$

**Variance of independent random variables:** This relation does not hold true if the variables are not independent.

$$\boxed{\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)}$$

**Variance of the binomial:**

- $X$ : binomial variable with parameters $n, p$, with $k$ successes in $n$ independent trials.
- Let $X_i = 1$, if $i^{\text{th}}$ trial was a success and $X_i = 0$ otherwise, be indicator events. Thus we can say $X = X_1 + \cdots + X_n$.
- $\text{var}(X) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n\text{var}(X_1)$

  $$\boxed{\text{var}(X) = n(p)(1-p)}$$

**Note:** The Hat Problem (Done).

**Inclusion Exclusion Principle**
$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) = \sum_i \mathbf{P}(A_i) - \sum_{i_1<i_2} \mathbf{P}(A_{i_1} \cap A_{i_2}) + \sum_{i_1<i_2<i_3} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n-1}\mathbf{P}(\bigcap_{k=1}^{n} A_k)$

**Variance of Geometric P.M.F.**

- Given $X > 1$, $X-1$ has the same geometric P.M.F.
- Given $X > 1$, conditional P.M.F. of $X$: same as unconditional P.M.F. of $X+1$.

$$\mathbf{E}[X|X>1] = 1 + \mathbf{E}[X]$$

$$\mathbf{E}[X^2|X>1] = \mathbf{E}[(X+1)^2]$$

- **Divide and Conquer:** By the virtue of total expectation theorem, we can say

$$\mathbf{E}[X^2] = \mathbf{P}(X=1).\mathbf{E}[X^2|X=1] + \mathbf{P}(X>1).\mathbf{E}[X^2|X>1]$$

$$= p.1 + (1-p)(\mathbf{E}[X^2] + 2\mathbf{E}[X] + 1)$$

$$= p + (1-p)(\mathbf{E}[X^2] + \frac{2}{p} + 1)$$

- Using algebra, we get

$$\boxed{\mathbf{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}}$$

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

$$\boxed{\text{var}(X) = \frac{1-p}{p^2}}$$

**Independent Events vs Random Variables**
$n$ events are said to be independent if their associated indicator variables are independent.

## 3.8 Lecture 8: Continuous Random Variables - I

**Probability Density Functions:** The only difference from the discrete case is that these are defined on real values and probabilities of individual points are zero.

- $\boxed{\mathbf{P}(a \le X \le b) = \int_a^b f_X(x)dx}$
- $f_X(x) \ge 0$
- $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1$
- $\mathbf{P}(X = x) = 0$
- **Note:** Inclusion or exclusion of end points does not really matter.

**Uniform Random Variable:**

- Intuitively this can be thought of as the area under a straight line between two points, $a, b$. Equal portions of the line span equal area, and thereby have the same probabilities. The height of the rectangle formed is $\frac{1}{b-a}$.
- A more general P.D.F. is the piece wise constant. Which has many such rectangles. One thing to note is a P.D.F. doesn't necessarily have to be continuous.

**Expectation/mean of a continuous random variable:** Like we defined the expectation as a weighted summation in the discrete case, here it becomes a weighted integral.

- $\boxed{\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx}$
- **Assumption:** The expectation needs to me mathematically well defined.

$$\int_{-\infty}^{\infty} |x|\ f_X(x)\ dx < \infty$$

- All the properties of expectation also hold true for the continuous case, including the **linearity property.**
- **Expected Value Rule:** $\boxed{\int_{-\infty}^{\infty} g(x) f_X(x)dx}$
- All the definitions and properties of **variance** are also the same.

**Mean and Variance of Uniform Distribution**

- $\mathbf{E}[X] = \int\limits_{-\infty}^{\infty} x f_X(x)dx = \int\limits_{a}^{b} x \frac{1}{b-a} dx = \boxed{\frac{a+b}{2}}$
- $\mathbf{E}[X^2] = \int\limits_{-\infty}^{\infty} x f_X(x)dx = \int\limits_{a}^{b} x^2 \frac{1}{b-a} dx = \frac{1}{b-a}\left(\frac{b^3-a^3}{3}\right)$

- $\mathrm{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \boxed{\dfrac{(b-a)^2}{12}}$
- $\sigma = \sqrt{\mathrm{var}(X)} = \boxed{\dfrac{b-a}{2\sqrt{3}}}$

### Exponential Random Variable

- It's probability density function is determined by a single parameter $\lambda \in \mathbb{Z}^+$.
- The form of the P.D.F. is as shown here

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

- Also called **inverse exponential**, it starts at $\lambda$ and decays at the rate of $\lambda$. The larger $\lambda$ is, the faster it decays.
- The shape of this distribution is pretty similar to the geometric distribution, only that the latter is discrete.
- The tail probability of an exponential random variable is given by

$$\mathbf{P}\{X \geq a\} = \int_a^\infty \lambda e^{-\lambda x} dx = \boxed{e^{-\lambda a}}$$

  Putting $a = 0$, we can verify this is a P.D.F. since it then equals to 1.

- The expectation of this random variable is given by

$$\mathbf{E}[X] = \int_0^\infty x.\lambda e^{-\lambda x} dx = \boxed{\frac{1}{\lambda}}$$

- The variance of exponential random variable is given by

$$\mathbf{E}[X^2] = \int_0^\infty x^2.\lambda e^{-\lambda x} dx = \boxed{\frac{2}{\lambda^2}}$$

  Thus we can calculate

$$\mathrm{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \boxed{\frac{1}{\lambda^2}}$$

- Has the same property of **memorylessness**.
- Generally helps us model the time we have to wait for something to occur in a process, in real world phenomena.

### Cumulative Distribution Functions

- **Continuous case**

$$\boxed{F_X(x) = \mathbf{P}\{X \leq x\} = \int_{-\infty}^x f_X(t)dt}$$

- **Discrete case**

$$\boxed{F_X(x) = \mathbf{P}\{X \leq x\} = \sum_{t \leq x} f_X(t)}$$

  This usually takes the form of a **staircase function** where the size of the jump is equal to the corresponding value of the P.M.F.

- C.D.F.s are useful because they have a lot of information about the random variable and we can calculate pretty much anything that we want to. For the continuous case, it can even recover the P.D.F.
- **Calculating a P.D.F**

$$\boxed{\frac{dF_X(x)}{dx} = f_X(x)}$$

  **Note:** This formula will only be correct for places where the derivative exists.

### General Properties of C.D.F

- It is a **non-decreasing** function *i.e.*

$$\boxed{y \geq x \Rightarrow F_X(y) \geq F_X(x)}$$

- We can say the following about the asymptotic limits of cumulative distributive functions

$$\boxed{\lim_{x\to\infty} F_X(x) = 1}$$

$$\boxed{\lim_{x\to-\infty} F_X(x) = 0}$$

### Normal (Gaussian) Random Variables

- **Importance of Normal Random Variables**
  - They play a key role in probability theory, like in the case of *central limit theorem*.
  - They have nice analytical properties that make them useful in various applications.

  - They are a good model of randomness or noise, whenever that noise is due to the addition of small independent noise terms.

- **The Standard Normal Form**

$$\boxed{N(0,1): \quad f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}}$$

- It's expectation is $\mathbf{E}[X] = 0$ since the distribution is symmetric. The variance can be found by integrating by parts $\text{var}(X) = 1$. This gives us the notation we have been using $N(0,1)$.

- **General Normal Form** Where $\sigma$ is a positive parameter.

$$\boxed{N(\mu,\sigma^2): \quad f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}}$$

  This curve will not necessarily be symmetric around zero and we will have some control over it's width.

- $\mu$ is the point about which the curve is symmetric, which by definition makes it the mean $\mathbf{E}[X] = \mu$. The variance can be found out through calculus and turns out to be $\sigma^2$.

- A nice property of these distributions is that they behave elegantly when we form linear functions of them making them analytically convenient.

- Let $Y = aX + b$, where $X \approx N(\mu,\sigma^2)$. Following linearity we can say that $\mathbf{E}[Y] = a\mu + b$ and $\text{var}(Y) = a^2\sigma^2$. But the most important fact is that $Y$ is also a normal random variable.

$$\boxed{Y \approx N(a\mu + b, a^2\sigma^2)}$$

  **Note:** For $a = 0$, $Y = b$ is a discrete and degenerate normal random variable. (convention)

- C.D.F. of the standard normal form is denoted by

$$\varphi(y) = F_Y(y) = \mathbf{P}\{Y \le y\}$$

- Expressing a normal random variable in standard form.

$$\boxed{X = \mu + \sigma Y}$$

  (Linearity, an amazing thing.)

## 3.9 Lecture 9: Continuous Random Variables - II

**Conditional P.D.F.s**

$$f_{X|A}(x).\delta \approx \mathbf{P}\{\, x \le X \le x + \delta \mid A \,\}$$

$$\boxed{\mathbf{P}\{\, X \in B \mid A \,\} = \int_B f_{B|A}(x)\, dx}$$

**Conditional P.D.F when $X \in A$**

$$\boxed{f_{X|X\in A}(x) = \begin{cases} 0, & \text{if } x \notin A \\ \frac{f_X(x)}{\mathbf{P}\{A\}}, & \text{if } x \in A \end{cases}}$$

**Conditional Expectation**

$$\boxed{\mathbf{E}[X|A] = \int x f_{X|A}(x)\, dx}$$

**Expected Value Rule**

$$\boxed{\mathbf{E}[g(X)|A] = \int g(x)\, f_{X|A}(x)\, dx}$$

**Memorylessness of Exponential Random Variable:** We will establish this property by the means of an example. Suppose that it is known that life bulbs have a lifetime until they burn out which is given by an exponential random variable. Given a choice to buy a new bulb or an old one, which one should you buy?

- Let $T$ : exponential($\lambda$) denote the life time of the bulb, a random variable, exponential with some given parameter $\lambda$.

- From our earlier calculation, we know

$$\mathbf{P}\{T > x\} = e^{-\lambda x}, \quad \forall x \ge 0$$

- We are told a certain light bulb is operating for some $t < T$ without failing. We can now have a random variable $X = T - t$ which depicts the time remaining till the light bulb burns out.

- Now we will find out the our used bulb last atleast $x$ units more, or in other words, the light bulb remains alive for $t + x$ units.

$$\mathbf{P}\{X > x | T > t\} = \frac{\mathbf{P}\{T - t > x, T > t\}}{\mathbf{P}\{T > x\}}$$

$$= \frac{\mathbf{P}\{T > t+x, T > t\}}{\mathbf{P}\{T > x\}} = \frac{\mathbf{P}\{T > t+x\}}{\mathbf{P}\{T > x\}}$$

$$= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = \boxed{e^{-\lambda x}}$$

Since this is the same as the probability of a new bulb running for $x < T$, we can say that their lives are still probabilistically the same.

**Consequence of memorylessness**

- We know that the p.d.f. is defined by

  $$f_T(x) = e^{-\lambda x}, \quad \text{for } x \geq 0$$

  From which we can calculate the probability that $T$ lies in a small interval.

  $$\mathbf{P}\{0 \leq T \leq \delta\} = f_T(0)\delta = \lambda\delta, \quad \text{for } \delta \to 0$$

- Now we are told that the bulb has been alive for $t$ time units, then the probability that it burns out during the next $\delta$ time units is given by

  $$\mathbf{P}\{t \leq T \leq t+\delta \mid T > t\}$$

  Since, a still alive bulb is probabilistically identical to a new one, this conditional probability is the same as the non-conditional one given above.

  $$\mathbf{P}\{t \leq T \leq t+\delta \mid T > t\} = \boxed{\lambda\delta}$$

- This can be interpreted as there being independent events after every $\delta$ time steps, each having

  $$\boxed{\mathbf{P}\{\text{success}\} = \lambda\delta}$$

  making it a continuous analog of the geometric random variable.

**Total Probability Theorem**

$$\boxed{f_X(x) = \mathbf{P}\{A_1\}f_{X|A_1}(x) + \cdots + \mathbf{P}\{A\}f_{X|A_n}(x)}$$

**Total Expectation Theorem** Multiplying by $x$ and integrating both sides we get

$$\boxed{\mathbf{E}[X] = \mathbf{P}\{A_1\}\mathbf{E}[X|A_1] + \cdots + \mathbf{P}\{A_n\}\mathbf{E}[X|A_n]}$$

**Mixed Distributions**

- Given that we have 1 dollar and the opportunity to play in a lottery.

  $$X = \begin{cases} \text{uniform on } [0,2], & \text{with probability } 1/2 \\ 1, & \text{with probability } 1/2 \end{cases}$$

  **Note:** This variable cannot be continuous since at $X = 1$, the probability is not 0 and it is not discrete because it takes values in a continuous range also, thus it is called a **mixed random variable.**

- C.D.F.s used to define such variables. $X$ is mixed when

  $$X = \begin{cases} Y, & \text{discrete with probability } p \\ Z, & \text{continuous with probability } 1-p \end{cases}$$

  Using the total probability theorem, we can write C.D.F as

  $$F_X(x) = p.\mathbf{P}\{Y \leq x\} + (1-p).\mathbf{P}\{Z \leq x\}$$

  $$\boxed{F_X(x) = p.F_Y(x) + (1-p).F_Z(x)}$$

- Expectation can be given by

  $$\boxed{\mathbf{E}[X] = p.\mathbf{E}[Y] + (1-p).\mathbf{E}[Z]}$$

**Joint continuous random variables and joint P.D.F.s:** The continuous analogue of the joint P.M.F.s given by $f_{XY}(x,y)$. The probability of two events $X$ and $Y$ occurring can then be given by

$$\boxed{\mathbf{P}\{(X,Y) \in B\} = \iint_{(x,y)\in B} f_{XY}(x,y)\, dx\, dy}$$

**Def.** Two random variables are called **jointly continuous** if they can be described by a joint P.D.F.

Given that we integrate over a small area that varies with the parameter $\delta$, we get an interpretation of joint P.D.F.s in terms of probabilities of small rectangular areas.

$$\boxed{\mathbf{P}\{a \leq X \leq a+\delta,\ c \leq Y \leq c+\delta\} \approx f_{XY}(a,c).\delta^2}$$

$f_{XY}(x,y)$ : it is a probability density or the probability per unit area. Analogously, we can also say that *points* and *curves* will have 0 probability, since the volume under then is 0.

**Warning:** Looking at the case where $X = Y$, we can see that their joint probabilities can be defined by a straight line. This is the case where they are individually continuous but not jointly continuous. (Implications of linear dependence of sub spaces.)

**Marginal probability functions**

$$\boxed{f_X(x) = \int f_{XY}(x,y)\ dy}$$

$$\boxed{f_X(x) = \iint f_{XYZ}(x,y,z)\ dy\ dz}$$

**Expected Value Rule**

$$\boxed{\mathbf{E}[g(X,Y)] = \iint g(x,y)\ f_{XY}(x,y)\ dx\ dy}$$

**Note:** All the properties off discrete hold true here, some with minor changes.

**Joint C.D.F.**

$$\boxed{F_{XY}(x,y) = \mathbf{P}\{x \leq X, y \leq Y\} = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{XY}(x,y)\ dx\ dy}$$

$$\boxed{f_{XY}(x,y) = \frac{\partial^2 F_{XY}}{\partial x \partial y}(x,y)}$$

**Buffon's Needle and Monte Carlo Simulation**

- **Problem:** We have an infinite plane with parallel lines $d$ distance apart. Given a needle of length $l$ and we throw it at random on the plane. What is the probability that the needle intersects a line?
- **Assumption:** We will assume $l < d$.
- **Model the experiment:** This involves
  - Defining a *sample space.*
  - Defining the relevant *random variables.*
  - Choosing the right *probability function.*
  - Identifying the *event of interest.*
  - *Calculating* the result.
- The sample space of our experiment will be defined by all the possible combinations of the two random variables $(X, \Theta)$. Where $X$ is measuring the vertical distance from the center of the needle to the nearest line and $\Theta$ is measuring the acute angle that the needle makes when it falls on the line or is extended to meet it.
- The limits of random variables that we can infer from observation.

$$X: \quad 0 \leq x \leq \frac{d}{2}$$

$$\Theta: \quad 0 \leq \theta \leq \frac{\pi}{2}$$

- A reasonable probability function to model randomness is the uniform distribution where each outcome of the sample space ($(X, \Theta)$ pair) is equally. (Below are obvious results using the formula to calculate uniform probabilities.)

$$X: \quad f_X(x) = \frac{2}{d}$$

$$\Theta: \quad f_\Theta(\theta) = \frac{2}{\pi}$$

- **Assumption:** $X$ and $\Theta$ are independent. This is again reasonable as the distance from a line has nothing to do with the orientation after the needle has fallen and vice versa.
- We can thus easily define a joint P.M.F for our model as follows

$$f_{X\Theta}(x,\theta) = f_X(x).f_\Theta(\theta) = \frac{4}{\pi d}$$

This concludes our probabilistic model.

- **Identifying Problem of Interest:** We want to calculate **P**{intersection}, this can happen if the projection of the needle is larger than or equal the vertical distance from the line $x \leq \frac{l}{2} sin\theta$ or more generally

$$X \leq \frac{l}{2} sin\Theta$$

- Now all that is left is calculation for which we need to plug in the limits in our joint probability function.

$$\mathbf{P}\{X \leq \frac{l}{2} sin\Theta\} = \int_0^{\frac{pi}{2}} \int_0^{\frac{l}{2} sin\theta} \frac{4}{\pi d}\ dx\ d\theta = \boxed{\frac{2l}{\pi d}}$$

- Suppose we had to estimate $pi$, we have the information for $d$ and $l$. We can then think of performing a *simulation* where we toss a needle $n$ times and then take the probability as the frequency of intersections over this number. This information can help us approximate $pi$ with great accuracy. Such simulation techniques are called **Monte Carlo Methods.**

- These methods can be used to determine quantities where conventional methods prove to be very complex to solve.

## 3.10 Lecture 10: Continuous Random Variables - III

**Conditional P.D.F.s given another random variable**

- **Definition**

$$\boxed{f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}, \quad \text{if } f_Y(y) > 0}$$

- Given the conditional definition for events, we can just substitute the event $A$ by a random variable $Y \approx y$.

$$\mathbf{P}\{\, x \le X \le x+\delta \;\mid y \le Y \le y+\epsilon \,\}$$

$$= \frac{f_{XY}(x,y)\delta\epsilon}{f_Y(y)\epsilon} = \boxed{f_{X|Y}(x|y).\delta}$$

- To get past the restriction of not being able to define conditional probabilities when$f_Y(y) = 0$ we use this definition

$$\boxed{\mathbf{P}\{\, X \in A \mid Y = y \,\} = \int_A f_{X|Y}(x|y)\,dx}$$

**Total Probability Theorem**

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y)\,dy}$$

**Conditional Expectation**

$$\boxed{\mathbf{E}[X|Y=y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\,dx}$$

**Total Expectation Theorem**

$$\boxed{\mathbf{E}[X] = \int_{-\infty}^{\infty} f_Y(y)\,\mathbf{E}[X|Y=y]\,dy}$$

**Expected Value Rule**

$$\boxed{\mathbf{E}[g(x)|Y=y] = \int_{-\infty}^{\infty} g(x)\, f_{X|Y}(x|y)\,dx}$$

**Independence**

$$\boxed{f_{XY} = f_X(x).f_Y(y)}$$

Or in an equivalent manner

$$f_{X|Y}(x|y) = f_X(x)$$

**Consequences of Independence** Same as discrete case

- $\mathbf{E}[XY] = \mathbf{E}[X].\mathbf{E}[Y]$
- $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$
- $\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(x)].\mathbf{E}[h(Y)]$

**Stick Breaking Example**

- Given a stick of length $l$, we break it at some random location, which corresponds to some random variable

$$X: \quad \text{uniform in } [0,l]$$

- We now break it again as follows

$$Y: \quad \text{uniform in } [0,X]$$

  which means that the conditional distribution is uniform.

- From these two we can clearly see that

$$f_X(x) = \frac{1}{l}$$

$$f_{Y|X}(y|x) = \frac{1}{x}$$

  Using these two we can calculate the joint p.d.f. as follows

$$f_{XY}(x,y) = f_{Y|X}(y|x).f_X(x)$$

$$\boxed{f_{XY}(x,y) = \frac{1}{lx} \quad 0 \le y \le x \le l}$$

- Now, we wish to calculate the marginal of $y$, (notice limits of $x$)

$$\boxed{f_Y(y) = \int_x f_{XY}(x,y)\,dx = \int_y^l \frac{1}{lx}\,dx = \frac{1}{l}\log\left(\frac{l}{y}\right)}$$

- Calculation of expectation

$$\mathbf{E}[Y] = \int_0^l y\,\log\left(\frac{l}{y}\right)\,dy$$

- (Alternative) Using Total expectation theorem

$$\mathbf{E}[Y] = \int_x f_X(x)\,\mathbf{E}[Y|X=x]\,dx$$

We know that the conditional expectation given above is on a uniform distribution thus it is nothing but one-half of the range *i.e* $\frac{x}{2}$. Thus

$$\mathbf{E}[Y] = \int_0^l \frac{1}{l}.\frac{x}{2}\,dx$$

- Although the integration is straight forward, we can think of it as

$$\boxed{\frac{1}{2}\int_0^l f_X(x).x\,dx = \frac{1}{2}\mathbf{E}[X] = \frac{l}{4}}$$

- Intuitively, if we break a stick twice, the first expected value is bound to be half the length and the second one, the fourth of the original length.

### Independent Standard Normals

- Noise that shows up at different parts of the system is often considered independent and modeled by Normal distributions.
- Given two independent standard normal distributions, we can write their joint probability density function as

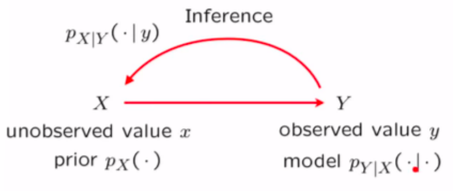
$$\boxed{f_{XY}(x,y) = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}}$$

For any circle represented by $x^2+y^2=r^2$, the joint p.d.f. takes a constant value. These circles are known as **contours**.

### Independent Normals

- Considering two independent normal random variables, we can write their joint p.d.f. as

$$\boxed{\frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right\}}$$



### Bayes' rule - a theme with variations

- Think of $X$ as an unknown state of the world and $Y$ being a noisy observation of $X$. The conditional p.m.f tells us the distribution of $Y$ under each possible state of the world. Once we observe the value of $Y$, we obtain some information about $X$ and we use this information about the likely values of $X$. Mathematically, rather than relying on a prior for $X$, we for its conditional P.M.F. given the particular observation we have seen.
- **Bayes' rule for random variables**

$$\boxed{p_{X|Y}(x|y) = \frac{p_X(x)\,p_{Y|X}(y|x)}{p_Y(y)}}$$

The conditional distribution of $X$ is called the **posterior** for which we rely on the **prior** of $X$.

- Here the marginal of $Y$ or $p_Y(y)$ can be found with the help of join probability mass function which is already given to us

$$\boxed{\sum_{x'} p_X(x')\,p_{Y|X}(y|x')}$$

- The same arguments can be made for the case of continuous random variables.

**Mixed Bayes' Rule:** Here $K$ is a discrete random variable and $Y$ is a continuous random variable.

$$\boxed{p_{K|Y}(k|y) = \frac{p_K(k)\,f_{Y|K}(y|k)}{f_Y(y)}}$$

$$\boxed{f_Y(y) = \sum_{k'} p_K(k')\, f_{Y|K}(y|k')}$$

This is specially useful when the noise is continuous whereas our data is discrete. The reverse can also be done by adjusting the formula accordingly.

$$\boxed{f_{Y|K}(y|k) = \frac{f_Y(y)\, p_{K|Y}(k|y)}{p_K(k)}}$$

$$\boxed{\int_y f_Y(y')\, p_{K|Y}(k|y')\, dy'}$$

**Detection of a Binary Signal**