Intro to Probability

(Quick Reference)

Parameswaran Raman • 2015 • UC Santa Cruz

Last Revision: December 19, 2015

Table of Contents

Abstract

I prepared these notes as a resource to help myself; and anyone else interested, in quickly reviewing concepts in this course. My goal is to provide a very concise, end-to-end resource that covers all the important material discussed in a course on this topic. If you spot any errors or have any suggestions, please contact me directly at params@ucsc.edu.

Thanks for reading!

1 Basics of Probability

1.1 Definition of Probability and Notations

Notations:

- *Sample Space*: Collection of all possible outcomes. Denoted by S or Ω.
- *Outcome*: An element of the sample space S. Denoted by $s \in S$.
- *Event*: Subset of the sample space $A \subset S$ that defines an experiment of interest. In other words, event is a set of outcomes.
- *Finite, Infinite Sets*: Finite sets are those with a discrete (integer) number of elements. Infinite Sets can be either *countable* (those with a one-one correspondence with $\{1, 2, 3, \ldots, \}$) or *uncountable.*

Example: Consider the experiment involving tossing a fair coin twice and let the event of interest be defined as "obtaining at least one head".
Then, Sample Space, $S = \{HH, HT, TH, TT\}$.
Outcome is any possible element from S. For e.g.: HH, TH, etc.
Event, A= $\{HH, HT, TH\}$.

Disjoint Events A and B are disjoint or "mutually exclusive" if A and B have no outcomes in common (i.e.: $A \cap B = \emptyset$).

Disjoint Events (generalization): $A_1, \ldots, A_n$ is a collection of disjoint events iff:

$$A_i \cap A_j = \emptyset \quad \forall i, j \mid i \neq j$$

De-Morgan's Laws:

$$\left(\bigcup_{i \in I} A_i\right)^C = \bigcap_{i \in I} A_i^C$$

$$\left(\bigcap_{i \in I} A_i\right)^C = \bigcup_{i \in I} A_i^C$$

Probability: Function over S that measures the likelihood of events. A probability for an event A on a set S is a specification of the measure $\Pr(A)$ such that the below axioms are satisfied.

- **Axiom 1**: For every A, $\Pr(A) \geq 0$
- **Axiom 2**: $\Pr(S) = 1$
- **Axiom 3**: For every *infinite* sequence of "disjoint" events, $A_1, A_2, \ldots, A_n$, $A_i \subset S$,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr\left(A_i\right)$$

- **Remarks**: Axiom 3 also holds for n *finite* events, but only when "disjoint",

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr\left(A_i\right)$$

1.2 Other basic properties

- If $A \subset B$, then $\Pr(A) \leq \Pr(B)$
- For every A, $0 \leq \Pr(A) \leq 1$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- $\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B) + \Pr(C) - \Pr(A \cap B) - \Pr(B \cap C) - \Pr(A \cap C) + \Pr(A \cap B \cap C)$
- **Bonferroni Inequality**: For any events $A_1, \dots, A_N$,

$$\Pr\left(\bigcap_{i=1}^{N} A_i\right) \geq 1 - \sum_{i=1}^{N} \Pr\left(A_i^C\right)$$

- **Boole's Inequality**: For any events $A_1, \dots, A_N$,

$$\Pr\left(\bigcup_{i=1}^{N} A_i\right) \leq \sum_{i=1}^{N} \Pr\left(A_i\right)$$

2 Counting Methods

Multiplication Rule: Suppose an experiment has k parts ($k \geq 2$); such that the i-th part of the experiment has n_i possible outcomes ($i = 1, \dots, k$), and that *all outcomes in each part can occur regardless of which specific outcomes have occurred in the other parts*. Then, the sample space S will contain vectors of the form $(u_1, u_2, \dots, u_k)$, where u_i is one of the possible outcomes. The total number of vectors is $n_1 n_2 \dots n_k$.
This can be diagrammatically seen below:

Permutations: Suppose k items have to be selected from n items. Each outcome is a "permutation" of k items and the total number of permutations is given by:

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

Remarks:

- when sampling without replacement: Total number of permutations is $P_{n,k}$
- when sampling with replacement: Total number of permutations is n^k

Combinations: Suppose k items have to be chosen from n items. This can be done in number of ways given by:

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

Remark:

- $\binom{n}{k}$ are called the binomial coefficients and appear in the binomial theorem:

$$(x+y)^n = \sum_{k=1}^{n} \binom{n}{k} x^k y^{n-k}$$

- The binomial coefficient $\binom{10}{3}$ is equivalent to the binomial coefficient $\binom{10}{7}$

Multinomial Coefficients: Provides a way of splitting n elements into k ($k \geq 2$) groups such j-th group gets n_j elements, and $\sum_{j=1}^{n} n_j = n$. The number of ways such a split can be accomplished is given by:

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

Remarks:

- $\binom{n}{n_1, n_2, \ldots, n_k}$ are called the multinomial coefficients and appear in the multinomial theorem (which is a generalization of the binomial theorem)
- The binomial coefficient $\binom{10}{3}$ is equivalent to the multinomial coefficient $\binom{10}{3,7}$

3 Conditional Independence

Multiplication Rule: In general for n events,

$$\Pr(A_1 \cap A_2 \cap \dots A_n) = \Pr(A_1). \Pr(A_2|A_1). \Pr(A_3|A_1 \cap A_2) \dots \Pr(A_n|A_1 \cap A_2 \dots A_{n-1})$$

Independence of Events: Knowing that one event has occurred does not influence occurrence of another event. ie. A and B are independent if:

$$\Pr(A|B) = \Pr(A)$$
$$\Pr(B|A) = \Pr(B)$$

This also implies:

$$\Pr(A \cap B) = \Pr(A) \Pr(B)$$

Remarks:

- If A and B are however not independent, then in order to expand the intersection, the general multiplication rule (shown above) has to be followed.
- Disjoint events are not the same as Independent events

3.1 Conditional Independence

Given n events, then $A_1, \dots, A_n$ are conditionally independent given B, with $\Pr(B) > 0$, if:

$$\Pr\left(\bigcap_{i \in I} A_i|B\right) = \prod_{i \in I} \Pr(A_i|B)$$

where, I represents every subset of $\{1, 2, \dots, k\}$.

For example, events A, B and C are conditionally independent given B ($\Pr(B) > 0$) if:

- $\Pr(A_1 \cap A_2|B) = \Pr(A_1|B) \Pr(A_2|B)$
- $\Pr(A_2 \cap A_3|B) = \Pr(A_2|B) \Pr(A_3|B)$
- $\Pr(A_1 \cap A_3|B) = \Pr(A_1|B) \Pr(A_3|B)$
- $\Pr(A_1 \cap A_2 \cap A_3|B) = \Pr(A_1|B) \Pr(A_2|B) \Pr(A_3|B)$

3.2 Other basic properties

- **Generalization to independence of n events**: Let $A_1, \dots, A_k$ be a collection of events. These events are independent if for every subset $A_{i1}, \dots A_{ij}$ of j of these events ($j = 1, 2, \dots, k$,

$$\Pr(A_{i1} \cap \dots \cap A_{ij}) = \Pr(A_{i1}) \dots \Pr(A_{ij})$$

- If A, B are independent, then the following independence relations hold:
  - A and B^C are independent
  - A^C and B are independent
  - A^C and B^C are independent

- Let A_1, A_2 and B be events such that $\Pr(A \cap B) > 0$. Then, A_1 and A_2 are conditionally independent given B iff:

  $$\Pr(A_2|A_1 \cap B) = \Pr(A_2|B)$$

  or,

  $$\Pr(A_2|A_1, B) = \Pr(A_2|B)$$

3.3 Baye's Theorem

Partitions: If $B_i \subset S$ such that $B_i \cap B_j \neq 0 \quad \forall i \neq j$ and $\bigcup_{i=1}^{k} B_i = S$, then the events B_i are called *partitions*. Partitions are useful when dealing with conditional probability and form the basis of the Baye's Theorem. They can be visualized by the diagram below:

Law of Total Probability: Given partitions $B_i \quad (i = 1, 2, \dots, K)$ for a set A,

$$\Pr(A) = \underbrace{\Pr\left(\bigcup_{i=1}^{K}(A \cap B_i)\right) = \sum_{i=1}^{K} \Pr(A \cap B_i)}_{\text{union of disjoint events}} = \sum_{i=1}^{K} \Pr(B_i)\Pr(A|B_i)$$

Baye's Theorem: Let $B_1, \dots, B_K$ are partitions of S such that $\Pr(B_j) > 0, \quad (j = 1, 2, \dots K)$. Further, assume another event A such that $\Pr(A) > 0$. Then,

$$\Pr(B_i|A) = \frac{\Pr(B_i)\Pr(A|B_i)}{\underbrace{\sum_{j=1}^{K} \Pr(B_j)\Pr(A|B_j)}_{\text{law of total probability}}} = \frac{\Pr(B_i)\Pr(A|B_i)}{\Pr(A)}$$

Remarks:

- Baye's Theorem can be thought of a tool used to reverse the conditional probability. ie. given $\Pr(A|B)$, it provides a way to compute $\Pr(B|A)$ (assuming the assumptions of partitions hold).
- In simple terms of two events A and B, Baye's Theorem says:

  $$\Pr(B|A) = \frac{\Pr(B)\Pr(A|B)}{\Pr(A)}$$

4 Random Variables

4.1 Random Variables

A *random variable* X is a real-valued function on S that assigns a real number $X(S) = x$ to each possible outcome, $s \in S$. In other words,

$$X : S \to D$$

where, D could be countably-finite (corresponding to *discrete* random variables) or countably-infinite (corresponding to *continuous* random variables). Note that X above refers to the random variable whereas x refers to its realization (or value that it takes).

Example: Consider the event of tossing a coin three times and obtaining \$1 for each head and $-$\$1 for each tail. Let X denote the random variable representing the amount of money that can be earned. The various outcomes possible for such an experiments along with the possible values that the random variable X can take is summarized in the table below:

Outcome	$X = x$
1 1 1	3
1 1 0	1
1 0 0	-1
1 0 1	$\ldots$
$\ldots$	$\ldots$

ie: $X : S \to D = \{-3, -1, 1, 3\}$
Note that, here the elements of the sample set S are all equally likely, but in general they do not have to be (e.g.: consider an unbalanced coin).

4.2 Discrete and Continuous Probability Functions (p.m.f, p.d.f)

Random Variables are characterized by *probability distributions.*

Discrete Probability Distributions: A random variable X has a discrete distribution (termed as *probability mass function*), or X is a discrete random variable if it takes at most a countable number of values. The probability function (p.f) or probability mass function (p.m.f)[1] of a discrete random variable X is:

$$f_X(x) = \Pr(X = x)$$

Properties:

- *Probability values are bounded between 0 and 1*: $0 \leq f_X(x) \leq 1$
- *Probability mass lies entirely inside the domain*: $f_X(x) = 0 \quad \forall x \notin D$
- *Probability mass must add upto 1*: $\sum_{x \in D} f_X(x) = 1$
- *Definition of probability associated with an event*: $\Pr(X \in A) = \sum_{x \in A} f_X(x)$

[1] In several places, probability function (p.f) is used as a generic terminology to denote both p.m.f and p.d.f

Example I: Consider a random variable X following the Uniform Distribution,this means the the r.v. takes the values $X = x$, $x = \{1, 2, \ldots, K\}$, where all values of x are equally likely. The p.f. is given by:

$$f_X(x) = \Pr(X = x) = \begin{cases} \frac{1}{k}, & x = 1, 2, \ldots, K \\ 0, & \text{otherwise} \end{cases}$$

Example II: Bernoulli Distribution Suppose an event A happens with probability p. Let X be the random variable that denotes such an event.

$$X = \begin{cases} 1, & \text{if } A \text{ happens} \\ 0, & \text{if } A^C \text{ happens} \end{cases}$$

then the probability distribution of X is given by:

$$f_X(x) = \begin{cases} (1-p), & \text{when } x = 0 \\ p, & \text{when } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

Continuous Probability Distributions: A random variable X has a continuous probability distribution if there exists a non-negative function f (called the *probability density function*) such that:

$$\Pr(a \le X \le b) = \int_a^b f(x)dx$$

This is also shown in the picture below:
Note that if X is a continuous random variable, $\Pr(X = x) = 0$.

Properties:

- *Density is non-negative*: $f(x) \ge 0$
- *Density has to add up to one*: $\int_{-\infty}^{\infty} f(x)dx = 1$

Example III: Provide an example.

4.3 Cumulative Density Functions (c.d.f)

Cumulative Density Function for any random variable X is given by:

$$F(x) = \Pr(X \le x)$$

Properties:

- $0 \le F(x) \le 1 \quad \forall x$
- *$F(x)$ is non-decreasing*:
  If $x_1 < x_2$, $\implies \{X \le x_1\} \subset \{X \le x_2\}$, then $\Pr(X \le x_1) \le \Pr(X \le x_2) \implies F(x_1) \le F(x_2)$
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
- c.d.f need *not be continuous* (infact, for discrete distributions they will not be so).
- *Derivatives of cdf's are the pdf's* (this applies to continuous random variables only; for discrete random variables the derivative doesn't exist).

4.4 Quantile Function

Given a random variable X, $F^{-1}(p)$ is defined as the p-th quantile of X, and $F^{-1}(.)$ is called the *quantile* function.

Remarks:

- *median* is a type of quantile.
- In the case of discrete random variables, the p-quantile is the smallest x such that $F(x) \geq p$.

4.5 Joint Density Functions

Discrete joint p.f: In case of discrete random variables X and Y, the joint p.m.f is given by:

$$f_{X,Y}(x, y) = \Pr(X = x, Y = y)$$

and,

$$\Pr((X, Y) \in A) = \sum_{(X,Y)} f_{X,Y}(x, y) = 1$$

Continuous joint p.f: In case of continuous random variables X and Y, the joint p.d.f is given by:

$$f_{X,Y}(x, y) = f(x, y)$$

and,

$$\Pr((X, Y) \in A) = \iint\limits_{A} f(x, y)dxdy$$

Properties:

- $f(x, y) \geq 0 \quad \forall x, y$
- $\iint\limits_{\mathbb{R}^2} f(x, y)dxdy = 1 \Leftrightarrow \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x, y)dxdy$. Intuitively, this denotes the probability of the region corresponding to the volume below the surface of a region.

Discrete Joint c.d.f: Extending the notion of c.d.f's to joint distributions (involving X and Y), joint c.d.f is defined as:

$$F(x, y) = \Pr(X \leq x, Y \leq y)$$

Continuous Joint c.d.f: For continuous random variables X and Y, the joint c.d.f is given by:

$$F(x, y) = \Pr(X \leq x, Y \leq y) = \int\limits_{-\infty}^{x} \int\limits_{-\infty}^{y} f(u, v)dudv$$

which implies,

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F(x, y)}{\partial y \partial x}$$

In other words, *derivatives of joint c.d.f's are the joint p.d.f's.*

Properties:

- $\Pr(a < X \leq b, c \leq Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c)$

- *Marginal c.d.f for x* is given by: $F_X(x) = \Pr(X \leq x) = \Pr(X \leq x, Y \in (-\infty, \infty)) = \lim_{y\to\infty} F(x, y)$
- Similarly, *Marginal c.d.f for y* is: $F_Y(y) = \Pr(Y \leq y) = \lim_{x\to\infty} F(x, y)$

4.6 Marginal Density Functions

Joint distributions deal with probability densities over multiple random variables. Extracting the densities over a subset of these random variables is termed as *marginalizing* over the desired subset. This gives us *marginal probability density.* Again, we can separate the discrete and continuous cases:

Discrete marginal p.f:

$$f_X(x) = \sum_Y f(x, y)$$

$$f_Y(y) = \sum_X f(x, y)$$

Continuous marginal p.f:

$$f_X(x) = \int\limits_{-\infty}^{\infty} f(x, y)dy$$

$$f_Y(y) = \int\limits_{-\infty}^{\infty} f(x, y)dx$$

Independence: Two random variables are independent if they produce independent events,

$$\Pr(X \in A, Y \in B) = \Pr(X \in A) \Pr(Y \in B)$$

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x) \Pr(Y \leq y)$$

In other words, we can also define the independence condition using c.d.f's as $\forall x, y$,

$$F(x, y) = F_X(x)F_Y(y)$$

For continuous random variables, we can also define the independence condition based on p.d.f's,

$$f(x, y) = f_X(x)f_Y(y)$$

Thus, as can be seen above there are several ways to verify the independence of random variables (using both c.d.f's as well as p.d.f's).

4.7 Conditional Density Functions

Discrete Conditional density function: Consider discrete random variables X, Y with p.f $f_X(x)$ and $f_Y(y)$ respectively, and joint p.f. $f(x, y)$, then the *conditional p.m.f* is defined as:

$$\Pr(X = x, Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} = \frac{f(x, y)}{f_Y(y)}$$

This is a new probability distribution by itself, infact this new conditional density can be defined in the similar way as we defined p.d.f's earlier:

$$g_X(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{f(y)}, & f(y) > 0 \\ 0, & \text{otherwise} \end{cases}$$

Continuous Conditional density function: Likewise, when X and Y are continuous random variables, we can define the *conditional p.d.f* as:

$$g_X(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)}, & f_Y(y) > 0 \\ 0, & \text{otherwise} \end{cases}$$

It can be shown that $g_X(x|y)$ is a valid p.f (obeys the usual properties of a p.f), both in the discrete and continuous cases.

Baye's Theorem for conditional densities:

$$g(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{g(y|x)f_X(x)}{f_Y(y)}$$

4.8 Multi-variate Distributions

When only two random variables are involved, the distributions are called *bi-variate* distributions. *Multivariate* case is the generalization to n random variables.
TODO

4.9 Functions of a random variable, Transformation Theorems

TODO

5 Markov Chains

TODO

6 Expectations and Variances

6.1 Definition of Expectation and Properties

If X is given to be a discrete random variable with p.m.f $f_X(x)$, then its expected value is defined as:

$$\mathbb{E}[X] = \sum_X x f_X(x) = \sum_X x \Pr(X = x)$$

Remarks:

- Expectation can be thought of as a *weighted average*.
- It could happen that $\mathbb{E}[X]$ does not exist. Therefore, $\sum_X |x| f_X(x) < \infty$ is the condition for expectation of X to be well-defined. For e.g.: in the case of *cauchy* distribution, expectation is not defined.

Extending the definition to a continuous random variable X with p.d.f $f(x)$, the expected value is defined analogously as:

$$\mathbb{E}[X] = \int\limits_{-\infty}^{\infty} x f(x) dx$$

Properties:

- Consider $Y = r(X)$ be a random variable which is a function of another random variable X having the p.d.f as $f(x)$. There are two ways of computing the expected value of Y, $\mathbb{E}[Y]$:
  - $\mathbb{E}[Y] = \int\limits_{-\infty}^{\infty} y\, g(y) dy$, where $g(y)$ is the p.d.f of Y (which has to be computed first)
  - $\mathbb{E}[Y] = \int\limits_{-\infty}^{\infty} r(x) f(x) dx$, where $f(x)$ is the p.d.f of X (which is already available)
- If 'a' is a constant, then $\mathbb{E}[a] = a$.
- If $Y = aX + b$, then $\mathbb{E}[Y] = a\mathbb{E}[X] + b$
- If 'a' is a constant such that $\Pr(X \geq a) = 1$, then $\mathbb{E}[X] \geq a$.
- If 'b' is a constant such that $\Pr(X \leq b) = 1$, then $\mathbb{E}[X] \leq b$.
- *Linearity of Expectations*: If $X_1, \ldots, X_n$ are random variables then,

  $$\mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

  Note that no *independence* assumptions are required for this to hold.

6.2 Jensen's Inequality

Let g be a *convex* function and X a random variable, then:

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

Pf: Uses Taylor's expansion of g(x) around $\mu = \mathbb{E}[X]$.

Convex Functions: A function g is convex if, $\forall \alpha \in [0,1]$ and $\forall x, y$ the following holds:

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y)$$

6.3 Definition of Variance and Properties

If X is a random variable, the variance of X is defined as:

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

where $\mu = \mathbb{E}[X]$.

Therefore, **in the discrete case**:

$$\mathbb{V}[X] = \sum_X (x-\mu)^2 f(x)$$

and **in the continuous case**:

$$\mathbb{V}[X] = \int\limits_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Properties:

- Variance cannot be negative: $\mathbb{V}[X] \geq 0$
- Variance is a number, not a random variable
- Variance tells about "dispersion/distance" while mean/expectation tells us about the "location"[2].
- Standard Deviation is defined as: $\sqrt{\mathbb{V}[X]}$
- If $\mathbb{E}[X]$ does not exist $\implies$ $\mathbb{V}[X]$ does not exist as well. On the other hand, if $\mathbb{E}[X]$ exists $\not\Longrightarrow$ $\mathbb{V}[X]$ exists.
- If 'a' is a constant, then $\mathbb{V}[a] = 0$ (Contrast this with the expectation).
- If 'a' and 'b' are constants, $\mathbb{V}[aX + b] = a^2\ \mathbb{V}[X]$
- *Linearity of Variances*: If $X_1, X_2, \ldots, X_n$ are "independent" random variables, then $\mathbb{V}[X_1 + X_2 + \ldots + X_n] = \mathbb{V}[X_1] + \mathbb{V}[X_2] + \ldots + \mathbb{V}[X_n]$. In other words:

  $$\mathbb{V}\Bigg[\sum_{i=1}^n X_i\Bigg] = \sum_{i=1}^n \mathbb{V}[X_i]$$

  Note that here the *independence* assumptions are absolutely necessary (as opposed to linearity for expectations) and when the independence doesn't hold, *covariances* also have to be taken into account.

[2] This is also why in statistics mean and variance are referred to as location and scale parameters

- If X, Y are two random variables which are *not necessarily independent*, then:

  $$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\, cov(X,Y)$$

  In case of three random variables X_1, X_2 and X_3:

  $$\mathbb{V}[X_1+X_2+X_3] = \mathbb{V}[X_1] + \mathbb{V}[X_2] + \mathbb{V}[X_3] + 2\, cov(X_1,X_2) + 2\, cov(X_2,X_3) + 2\, cov(X_1,X_3)$$

  and in general for $X_1, \ldots, X_n$:

  $$\mathbb{V}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{V}[X_i] + 2\, \sum_{i<j}\sum_{j} cov(X_i, Y_j)$$

Why are these linearity relations useful?
Often computing the expectations and variances of some quantities get complicated easily and expressing them as a sum reveals possible alternatives. Consider for example a random variable X distributed according to the Binomial distribution and let us say we are interested in computing the variance of X. Instead of working out the variance of binomial distribution the formal way, we could use the fact that a Binomial random variable X is a collection of i.i.d Bernoulli random variables X_i. This observation lets us use the linearity of variances.

$$X \sim \text{Binomial}(n,p)$$

$$\mathbb{V}[X] = ?$$

We can express a binomial r.v. as a collection of n i.i.d bernoulli r.v. s. Therefore, using the linearity of variances,

$$\mathbb{V}[X] = \mathbb{V}[X_1] + \mathbb{V}[X_2] + \ldots + \mathbb{V}[X_n]$$

$$X_i \sim \text{Bernoulli(p)}$$

Finally,

$$\mathbb{V}[X] = n\ \mathbb{V}[X_i] = np(1-p)$$

6.4 Conditional Expectation

Consider random variables X and Y, $g_Y(Y|X)$ being the conditional distribution of Y given X. Then the *conditional expectation* is defined as follows:
Discrete Case:

$$\mathbb{E}[Y|X=x] = \sum_{Y} y\ g_Y(y|x) = \sum_{Y} y\, \Pr(Y=y|X=x)$$

Continuous Case:

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y\ g_Y(y|x)dy$$

Properties:

- $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$

6.5 Conditional Variance

Using the same setup as for conditional expectations, we can define *conditional variance* as follows:

$$\mathbb{V}[Y|X=x] = \mathbb{E}\big[Y^2|X=x\big] - \Big(\mathbb{E}\big[Y|X=x\big]\Big)^2$$

6.6 Covariance, Correlation, Schwarz Inequality

Covariance and *Correlation* both measure linear association between random variables.

Covariance: Consider random variables X, Y, then the *covariance* between them is defined as:

$$cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

Substituting the above, we arrive at another expression for the covariance:

$$cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Correlation: Correlation between two random variables X and Y is given by:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$$

Correlation ρ takes values between -1 and 1. Values closer to 1 or -1 indicate "high linear relationship" between the variables while values closer to 0 indicate "low linear relationship". Note that correlation does not say anything about *non-linear association* though.

Schwarz Inequality: Given two random variables U and V:

$$\Big(\mathbb{E}[U\ V]\Big)^2 \leq \mathbb{E}[U^2]\ \mathbb{E}[V^2]$$

Other useful properties/theorems:

- If X, Y are independent random variables with finite variances, then: $cov(X,Y) = 0$ and $\rho(X,Y) = 0$
- In the case of a single random variable, covariance reduces to variance: $cov(X,X) = \mathbb{E}[X^2] - \mu_X\mu_Y = \mathbb{V}[X]$
- If X, Y are random variables with finite variances σ_X^2 and σ_Y^2, then:

$$cov(X,Y)^2 \leq \sigma_X^2\sigma_Y^2$$

$$-1 \leq \rho(X,Y) \leq 1$$

- If X is a random variable with finite variance, and Y is another random variable defined as $Y = aX + b$ with constants 'a', 'b' ($a \neq 0$), then:

$$a > 0 \implies \rho(X,Y) = 1$$

$$a < 0 \implies \rho(X,Y) = -1$$

As a matter of fact, the above property is the *main reason why linear regression works in practice.*

6.7 Moment Generating Function

Moment of random variable: $\mathbb{E}[X^k]$ is termed as the k-th moment of a random variable X.

Remarks:

- k-th moment exists iff $\mathbb{E}[|X|^k] < \infty$
- If X is bounded (i.e.: $\Pr(a \leq X \leq b) = 1$ for some constants a, b), then: *all the moments of X exist.* Note that this condition is not satisfied in the case of normal distribution, as a result of which not all its moments exist.
- If the higher order moments exist, then the lower order moments must also exist. i.e.:
  If $\mathbb{E}[|X|^k] < \infty$ for some k, then:
  $\mathbb{E}[|X|^j] < \infty$ for every $j > 0$ such that $j < k$

Central Moment of random variable: $\mathbb{E}[(X-\mu_X)^k]$ is termed as the k-th central moment of a random variable X.

Why are we interested in these moments?
Each of these moments gives us some information about the random variable and the distribution underlying it. For instance, the first order moment gives us the expected value, second order central moment gives us variance and the third order central moment provides a measure of the "skewness" of the distribution. Similarly, questions about "symmetry" and "spread" of the distribution can be answered by looking at the moments. This is also the reason we study the *moment generating function* which helps generate any number of moments for a distribution. However, it is important to keep in mind that certain distributions like the *cauchy distribution* do not have any moments, but we still work with them.

Moment Generating Function (m.g.f): Given a random variable X, the m.g.f of X is given by:

$$\psi(t) = \mathbb{E}[e^{tX}]$$

where $t \in \mathbb{R}$.

Using m.g.f to generate moments: If X is a random variable whose m.g.f $\psi(t)$ is finite for all t in some open interval around $t = 0$, then $\forall n > 0$, the n-th moment of X is derived as follows:

$$\mathbb{E}[X^n] = \left.\frac{\partial^n \psi(t)}{\partial t^n}\right|_{t=0} = \psi^{(n)}(0)$$

In simple words, the n-th derivative of the m.g.f of X evaluated at the point $t = 0$ gives the n-th moment for X.

Properties:

- Let X be a random variable with m.g.f ψ_1. Also, let Y be another random variable such that $Y = aX + b$ where 'a', 'b' are constants. If ψ_2 is the m.g.f of Y, then:

  $$\psi_2(t) = e^{bt}\psi_1(at)$$

  for all t such that $\psi_1(at)$ is finite:

- If $X_1, X_2, \ldots, X_n$ are n independent random variables (need not be identically distributed) and ψ_i is the m.g.f of X_i. Also, let Y be another random variable such that $Y = X_1 + X_2 + \ldots + X_n$. Then:

  $$\psi(t) = \prod_{i=1}^{n} \psi_i(t)$$

  for all t such that $\psi_i(t)$ is finite. Here, $\psi(t)$ is the m.g.f of Y.
  This theorem is very useful in practice especially in the context of *multivariate gaussian distributions*.

6.8 The Normal Distribution and useful results

A random variable X has a normal distribution (gaussian distribution) with mean μ and variance σ^2 if X has the p.d.f defined by:

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

where $-\infty < x < \infty$.

Why is normal distribution so attractive? Normal (or Gaussian) distribution is one of the most commonly used distribution in mathematics and statistics. It has several nice mathematical properties which make it highly applicable to several areas. It is *unimodal*, *symmetric*, *closed under addition operation* to mention just a few. It is also closely related to the *Central Limit Theorem.* A lot of statistics (inference, hypothesis testing, etc) is based on gaussian distribution.

Properties:

- The normal distribution defined by $f(x|\mu, \sigma^2)$ is a valid p.d.f
- The m.g.f of X such that $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by:

  $$\psi(t) = \exp\left\{ \mu t + \frac{1}{2}\sigma^2 t^2 \right\}$$

- If X is a random variable such that $X \sim \mathcal{N}(\mu, \sigma^2)$, then: $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$. Note that proof for this follows from using the m.g.f for normal distribution and taking appropriate derivatives.
- **Standard Normal Distribution**: When $\mu = 0$, $\sigma^2 = 1$, then we have the *standard normal distribution* where $X \sim \mathcal{N}(0, 1)$. The p.d.f is given by:

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

  where $-\infty < x < \infty$.
  Add diagrams showing symmetry, cdf etc.

  We use $\phi(x)$ to denote the p.d.f of a standard normal distribution and $\Phi(x)$ to denote the c.d.f of a standard normal distribution respectively.

  $$\Phi(x) = \Pr(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{t^2}{2} dt \right\}$$

This integration is not tractable as there is an infinite lower-bound (c.d.f of a standard normal cannot be computed in closed form) and hence *numerical integration methods* are needed. Therefore common ways to compute this c.d.f are:

– Use R or some other scientific programming language to do the numerical integration (upto certain approximation)

– Use the z-tables (this is the most widely used method)

– Use calculator

What do we do if we do not have the distribution in standard normal form? In that case, the normal distribution can be standardized by subtracting the mean and dividing by the standard deviation. After that the z-tables can be used.

- **Linear Transformations**: If X, Y are random variables such that $X \sim \mathcal{N}(\mu, \sigma^2)$, and $Y = aX + b$, where $a, b \in \mathbb{R}$, $a \neq 0$, then:
$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$
- **Standardizing the normal distribution**: If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$, then:
$$Z \sim \mathcal{N}(0, 1)$$
and the c.d.f of X is given by:
$$F(x) = \Phi(\frac{x-\mu}{\sigma}) = \Phi(Z)$$
- **Linear Combinations**: If $X_1, X_2, \ldots, X_n$ are independent random variables such that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then:
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\Big)$$
More generally,
$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\Big(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\Big)$$
- **Sample Mean**: Using the above result, we can compute the *sample mean* of n samples $X_1, \ldots, X_n$ drawn i.i.d from $\mathcal{N}(\mu, \sigma^2)$ as:
$$\mathbb{E}\big[\overline{X}_n\big] = \frac{X_1 + \ldots + X_n}{n} = \mu$$
and the *sample variance* as:
$$\mathbb{V}\big[\overline{X}_n\big] = \frac{\sigma^2}{n}$$
Therefore, the resulting sample is distributed according to the following normal distribution:
$$\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$
Intuitively what this implies is that as we collect more samples (i.e.: n increases), the variance in our estimate decreases.

6.9 Log-Normal Distribution

6.10 Markov Inequality

6.11 Chebyshev's Inequality

6.12 Law of Large Numbers

6.13 Central Limit Theorem and Interpretations

7 Stochastic Process

7.1 Definition of Stochastic Process, Properties and Examples

7.2 Poisson Process

8 Probability Distributions

8.1 Bernoulli

8.2 Binomial

8.3 Multinomial

8.4 Beta

8.5 Gamma

8.6 Dirichlet

8.7 Poisson

8.8 Gaussian (Normal)

8.9 Log-Normal

8.10 Exponential

8.11 Negative Binomial

8.12 Geometric

8.13 Hyper-Geometric

8.14 Exponential Family notation

9 Appendix: Useful tools and identities

9.1 Product and Quotient rule for derivatives

9.2 Integration by parts

9.3 Integrals of common functions

9.4 Sum of natural numbers series

9.5 Arithmetic, Geometric Series and sums

9.6 Power Series

9.7 Taylors Series

9.8 Completing the squares

10 Acknowledgements