Chapter 3: Random Variables

Chiranjit Mukhopadhyay
Indian Institute of Science

3.1 Introduction

In the previous Chapter on Elementary Probability Theory, we learned how to calculate probabilities of non-trivial events *i.e.* event $A \neq \phi$ or Ω. While they are useful in elementary probability calculations, for practical applications and in particular for development of statistical theory, we are typically interested in modeling the distribution of values of a "variable" in a real or hypothetical population. In this chapter we shall learn to do so, where we shall further learn how to define concepts like mean, standard deviation, median etc. of a variable of interest in a population. But before getting into the details we need to first formally define what we mean by a "variable", leading to our first definition.

Definition 3.1: A **random variable** (r.v.) X is a function which maps the elements of the sample space Ω to the set of real numbers $\Re$. [1]

Mathematically, $X : \Omega \to \Re$, and the range of the r.v. X is denoted by $\mathcal{X}$. X is a variable because its value $X(\omega)$ depends on the input ω, which varies over the sample space Ω; and this value is random because the input ω is random, which is the outcome of a chance experiment. A few examples will help clarify the point.

Example 3.1: Consider the experiment of tossing a coin three times. For this experiment the sample space $\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$. Let $X(\omega)$ =No. of H's in ω, which in words represents the number of Heads in the three tosses. Thus $X(HHH) = 3$, ..., $X(THT) = 1$, ..., $X(TTT) = 0$, and $\mathcal{X} = \{0, 1, 2, 3\}$. $\bigtriangledown$

Example 3.2: Consider the experiment of rolling a dice twice. For this experiment the sample space $\Omega = \{(1,1), \ldots, (1,6), \ldots, \ldots, (6,1), (6,6)\} = \{$ordered pairs $(i,j) : 1 \leq i \leq 6, 1 \leq j \leq 6, i$ and j integers$\}$. Let $X(\omega) = X((i,j)) = i + j$, which in words represents the sum of two faces. Thus $X((1,1)) = 2$, ..., $X((3,4)) = 7$, ..., $X((6,6)) = 12$, and $\mathcal{X} = \{2, 3, \ldots, 11, 12\}$. $\bigtriangledown$

Example 3.3: Consider the experiment of throwing a dirt into a dartboard with radius r. If

[1] It should be noted that any such function $X : \Omega \to \Re$ does not qualify to be called a random variable. Recall that typically the sample space Ω is considered along with a collection of events $\mathcal{A}$, a σ-field of subsets of Ω. Now consider the σ-field generated by all finite unions of intervals of the form $\cup_{i=1}^{n}(a_i, b_i]$, where $-\infty < a_1 < b_1 < a_2 < b_2 \cdots < a_n < b_n < \infty$, in $\Re$. This σ-field is called the Borel σ-field in $\Re$ and is denoted by $\mathcal{B}$. Now consider a function $X : \Omega \to \Re$. Such a function is called a random variable if $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$, $\forall B \in \mathcal{B}$. The reason for this is otherwise we may not be able to define $P(X \in B)$ $\forall B \in \mathcal{B}$, because as has been mentioned in the previous chapter, $P(A)$ remains undefined for $A \subseteq \Omega \notin \mathcal{A}$. But as in the previous chapter where we had pretended as if such pathologies do not exist and proceeded with $\mathcal{A} = \wp(\Omega)$, the power set of Ω, here also we shall do the same with $\mathcal{A} = \wp(\Omega)$ and $\mathcal{B} = \wp(\Re)$ and X as any function from Ω to $\Re$, without getting bogged down with rigorous mathematical treatment of the subject.

the bull's eye or the center of the dartboard is taken as the origin $(0,0)$ then assuming that the dirt always lands somewhere on the dartboard, the sample space $\Omega = \{(x,y) : x^2+y^2 \leq r^2\}$. Let $X(\omega) = X((x,y)) = \sqrt{x^2+y^2}$, which in words represents the distance of the landed dirt from the bull's eye. Then for this X, $\mathcal{X} = [0,r]$. $\bigtriangledown$

Once a r.v. X is defined and its range $\mathcal{X}$, the set of possible values that X can take, is identified, the immediate next question that arises is that of its probability distribution. By that it is meant that we would next like to know with what probability or what kind of frequency is X taking a value $x \in \mathcal{X}$. Once we are able to answer that, since by definition X is real-valued, the next natural questions are then, what is the average or mean value taken by X, what is the variability or more precisely the standard deviation of the values taken by X, what is the median of values taken by X, what can we say about the value $x_{0.9}$ (say) such that 90% of the time X will be $\leq x_{0.9}$, etc.. Once these things are appropriately defined and computed for a r.v. X, they will then give us the notion and numerical values of these concepts for the population of possible values of a r.v. X.

In statistical applications we have a (real or hypothetical) population of values of some variable X, like say for example height, weight, age, income etc. of interest, which we would like to study. For this purpose we shall typically collect a sample and observe these variables of interest for the individuals (also called sampling units) chosen in the sample, based on which we would like to extrapolate or infer about different features of the population of X values. But for doing that, say for example for saying something like, in the population, median age is 25 years, or standard deviation of heights is 4", or mean income is Rs.100,000; we first need to concretely define these concepts themselves in the population before learning how to use the sample to infer about them, in which we are eventually interested in. Here we shall learn how to define these notions in the population by studying the totality of values a variable X can take without referring to a sample of such values. It turns out that the way one can define these notions for a r.v. X, starting with its probability distribution, that gives which value occurs how frequently, depends on the nature of $\mathcal{X}$ and there are at least two different cases that needs separate treatment[2] - discrete and continuous, which are taken up in the next two sections.

3.2 Discrete R.V.

Definition 3.2: A random variable X is called **discrete** if its range $\mathcal{X}$ is countable.

A set is called countable if it is either finite or countably infinite. A set $\mathcal{X}$ is called countably infinite if there is a one-to-one and onto function $f : \mathcal{P} \to \mathcal{X}$, where $\mathcal{P}$ is the set of positive integers $\{1,2,\ldots\}$. Like for example, $\mathcal{X} = \{0,1,2,\ldots\}$, the set of non-negative integers, is countably infinite as $f(n) = n-1$ is a one-to-one and onto function $f : \mathcal{P} \to \mathcal{X}$; $\mathcal{X} = \{2,4,\ldots\}$, the set of positive even integers, is countably infinite as $f(n) = 2n$ is a one-

[2]Of course there are ways to mathematically handle all kinds of random variables in a unified manner, which we shall learn in due course. But treating two separate cases are conceptually much easier for the beginners, and thus like most standard text books here also the same approach is adopted.

to-one and onto function $f : \mathcal{P} \to \mathcal{X}$; $\mathcal{X} = \{0, \pm 1, \pm 2, \ldots\}$ the set of integers is countably infinite as $f(n) = \begin{cases} n/2 & \text{if } n \text{ is even} \\ -(n-1)/2 & \text{if } n \text{ is odd} \end{cases}$ is a one-to-one and onto function $f : \mathcal{P} \to \mathcal{X}$; the set of rational numbers $\mathcal{Q}$ is countably infinite as $\mathcal{Q} = \cup_{n=0}^{\infty} A_n$, where $A_0 = \{0\}$ and for $n \geq 1$, $A_n = \{\pm m/n : m \in \mathcal{P}$ and m and n are relatively prime$\}$ are countable sets and countable union of countable sets is countable.

Thus for a discrete r.v. X, its range may be written as $\mathcal{X} = \{x_1, x_2, \ldots\}$. For such a r.v. the most obvious way to define its distribution would be to explicitly specify $P[X = x_n] = p_n$ (say) for $n \geq 1$, where for computing p_n one needs to go back to Ω to see which $\omega \in \Omega$ satisfies the condition $X(\omega) = x_n$, collect all such ω's into a set $A \subseteq \Omega$ and then define $p_n = P(A)$. In this process it is implicit that we are defining the event $[X = x_n]$ as $\{\omega \in \Omega : X(\omega) = x_n\}$. When the probability distribution of a (discrete) r.v. is defined by specifying $P[X = x]$ for $x \in \mathcal{X}$ then it is being specified through its probability mass function. In general the definition of a probability mass function is as follows.

Definition 3.3: A function $p : \mathcal{X} \to [0, 1]$ with a countable domain $\mathcal{X} = \{x_1, x_2, \ldots\}$ is called a **probability mass function** or **p.m.f.** if
a. $p_n \geq 0 \; \forall n \geq 1$, and
b. $\sum_{n \geq 1} p_n = 1$,
where $p_n = p(x_n)$.

Specifying the distribution of a (discrete) r.v. X by its p.m.f. means providing a function $p(x)$, which is a p.m.f. with $\mathcal{X}$ as its domain as in ***Definition 3***, with the interpretation that $p(x) = P[X = x]$. With this interpretation, $p(x)$ can be defined $\forall x \in \Re$ (not necessarily only for $x \in \mathcal{X}$) as $p(x) = \begin{cases} p_n & \text{if } x = x_n \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases}$. Now let us look at a couple of examples to examine how the distribution of a (discrete) r.v. may be specified by its p.m.f..

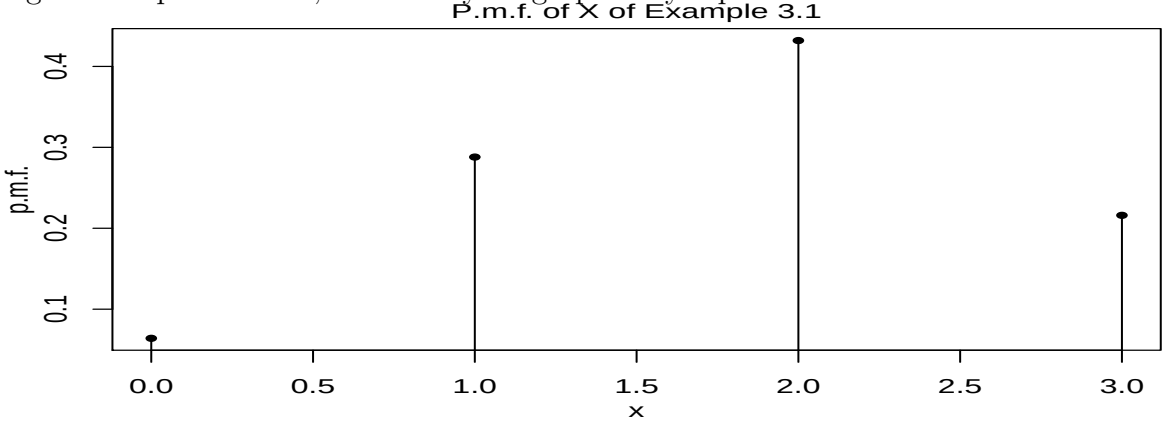
Example 3.1 (Continued): Here the r.v. X can only take values in $\mathcal{X} = \{0, 1, 2, 3\}$ and thus in order to obtain its p.m.f. we only need to figure out $P[X = x]$ for $x = 0, 1, 2, 3$. However for this purpose we first need to know the probabilities of each sample point $\omega \in \Omega$. Suppose the coin is biased with $P(H) = 0.6$, so that $P(T) = 0.4$, and the three tosses are independent. Then the probabilities of the 8 ω's are as follows:

ω	HHH	HHT	HTH	THH
Probability	$0.6^3 = 0.216$	$0.6^2 \times 0.4 = 0.144$	$0.6^2 \times 0.4 = 0.144$	$0.6^2 \times 0.4 = 0.144$
ω	TTH	THT	HTT	TTT
Probability	$0.6 \times 0.4^2 = 0.096$	$0.6 \times 0.4^2 = 0.096$	$0.6 \times 0.4^2 = 0.096$	$0.4^3 = 0.064$

Now after collecting the ω's corresponding to the four events $[X = x]$ for $x = 0, 1, 2, 3$ we get

$P[X = 0]$	$P[X = 1]$	$P[X = 2]$	$P[X = 3]$
$= P(\{TTT\})$	$= P(\{HTT, THT, TTH\})$	$= P(\{HHT, HTH, THH\}$	$= P(\{HHH\}$
$= 0.064$	$= 3 \times 0.096$	$= 3 \times 0.144$	$= 0.216$
	$= 0.288$	$= 0.432$	

This gives the p.m.f. of X, which may be graphically represented as follows: $\bigtriangledown$



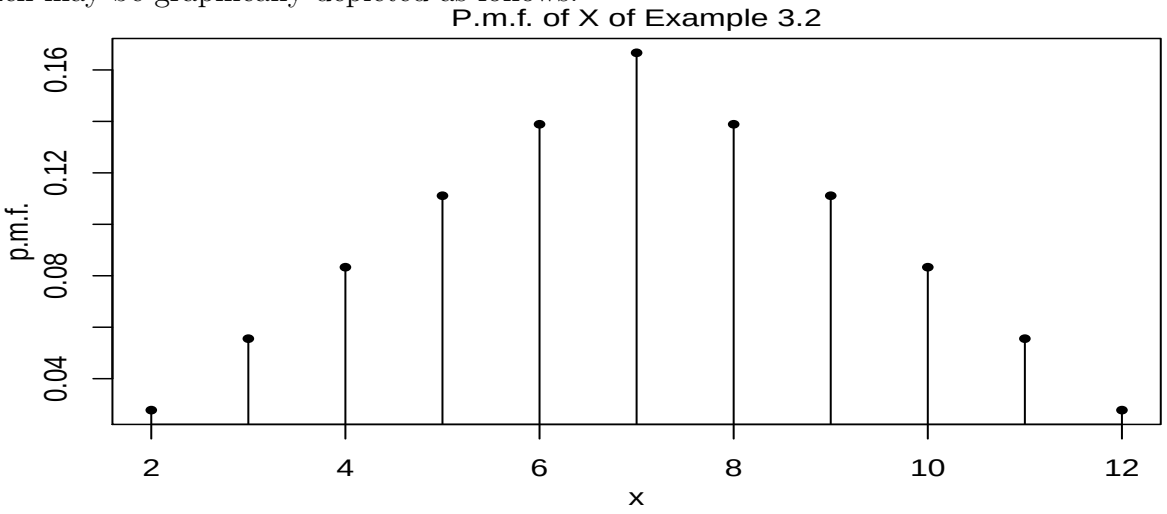
Example 3.2 (Continued): Here $\mathcal{X} = \{2, 3, \ldots, 11, 12\}$ and the probability of each of the 11 events $[X = x]$ for $x = 2, 3, \ldots, 11, 12$ are found by looking at the value X takes for each of the 36 fundamental outcomes as in the following table:

$X = i + j$ $j \downarrow \quad i \rightarrow$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Now if the dice is fair, then each of the 36 fundamental outcomes is equally likely with a probability of 1/36 each, so that by collecting or counting the number of these fundamental outcomes that lead to the event $[X = x]$ for $x = 2, 3, \ldots, 11, 12$ we obtain the p.m.f. of X as

x	2	3	4	5	6	7	8	9	10	11	12
$p(x) \times 36$	1	2	3	4	5	6	5	4	3	2	1

which may be graphically depicted as follows: $\bigtriangledown$



Example 3.4: Consider the experiment of keeping on tossing a coin till a Head appears. For this experiment, $\Omega = \{H, TH, TTH, TTTH, \ldots\}$. Define the random variable $X(\omega)$ =No. of T's in ω, which in words gives, the number of Tails till the first Head appears, or $X+1$ gives the number of tosses required to get the first Head in an experiment where a coin is tossed till a Head appears. Clearly $\mathcal{X} = \{0, 1, 2, \ldots\}$ which is not finite but countably infinite. Thus this r.v. is discrete. Now in order to obtain the p.m.f. of X, suppose the tosses are independent and let $P(H) = p$ for some $0 < p < 1$ so that $P(T) = 1-p = q$ (say). Then for $x = 0, 1, 2, \ldots$, $p(x) = P[X = x] = P[\underbrace{TT\cdots T}_{x-\text{many}} H] = q^x p$ gives the p.m.f. of X. Note that $p(x)$ is a legitimate p.m.f. because $q^x p > 0\ \forall x = 0, 1, 2, \ldots$ and $\sum_{x=0}^{\infty} q^x p = p[1 + q + q^2 + q^3 + \cdots] = \frac{p}{1-q} = 1$. $\bigtriangledown$

One of the main reasons for obtaining the probability distribution of a r.v. is to be able to compute $P[X \in A]$ for an arbitrary $A \subseteq \mathcal{X}$. While this is conceptually straight-forward to do so using the p.m.f. $p(x)$ of a (discrete) r.v. with the help of the formula $P[X \in A] = \sum_{x \in A} p(x)$, the validity of which easily follows from countable additivity of $P(\cdot)$, in practice, evaluating the summation may be a tedious task. For example, in **Example 1**, the probability of the event, "at most one head", may be expressed as $X \leq 1$, the probability of which is obtained as $P[X \leq 1] = P[X = 0] + P[X = 1] = 0.064 + 0.288 = 0.352$; in **Example 2**, the probability of the event, "sum not exceeding 9 and not less than 3", may be expressed as $3 \leq X \leq 9$, the probability of which is obtained as $P[3 \leq X \leq 9] = P[X = 3] + P[X = 4] + \cdots + P[X = 9] = (2 + 3 + 4 + 5 + 6 + 5 + 4)/36$ =29/36; and in **Example 4**, the probability of the event, "at least 10 tosses are required to get the first Head", may be expressed as $X \geq 9$, the probability of which is obtained as $P[X \geq 9] = 1 - P[X \leq 8] = 1 - p[1 + q + q^2 + \cdots + q^8] = 1 - p\frac{1-q^9}{1-q} = q^9$. A tool which facilitates such probability computation is called cumulative distribution function, which is defined as follows.

Definition 3.4: For a r.v. X its **cumulative distribution function** or **c.d.f.** is given by $F(x) = P[X \leq x]$ for $-\infty < x < \infty$.

First note that (unlike p.m.f.) the definition does not require X to be discrete. The notion of c.d.f. is well-defined for any r.v. X. [3] Next note that for a discrete r.v., computation of its c.d.f. amounts to calculation of all the partial sums in one go which are set aside in its c.d.f., which can then be invoked for easy probability calculations. Finally note that for a discrete r.v. X, its c.d.f. is an alternative to p.m.f. way of specifying its the probability distribution. Both convey the same information about the probability distribution but each one has its own use in exploring different features of the distribution. As the notion of c.d.f. is common across the board for all r.v., a general discussion on c.d.f. of an arbitrary random variable is provided in Appendix A, which the reader should read after learning the concepts associated with a continuous random variable in §3. We begin by working with a few examples involving the notion of c.d.f. of discrete random variables.

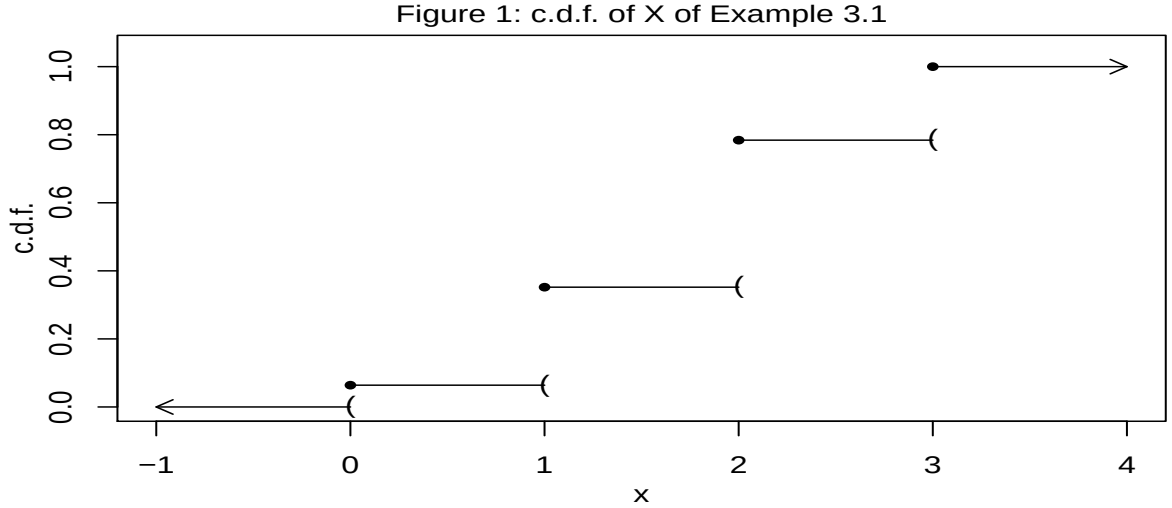
Example 3.1 (Continued): With the p.m.f. of X already figured out let us now compute its c.d.f. $F(x)$. For this we need to look at all possible ranges of values of X. First consider $-\infty < x < 0$. Since $X \geq 0$, clearly for $-\infty < x < 0$, $F(x) = P[X \leq x] =$

[3] As mentioned in footnote [2], c.d.f. is the vehicle through which all r.v.'s can be studied in a unified manner.

0. Next for $0 \leq x < 1$, $F(x) = P[X \leq x] = P[X = 0] = 0.064$; for $1 \leq x < 2$, $F(x) = P[X \leq x] = P[X = 0] + P[X = 1] = 0.064 + 0.288 = 0.352$; for $2 \leq x < 3$, $F(x) = P[X \leq x] = P[X \leq 1] + P[X = 2] = 0.352 + 0.432 = 0.784$; and finally for $3 \leq x < \infty$, $F(x) = P[X \leq x] = P[X \leq 2] + P[X = 3] = 0.784 + 0.216 = 1$. In summary $F(x)$ can be written as follows:

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 0.064 & \text{if } 0 \leq x < 1 \\ 0.352 & \text{if } 1 \leq x < 2 \\ 0.784 & \text{if } 2 \leq x < 3 \\ 1 & \text{if } 3 \leq x < \infty \end{cases}$$

whose graph when plotted against x looks as follows: $\triangledown$



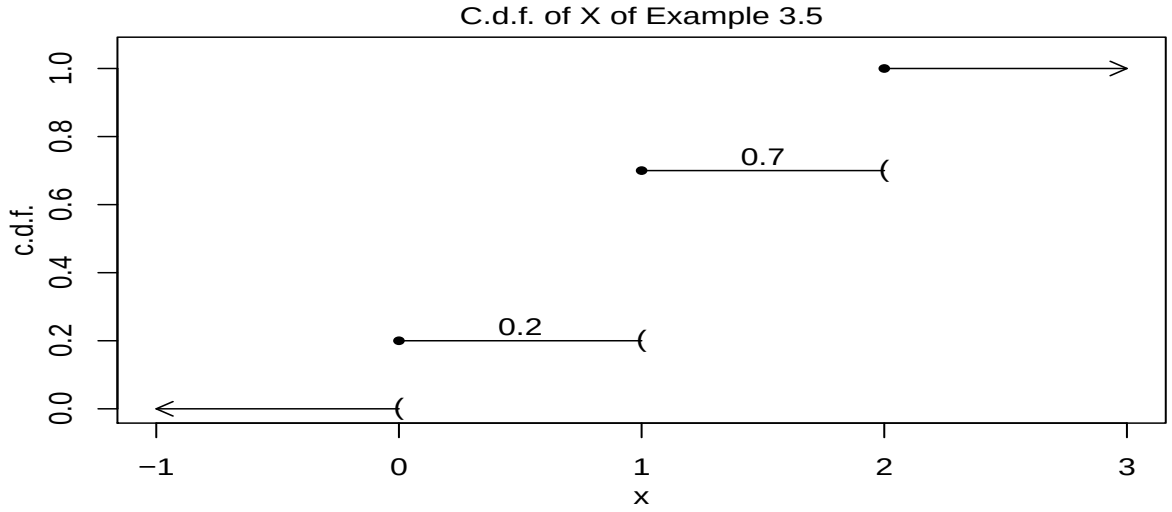
A couple of general remarks regarding the nature of the c.d.f. of a discrete r.v. are in order, which Figure 1 above will facilitate understand. In general the c.d.f. of a discrete r.v. looks as it is plotted in Figure 1 for **Example 1**. It is a discontinuous step function, with jumps at the points where it has a positive probability mass with the quantum of jump same as this probability mass and flat in between. As mentioned before, the general properties of an arbitrary c.d.f. have been systematically assorted in Appendix A, however it helps one intuitively understand two of the properties of the c.d.f. by studying it in this discrete case. The first one is that if there is a positive probability mass at a given point then the c.d.f. gets a jump with the amount of jump same as the probability mass and is discontinuous at that point, and vice-versa. The second point is that the r.v. has probability 0 of taking values in an open interval where the c.d.f. is flat, and vice-versa. Now let's see one of the major uses of the c.d.f. namely in probability computation.

Example 3.2 (Continued): Proceeding as in the previous example the c.d.f. of X in this example is given in the following table:

x	$-\infty < x < 2$	$2 \leq x < 3$	$3 \leq x < 4$	$4 \leq x < 5$	$5 \leq x < 6$	$6 \leq x < 7$
$F(x)$	0/36	1/36	3/36	6/36	10/36	15/36
x	$7 \leq x < 8$	$8 \leq x < 9$	$9 \leq x < 10$	$10 \leq x < 11$	$11 \leq x < 12$	$12 \leq x < \infty$
$F(x)$	21/36	26/36	30/36	33/36	35/36	36/36

Now for computing the probability of the event of interest $[3 \leq X \leq 9]$ one need not add 7 terms as before, where we had calculated this probability directly using the p.m.f.. $P[3 \leq X \leq 9] = P[X \leq 9] - P[X < 3]$ (this is because, if $A \subseteq B$, $B = A \cup (B - A)$ and $A \cap (A - B) = \phi$, and thus $P(B) = P(A) + P(B - A) \Rightarrow P(B - A) = P(B) - P(A)$; here take $A = [X < 3]$ and $B = [X \leq 9]$) $= P[X \leq 9] - P[X \leq 2] = F(9) - F(2) = (30 - 1)/36 = 29/36$. $\bigtriangledown$

Example 3.5: The c.d.f. of X denoting the number of cars sold by a sales-person on a given day is as follows:



The probability that the sales-person will be able to sell at lest one car on any given day is given by $P[X > 0] = 1 - P[X \leq 0] = 1 - F(0) = 1 - 0.2 = 0.8$. In general, since the probability mass at a given point equals the amount of jump in the c.d.f. at that point, it is easy to see that the X in this example has p.m.f.

x	0	1	2
$p(x)$	0.2	0.5	0.3

. $\bigtriangledown$

As can be seen from the above examples, the probability distribution of a discrete r.v. may be specified by either the p.m.f. or the c.d.f. and one can construct one of these given the other and thus both of them convey the same information. However for probability calculations it is typically easier to use the c.d.f. than the p.m.f.. On the other hand the graph of the p.m.f. typically gives a better intuitive feel about the probability distribution, like where most of the mass is concentrated, symmetry, number of modes etc., than the c.d.f.. Thus for a given distribution it is better to have both of them handy and use the one which is appropriate for a given task.

After having an understanding of a discrete probability distribution, we next turn our attention towards summary measures of such distributions. This scheme is analogous to the chapter on Descriptive Statistics, where after discussing frequency distributions, histograms and ogives, one next turns one's attention towards defining descriptive summary measures like mean, median. mode, standard deviation, skewness, kurtosis etc. that can be computed from the data. Here also we shall do exactly the same. However the key difference here is that we are defining these quantities for a population of values characterized by a r.v. as opposed to similar notions developed for a sample of observed values in the chapter on Descriptive Statistics. There are two classes of summary measures of a probability distribution that we

are interested in - moments and quantiles. Just like the distribution of a discrete r.v. may be characterized using either its p.m.f. or c.d.f., the summary measures attempting to capture general things like central tendency, dispersion or skewness can also be expressed using some functions of either the moments or the quantiles, and interestingly typically one requires the p.m.f. for the computation of the moments and c.d.f. for the quantiles.

3.2.1 Moments

Definition 3.5: For a positive integer k, the **k-th raw moment** of a discrete r.v. X with p.m.f. $p(x)$ is given by $\sum_{x\in\mathcal{X}} x^k p(x)$ which is denoted by $E[X^k]$, and the **k-th central moment** is given by $\sum_{x\in\mathcal{X}}(x-\mu)^k p(x)$ which is denoted by $E[(X-\mu)^k]$, where $\mu = E[X] = \sum_{x\in\mathcal{X}} xp(x)$, the first raw moment, is called the **Expectation** or **Mean** of X.

The intuitive understanding behind the above definition starts with the definition of Expectation or Mean μ. For this consider a r.v. X with the p.m.f.

x	1	2	3	4
$p(x)$	0.2	0.3	0.4	0.1

. According to ***Definition 3.5*** its mean is given by $\mu = 1\times0.2+2\times0.3+3\times0.4+4\times0.1 = 2.4$. Now let's see the rationale behind calling this quantity the "mean", when we already have a general understanding of the notion of mean. For this first recall that random variables are used for modeling a population of values, or the distribution of values in a population is expressed in terms of the probability distribution of an underlying random variable. Thus in this example we are attempting to depict a population where the only possible values are 1, 2, 3 and 4 with their respective relative frequencies being 20%, 30%, 40% and 10%, and μ is nothing but the mean of this population. If this population is finite having N elements, then it has $0.2N$ 1's, $0.3N$ 2's, $0.4N$ 3's and $0.1N$ 4's and thus naturally the mean value in this population should equal $(1\times0.2N + 2\times0.3N + 3\times0.4N + 4\times0.1N)/N = 2.4$. From this calculation it is immediate that the mean does not depend on the population size N and only depends on the relative frequency $p(x)$ of the value x in the population, and thus simply extending the "usual" definition of mean yields the formula $\mu = \sum_{x\in\mathcal{X}} xp(x)$. Before proceeding further it is illustrative to note a few properties of the expectation $E[X]$ of a r.v. X, which are as follows[4] (here c denotes a constant):

Property 1: $E[c] = c$
Property 2: $E[c + X] = c + E[X]$
Property 3: $E[cX] = cE[X]$

Once we accept that $E[X] = \sum_{x\in\mathcal{X}} xp(x)$, it is then natural to define $E[X^k]$ as $\sum_{x\in\mathcal{X}} x^k p(x)$ and $E[(X-\mu)^k]$ as $\sum_{x\in\mathcal{X}}(x-\mu)^k p(x)$. However a little bit of caution must be exercised before taking these formulæ to be granted. This is because X^k or $(X-\mu)^k$ are random variables in their own right and we have already defined the mean of a discrete r.v. as the sum of the product of the possible values it can take and their respective probabilities. Thus if $Y = X^k$ or $Y = (X-\mu)^k$, in order to find their mean one must figure out $\mathcal{Y}$, the set of

[4] A more comprehensive list of these properties along with other moments, not just the expectation, has been assorted in Appendix B for quick reference. The reason for this is, such a comprehensive list of properties of even just the expectation requires concepts that are yet to be introduced.

possible values Y can take, and its p.m.f. $p(y)$. Then its mean will be given by $\sum_{y\in\mathcal{Y}} yp(y)$. But it turns out that this coincides with $\sum_{x\in\mathcal{X}} x^k p(x)$ and $\sum_{x\in\mathcal{X}}(x-\mu)^k p(x)$ in the respective cases justifying the definitions of $E[X^k]$ and $E[(X-\mu)^k]$ as given in ***Definition 5***. Since it is so natural to define $E[g(X)]$ by $\sum_{x\in\mathcal{X}} g(x)p(x)$ for any function $g(\cdot)$, but it requires a proof starting with the definition of $E[X] = \sum_{x\in\mathcal{X}} xp(x)$, $E[g(X)] = \sum_{x\in\mathcal{X}} g(x)p(x)$ is also called the **Law of Unconscious Statistician**.

In practice, other than the first raw moment $E[X]$ or mean μ, the other moment that is extensively used is the second central moment[5] $E[(X-\mu)^2]$. This quantity is called **Variance** of the r.v. X and is denoted by $V[X]$ or σ^2. $\sqrt{V[X]}$ or σ is called the **Standard Deviation** of X. The intuitive idea behind calling it the variance, which is a measure of dispersion or measures how spread apart the values of the r.v. X are, is as follows. In order to measure dispersion one first needs a measure of location of the values as a reference point. Mean μ serves this purpose. Next one measures how far apart the values of X are from this reference point by considering the deviation $(X-\mu)$. Some of these are positive and some of these are negative and by virtue of the mean, they actually exactly cancel each other while averaging them out as has been noted in footnote 4. Thus in order to measure dispersion one needs to get rid of the sign of the deviation $(X-\mu)$. Simply ignoring the sign mathematically amounts to consideration of the absolute value $|X-\mu|$, which is a non-smooth function leading to mathematical difficulties later on. A smooth operation which gets rid of the sign without distorting the values too much is squaring[6]. This leads to the squared deviation $(X-\mu)^2$, whose average value is the variance. Since one changes the unit by squaring (length becomes area for example) the measure of dispersion expressed in the original unit of measurement is given by the standard deviation σ.

According to ***Definition 3.5***, $V[X] = \sigma^2 = \sum_{x\in\mathcal{X}}(x-\mu)^2 p(x)$. However there is a computationally easier formula for $V[X]$, which is as follows.

$$\begin{aligned} & V[X] \\ &= \sum_{x\in\mathcal{X}} (x-\mu)^2 p(x) \\ &= \sum_{x\in\mathcal{X}} (x^2 - 2\mu x + \mu^2) p(x) \\ &= \sum_{x\in\mathcal{X}} x^2 p(x) - 2\mu \sum_{x\in\mathcal{X}} xp(x) + \mu^2 \\ &= E[X^2] - 2\mu \times \mu + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned} \tag{1}$$

Formula (1) also gives the relationship between the second raw moment and the second central moment. Using binomial theorem it is easy to see that any k-th central moment can be expressed in terms of raw moments of k-th and lesser orders as in formula (1). Before illustrating numerical computation of means and variances, it is possibly better to first get

[5]Note that by **Property 2** the first central moment $E[(X-\mu)]$ equals 0.

[6]This is the standard way of getting rid of the sign in Statistics and we shall see that this technique of squaring for removing the sign is used extensively in the later part of the course.

the motivation for computation of these two quantities. This motivation comes from the following result.

Chebyshev's Inequality: For any r.v. X with mean μ and variance σ^2 and a constant c, $P(|X-\mu| < c\sigma) \geq 1 - \frac{1}{c^2}$.

Proof:

$$\begin{aligned}
&\sigma^2 \\
&= \sum_{x\in\mathcal{X}} (x-\mu)^2 p(x) \\
&\geq \sum_{x:\ x\in\mathcal{X}\ \&\ |x-\mu|\geq c\sigma} (x-\mu)^2 p(x) \qquad \text{(as we are adding positive quantities and the set} \\
&\qquad\qquad \{x:\ x\in\mathcal{X}\ \&\ |x-\mu|\geq c\sigma\} \text{ has at most the same elements as } \mathcal{X}) \\
&\geq c^2\sigma^2 \sum_{x:\ x\in\mathcal{X}\ \&\ |x-\mu|\geq c\sigma} p(x) \quad \text{(as for each } x\in\{x:\ x\in\mathcal{X}\ \&\ |x-\mu|\geq c\sigma\},\ (x-\mu)^2 \geq c^2\sigma^2) \\
&= c^2\sigma^2 P(|X-\mu|\geq c\sigma)
\end{aligned}$$

Above inequality implies that $P(|X-\mu| \geq c\sigma) \leq \frac{1}{c^2}$ which in turn yields the result by complementation law. $\bigtriangledown$

Chebyshev's Inequality states that if one knows the mean and variance of a random variable, then one can get an approximate idea about its probability distribution. Knowledge of the distribution requires knowledge of a function of some sort (like say for example a p.m.f. or a c.d.f.) which requires a lot more information storage than just two numbers like μ and σ. But once equipped with these two quantities, one can readily approximate the probability distribution of any r.v. using Chebyshev's Inequality. This gives us the motivation for summarizing the distribution of any r.v. by computing these two widely used moments. Now let us turn our attention to computation of these two moments.

Example 3.1 (Continued): Given

x	0	1	2	3
$p(x)$	0.064	0.288	0.432	0.216

as the p.m.f. of X, its mean $\mu = 0\times 0.064 + 1\times 0.288 + 2\times 0.432 + 3\times 0.216 = 1.8$. In order to compute the variance we shall use the short-cut formula (1), which requires $E[X^2]$ along with μ, which has just been found to be 1.8. $E[X^2] = 0^2\times 0.064 + 1^2\times 0.288 + 2^2\times 0.432 + 3^2\times 0.216 = 3.96$, and thus $\sigma^2 = 3.96 - 1.8^2 = 0.72$, and $\sigma = \sqrt{0.72} \approx 0.8485$. As an illustration of Chebyshev's inequality with $c = 1.5$, it may be stated that the probability that X lies between $1.8 \pm 1.5\times 0.8485 = 1.8 \pm 1.2725 \approx (0.53, 3.07)$ is at least $1 - \frac{1}{1.5^2} \approx 0.56$, while the actual value of this probability is 1-0.064=0.936. $\bigtriangledown$

Example 3.4 (Continued): Here the p.m.f. of X is given by $p(x) = q^x p$ for $x = 0, 1, 2, \ldots$ and thus in order to compute its mean and variance we need to find the sum of a couple of infinite series. First let us compute its mean μ which is same as

$$\begin{aligned}
&E[X] \\
&= \sum_{x=0}^{\infty} x q^x p \\
&= 0\times p + 1\times qp + 2\times q^2 p + 3\times q^3 p + 4\times q^4 p + \cdots
\end{aligned}$$

$$
\begin{aligned}
&= p\left[q + 2q^2 + 3q^3 + 4q^4 + \cdots\right] \\
&\quad p[\quad q \\
&= \qquad +q^2 \quad +q^2 \\
&\qquad +q^3 \quad +q^3 \quad +q^3 \\
&\qquad +q^4 \quad +q^4 \quad +q^4 \quad +q^4 \\
&\qquad \vdots \quad\ \vdots \quad\ \vdots \quad\ \vdots \quad] \\
&= p\left[\frac{q}{1-q} + \frac{q^2}{1-q} + \frac{q^3}{1-q} + \frac{q^4}{1-q} + \cdots\right] \\
&= \frac{p}{1-q}\frac{q}{1-q} \\
&= \frac{q}{p}
\end{aligned}
$$

For computation of the variance, according to (1) we first need to find the sum of the infinite series $\sum_{x=0}^{\infty} x^2 q^x p$, which is $E[X^2]$. Here we shall employ a different technique for evaluating this sum. First note that $\sum_{x=0}^{\infty} q^x = \frac{1}{1-q}$. Thus

$$
\begin{aligned}
&\tfrac{\partial}{\partial q}\textstyle\sum_{x=0}^{\infty} q^x \\
&= \textstyle\sum_{x=0}^{\infty} \frac{\partial}{\partial q} q^x \\
&= \textstyle\sum_{x=0}^{\infty} x q^{x-1} \\
&= \tfrac{\partial}{\partial q}\tfrac{1}{1-q} \\
&= \tfrac{1}{(1-q)^2}
\end{aligned}
\qquad \text{and} \qquad
\begin{aligned}
&\tfrac{\partial}{\partial q}\textstyle\sum_{x=0}^{\infty} x q^{x-1} \\
&= \textstyle\sum_{x=0}^{\infty} \frac{\partial}{\partial q} x q^{x-1} \\
&= \textstyle\sum_{x=0}^{\infty} x(x-1) q^{x-2} \\
&= \tfrac{\partial}{\partial q}\tfrac{1}{(1-q)^2} \\
&= \tfrac{2}{(1-q)^3}
\end{aligned}
\quad .
$$

Therefore

$$
\sum_{x=0}^{\infty} x^2 q^x = \frac{2q^2}{(1-q)^3} + \sum_{x=0}^{\infty} x q^x = \frac{2q^2}{(1-q)^3} + \frac{q}{(1-q)^2} = \frac{q+q^2}{(1-q)^3} \Rightarrow E[X^2] = \frac{q+q^2}{(1-q)^2}
$$

and thus

$$
V[X] = E[X^2] - (E[X])^2 = \frac{q+q^2}{(1-q)^2} - \frac{q^2}{(1-q)^2} = \frac{q}{p^2} \qquad \bigtriangledown
$$

Mean and variance/standard deviation respectively are the standard moment based measures of location and spread. There are a couple more summary measures which are also typically reported for a distribution. These are measures of skewness and kurtosis. Like the measures of location and spread, these measures are also not unique. However they have fairly standard moment based measures.

Skewness measures the symmetry of a distribution. For this it is very natural to consider the third central moment, usually denoted by α_3, of the distribution. That is $\alpha_3 = E[(X-\mu)^3] = \sum_{x\in\mathcal{X}}(x-\mu)^3 p(x)$, where μ denotes the mean of the distribution. Note that if a distribution is symmetric then its $\alpha_3 = 0$. For a distribution with a heavier right tail $\alpha_3 > 0$ and likewise $\alpha_3 < 0$ for a distribution with a long left tail. Thus the nature of the skewness of a distribution is readily revealed by the sign of α_3. However the exact numerical value of α_3 is also affected by the spread of a distribution, and thus a direct comparison of the α_3 values between two distributions does not provide any indication of whether one distribution

is more skewed than the other. Furthermore it is desirable to have the measure of skewness of a distribution as a pure number free of any units so that it remains unaffected by the scale of measurement. These considerations lead one to define the moment based measure of skewness as

$$\beta_1 = \frac{\alpha_3}{\sigma^3}$$

where σ is the standard deviation of the distribution. β_1 is called the **Coefficient of Skewness**.

Kurtosis measures the peakedness of a distribution. By peakedness one means how sharp or flat is the p.m.f.. Again here for this it is natural to consider $E[(X-\mu)^k]$ for some even power k. Since for $k = 2$, $E[(X-\mu)^k]$ already measures the spread or variance of the distribution, we use the next even k *i.e.* $k = 4$ for the kurtosis. Thus let $\alpha_4 = E[(X-\mu)^4] = \sum_{x\in\mathcal{X}}(x-\mu)^4 p(x)$. Just as in the case of skewness, here again α_4 by itself gets affected by the variability and is not unit free. This problem is circumvented by defining the **Coefficient of Kurtosis** as

$$\beta_2 = \frac{\alpha_4}{\sigma^4}$$

where σ is the standard deviation of the distribution. Now peakedness is not really an absolute concept like symmetry. By that we mean that just having the value of β_2 is not enough unless it can be compared with something which is a "standard" measure of peakedness. For this purpose one uses the most widely used (continuous) probability model called the Normal or Gaussian distribution whose density function looks like the ubiquitous so-called "bell curve". The Normal distribution or the bell-curve has a β_2 of 3, which serves the purpose of being used as the required bench-mark. Thus distributions with, $\beta_2 = 3$ are called **mesokurtic** meaning their p.m.f. has a peakedness comparable to the bell-curve; $\beta_2 > 3$ are called **leptokurtic** meaning their p.m.f. has a peakedness sharper than the bell-curve; and $\beta_2 < 3$ are called **platokurtic** meaning their p.m.f. has a peakedness flatter than the bell-curve.

3.2.2 Quantiles

A second class of summary measures of a distribution is expressed in terms of the quantiles.

Definition 3.6: For $0 < p < 1$, ξ_p is called the **p-th quantile** of a r.v. X if
a. $F(\xi_p) \geq p$, and
b. $F(\xi_p-) \leq p$
where $F(\cdot)$ is the c.d.f. of X and $F(\xi_p-) = \lim_{x\to\xi_p-} F(x)$ is the left-hand limit of $F(\cdot)$ at ξ_p.

p-th quantile of a distribution is nothing but its $100p$-th percentile. That is ξ_p is such a value that the probability of the r.v. taking a value less than or equal to that is about p and at least that is about $1-p$. Thus for example the **median** of a distribution would be denoted by $\xi_{0.5}$. However for a discrete r.v. (like say as in **Example 1**) there may not exist any value such that the probability of exceeding it is exactly 0.5 and falling below it

is also 0.5, and thus leaving the notion of median undefined unless defined carefully. The purpose of ***Definition 6*** is precisely that, so that no matter what p is, it will always yield an answer for ξ_p, which is what it should be for the easily discernible cases and the closest to the conceptual value for the not so easy ones. Thus though ***Definition 6*** might look overly complex for a simple concept, it is a necessary evil we have to live with, and the only way to appreciate it would be to work with it through a few examples and then examine whether the answer it has yielded makes sense.

Example 3.1 (Continued): As pointed out above, it appears that there is no obvious answer for the median of this r.v.. However from the c.d.f. of X given in page 6 just above its graph in Figure 1, we find that the number 2 satisfies both the conditions required by ***Definition 6*** for $p = 0.5$. Let's see why. Condition **a** stipulates that $F(\xi_{0.5}) \geq 0.5$ and here $F(2) = 0.784$ satisfies this condition, and $F(2-) = \lim_{x \to 2-} F(x) = 0.352 \leq 0.5$ satisfies condition **b**. Thus according to ***Definition 6***, the median of this r.v. is 2. Now let's examine why this answer makes sense. Consider a large number of trials of this experiment where in one trial you toss the coin three times and note down the number of Heads in that trial. Now after a large number of trials, say N, assumed to be odd, you will have about $0.064N$ many 0's, $0.288N$ many 1's, $0.432N$ many 2's and $0.216N$ many 3's. If you put these numbers in ascending order and then look for the number in the $(N+1)/2$-th position you will get a 2. Thus the answer for the median being 2 makes perfect sense! $\bigtriangledown$

Example 3.2 (Continued): Suppose we are interested in determining $\xi_{\frac{1}{6}}$ or the $16.\dot{6}$-th percentile of X. The c.d.f. of X is given at the bottom of page 6 in a tabular form. From this table it may be seen that $F(x) = \frac{1}{6}\ \forall x \in [4, 5)$. Therefore $\forall x \in (4, 5)$, $F(x) \geq \frac{1}{6}$ and $F(x-) = F(x) \leq \frac{1}{6}$. Thus any real number between 4 and 5 exclusive qualifies to be called as $\xi_{\frac{1}{6}}$ of X. Now $F(4) = \frac{1}{6}$ and $F(4-) = \frac{1}{12} \leq \frac{1}{6}$, and $F(5) = \frac{5}{18} \geq \frac{1}{6}$ and $F(5-) = \frac{1}{6}$. Thus the numbers 4 and 5 also satisfy the conditions of ***Definition 6*** for $p = \frac{1}{6}$ and are thus legitimate values for $\xi_{\frac{1}{6}}$ of X. It may also be noted that no other value outside the closed interval $[4, 5]$ satisfies both the conditions of ***Definition 6*** for $p = \frac{1}{6}$. Hence $\xi_{\frac{1}{6}}$ of X is not unique and any real number between 4 and 5 inclusive may be regarded as the $16.\dot{6}$-th percentile of X, and these are its only possible values. The reason why this answer makes sense can again be visualized in terms of a large number of trials of this experiment. The required value will be sometimes 4, sometimes 5 and sometimes in between. $\bigtriangledown$

Once we have learned how to figure out the quantiles of a distribution next let us turn our attention to quantile based summary measures of a distribution. We are mainly concerned with measures of location, spread and skewness, whose moment based measures are given by the mean, variance/standard deviation and β_1 respectively, as has been seen in §2.1. Analogous to the mean, the standard quantile based measure of location is given by the **Median** $\xi_{0.5}$. $\xi_{0.25}$, $\xi_{0.5}$ and $\xi_{0.75}$ are the three numbers which divide $\mathcal{X}$, the sample space of the r.v. X, in four equal parts in the sense that $P(X \leq \xi_{0.25}) \approx 0.25$, $P(\xi_{0.25} < X \leq \xi_{0.5}) \approx 0.25$, $P(\xi_{0.5} < X \leq \xi_{0.75}) \approx 0.25$ and $P(X > \xi_{0.75}) \approx 0.25$. Since they divide $\mathcal{X}$ in (approximately) four equal parts these three quantiles, $\xi_{0.25}$, $\xi_{0.5}$ and $\xi_{0.75}$ are together called **quartiles**. Analogous to the standard deviation, the standard quantile based measure of spread is given by the **Inter Quartile Range** (**IQR**) defined as $\xi_{0.75} - \xi_{0.25}$. For a

symmetric distribution $\xi_{0.25}$ and $\xi_{0.75}$ are equidistant from $\xi_{0.5}$. For a positively skewed distribution the distance between $\xi_{0.75}$ and $\xi_{0.5}$ is more than that between $\xi_{0.25}$ and $\xi_{0.5}$, and like wise for a negatively skewed distribution the distance between $\xi_{0.75}$ and $\xi_{0.5}$ is less than that between $\xi_{0.25}$ and $\xi_{0.5}$. Thus the difference between the two distances $\xi_{0.75} - \xi_{0.5}$ and $\xi_{0.5} - \xi_{0.25}$ appear to be a good indicator of skewness of a distribution. However just as in the case of moment based measure, this difference $\xi_{0.25} + \xi_{0.75} - 2\xi_{0.5}$ though captures skewness, remains affected by the spread of the distribution and is sensitive to the scale of measurements. Thus an appropriate quantile based measure of skewness may be found by dividing $\xi_{0.25} + \xi_{0.75} - 2\xi_{0.5}$ by the just described quantile based measure of spread IQR. Thus a quantile based **Coefficient of Skewness** is given by $\frac{\xi_{0.25}+\xi_{0.75}-2\xi_{0.5}}{\xi_{0.75}-\xi_{0.25}}$.

3.2.3 Examples

We finish our discussion on discrete r.v. after solving a few problems.

Example 3.6: If 6 trainees are randomly assigned to 4 projects find the mean, standard deviation, median, and IQR of the number of projects with none of the trainees assigned to it.

Solution: Let X denote the number of projects with none of the trainees assigned to it. Then X is a discrete r.v. with $\mathcal{X} = \{0, 1, 2, 3\}$. In order to find its moment and quantile based measures of location and spread we have to first figure out its p.m.f.

First lest us try to find $P[X = 0]$. In words, the event $[X = 0]$ means none of the projects is empty (here for the sake of brevity, the phrase "a project is empty" will be used to mean no trainee is assigned to it). Thus for $i = 1, 2, 3, 4$ let A_i denote the event, "i-th project is empty". Then $[X = 0] = (\cup_{i=1}^{4} A_i)^c$ and thus

$$
\begin{aligned}
& P[X = 0] \\
&= 1 - P\left(\cup_{i=1}^{4} A_i\right) && \text{(by the complementation law)} \\
&= 1 - \left\{\sum_{i=1}^{4} P(A_i) - \sum_{i \neq j} P(A_i \cap A_j) + \sum_{i \neq j \neq k} P(A_i \cap A_j \cap A_k) + P(A_1 \cap A_2 \cap A_3 \cap A_4)\right\} \\
& && \text{(by eqn. (2) of "Elementary Probability Theory")} \\
&= 1 - \left\{4 \times \frac{3^6}{4^6} - 6 \times \frac{2^6}{4^6} + 4 \times \frac{1}{4^6}\right\} && \text{(because, a) the event } A_i \text{ can happen in } 3^6 \text{ ways out} \\
& && \quad \text{of the possible } 4^6, \text{ and there are 4 of them;} \\
& && \quad \text{b) the event } A_i \cap A_j \text{ can happen in } 2^6 \text{ ways} \\
& && \quad \text{and there are } \binom{4}{2} = 6 \text{ of them;} \\
& && \quad \text{c) the event } A_i \cap A_j \cap A_k \text{ can happen in only} \\
& && \quad 1 \text{ way and there are } \binom{4}{3} = 4 \text{ of them; and} \\
& && \quad \text{d) the event } A_1 \cap A_2 \cap A_3 \cap A_4 \text{ is impossible)} \\
&= 0.380859
\end{aligned}
$$

Next let us figure out $P[X=1]$. There are $\begin{pmatrix} 4 \\ 1 \end{pmatrix} = 4$ ways of choosing the project that will remain empty. Now since $X=1$, the remaining 3 must be non-empty. Thus for these three non-empty projects, for $i=1,2,3$ let B_i denote the event, "i-th project is empty". Then the event, "the remaining three projects are non-empty" is same as $(\cup_{i=1}^{3} B_i)^c$, and thus the number of ways that can happen equals $3^6 \times [1 - P(\cup_{i=1}^{3} B_i)]$, as there are 3^6 ways of assigning the 6 trainees to the 3 projects. Now

$$
\begin{aligned}
& P\left(\cup_{i=1}^{3} B_i\right) \\
&= \sum_{i=1}^{3} P(B_i) - \sum_{i \neq j} P(B_i \cap B_j) + P(B_1 \cap B_2 \cap B_3) \quad \text{(by eqn. (2) of "Elementary Probability Theory")} \\
&= 3 \times \frac{2^6}{3^6} - 3 \times \frac{1}{3^6} \quad \text{(because, a) the event } B_i \text{ can happen in } 2^6 \text{ ways out of the possible } 3^6 \text{ and there are 3 of them;}
\end{aligned}
$$

b) the event $B_i \cap B_j$ can happen in only 1 way and there are 3 of them; and

c) the event $B_1 \cap B_2 \cap B_3$ is impossible)

Thus the total number of ways the event $[X=1]$ can happen is $4 \times (3^6 - 3 \times 2^6 + 3)$, and therefore $P[X=1] = \frac{4 \times \left(3^6 - 3 \times 2^6 + 3\right)}{4^6} = 0.527344$.

The event $[X=2]$ can happen in $6 \times (2^6 - 2)$ ways. This is because if exactly 2 projects are empty the remaining 2 must be non-empty and the 4 projects can be divided into two groups of 2 empty and 2 non-empty in $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ ways. Now there are 2^6 ways of assigning 6 trainees to the 2 non-empty projects, but since out of these there are 2 possibilities in which all the 6 trainees get assigned to the same project, the number of cases where they are assigned to the 2 projects such that none of the 2 projects is empty is $2^6 - 2$. Therefore $P[X=2] = \frac{6 \times (2^6 - 2))}{4^6} = 0.090820$.

The event $[X=3]$ can happen in $\begin{pmatrix} 4 \\ 1 \end{pmatrix} = 4$ ways. Choose one project out of 4 in $\begin{pmatrix} 4 \\ 1 \end{pmatrix} = 4$ ways and then all the 6 trainees are assigned to it. Thus $P[X=2] = \frac{4}{4^6} = 0.000977$. As a check note that

$$
\begin{aligned}
& P[X=0] + P[X=1] + P[X=2] + P[X=3] \\
&= 1 - \frac{1}{4^6}\left(4 \times 3^6 - 6 \times 2^6 + 4 - 4 \times 3^6 + 12 \times 2^6 - 12 - 6 \times 2^6 + 12\right) \\
&= 0.380859 + 0.527344 + 0.090820 + 0.000977 \\
&= 1
\end{aligned}
$$

Thus the p.m.f. of X is given by

x	0	1	2	3
$p(x)$	0.380859	0.527344	0.090820	0.000977

and therefore its mean $\mu = 0 \times 0.380859 + 1 \times 0.527344 + 2 \times 0.090820 + 3 \times 0.000977 = 0.711915$.

$E[X^2] = 0^2 \times 0.380859 + 1^2 \times 0.527344 + 2^2 \times 0.090820 + 3^2 \times 0.000977 = 0.899417$ and thus its standard deviation $\sigma = \sqrt{0.899417 - 0.711915^2} = \sqrt{0.392594} = 0.626573$.

$\xi_{0.5}$, the median of X is 1. This is because this is the only point where $F(1) = 0.908203 \geq 0.5$ and $F(1-) = 0.380859 \leq 0.5$, where $F(\cdot)$ is the c.d.f. of X. Likewise its first quartile $\xi_{0.25} = 0$, as $F(0) = 0.380859 \geq 0.25$ and $F(0-) = 0 \leq 0.25$; and the third quartile $\xi_{0.75} = 1$, as $F(1) = 0.908203 \geq 0.75$ and $F(1-) = 0.380859 \leq 0.75$; and therefore the IQR of X is 1. $\bigtriangledown$

Example 3.7: Let X denote the number of bugs present in the first version of a software. It has been postulated that $P[X = n] = c\frac{p^{n+1}}{n+1}$ for $n = 0, 1, 2, \ldots$ for some $0 < p < 1$.

a. Find the normalizing constant c in terms of the parameter p.

b. If for a certain software $p = 0.5$, what is the probability of it being free from any bug?

c. Find the expected number of bugs in a software having the postulated p.m.f..

Solution (a): The normalizing constant c must be such that $\sum_{n=0}^{\infty} P[X = n] = c\sum_{n=0}^{\infty} \frac{p^{n+1}}{n+1} = 1$. This basically amounts to finding the sum of the infinite series $\sum_{n=1}^{\infty} \frac{p^n}{n}$. Those of you who can recall the logarithmic series, should be able to immediately recognize that since for $|x| < 1$, $\log(1-x) = -\frac{x}{1} - \frac{x^2}{2} - \frac{x^3}{3} - \cdots$, $\sum_{n=1}^{\infty} \frac{p^n}{n} = -\log(1-p)$ as it is given that $0 < p < 1$. For those (like me) who cannot remember all kinds of weird formulæ here is how the series can be summed recalling the technique employed earlier in **Example 4**. For $0 < p < 1$

$$\sum_{n=0}^{\infty} p^n = \frac{1}{1-p} \Rightarrow \int \left\{\sum_{n=0}^{\infty} p^n\right\} dp = \sum_{n=0}^{\infty} \int p^n dp = \sum_{n=0}^{\infty} \frac{p^{n+1}}{n+1} = \int \frac{1}{1-p} dp + k = -\log(1-p) + k$$

for some constant k. For $p = 0$ since both $\sum_{n=0}^{\infty} \frac{p^{n+1}}{n+1}$ and $-\log(1-p)$ equal 0, $k = 0$. Therefore $\sum_{n=0}^{\infty} \frac{p^{n+1}}{n+1} = -\log(1-p)$ implying $c(-\log(1-p)) = 1$. Hence $c = \frac{1}{-\log(1-p)} = \frac{1}{\log\left(\frac{1}{1-p}\right)}$.

(b): For $p = 0.5$, $c = 1/\log 2 \approx 1.4427$, and we are to find $P[X = 0]$, the answer to which is $1.4427 \times 0.5 = 0.7213$.

(c): Here we are to find $E[X]$, which again basically boils down to summing an infinite series, which is done as follows.

$$\begin{aligned}
E[X] & \\
&= \sum_{n=0}^{\infty} nP[X = n] \\
&= c\sum_{n=0}^{\infty} n\frac{p^{n+1}}{n+1} \\
&= c\sum_{n=0}^{\infty} \left\{(n+1)\frac{p^{n+1}}{n+1} - \frac{p^{n+1}}{n+1}\right\} \\
&= c\sum_{n=0}^{\infty} p^{n+1} - c\sum_{n=0}^{\infty} \frac{p^{n+1}}{n+1} \\
&= \frac{p}{(1-p)\log\left(\frac{1}{1-p}\right)} - 1
\end{aligned}$$

$\bigtriangledown$

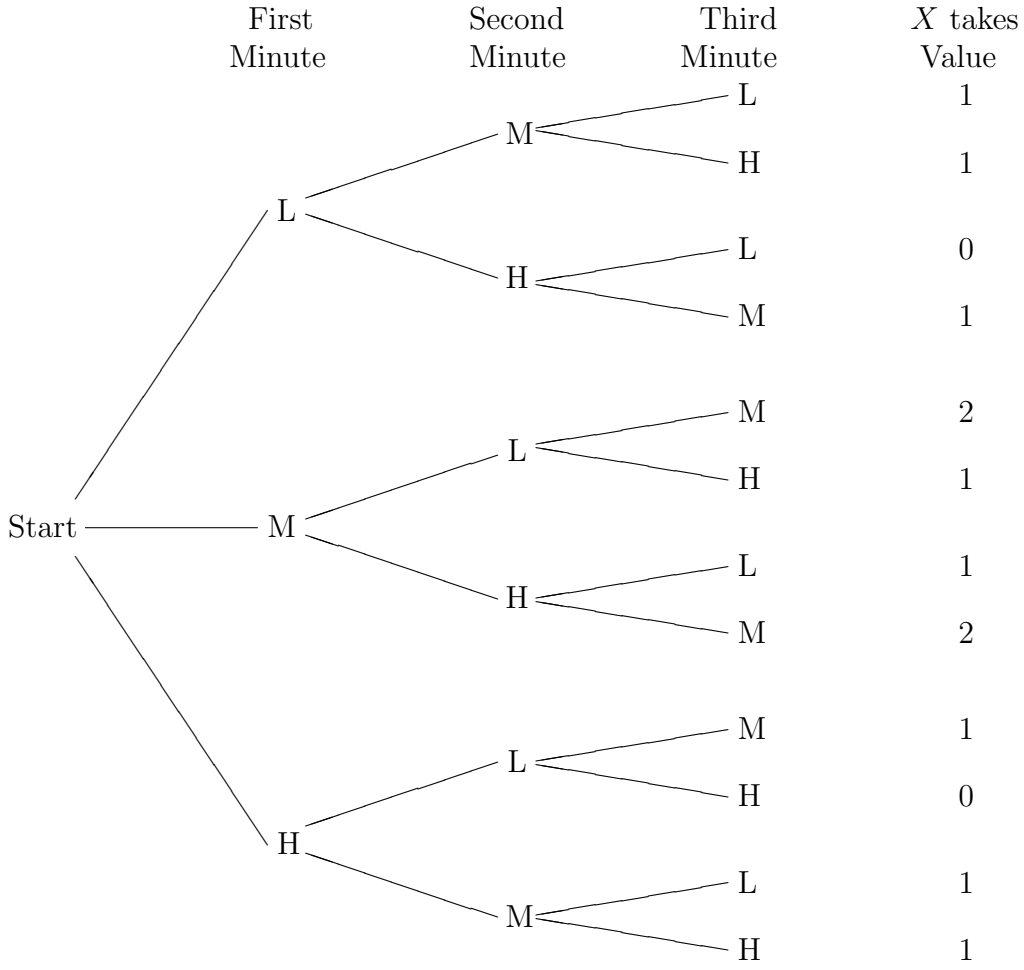
Example 3.8: The shop-floor of a factory has no control over humidity. The newly installed CNC machine however requires a humidity setting either at low, medium or high. The machine is therefore allowed to use its default program of starting at one of the humidity settings at random, and thereafter changing the setting every minute, again at random. Let X denote the number of minutes the machine runs in the medium setting in the first 3 minutes of its operation.Answer the following:

a. Plot the p.m.f. of X.

b. Sketch the c.d.f. of X.

c. Find the mean, standard deviation, median and IQR of X.

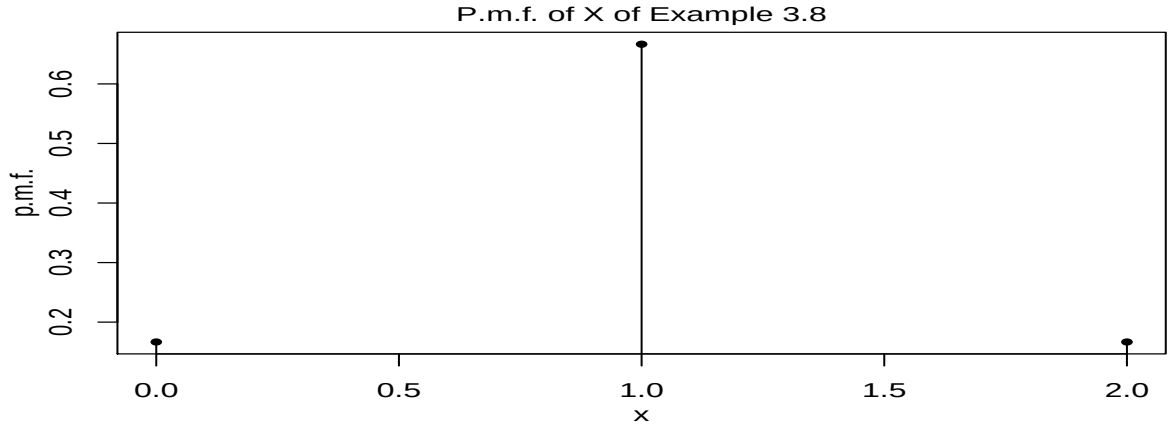
Solution (a): Denote Low, Medium and High by L, M and H respectively. Then the possible configurations for the first three minutes and hence the value of X can be figured out with the help of the following tree digram:



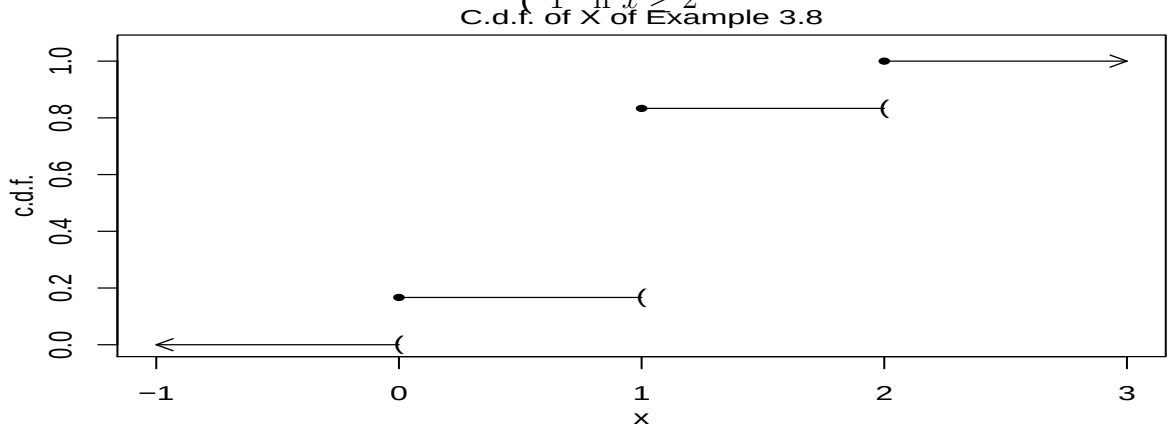
Thus it can be seen that there are a total of 12 possibilities and each has a probability of $\frac{1}{12}$. This may be verified as follows. Take any configuration, say MHL. $P(\text{MHL}) = P(\text{M in the First Minute}) \times P(\text{H in the Second Minute} \mid \text{M in the First Minute}) \times P(\text{L in the Third Minute} \mid \text{M in the First Minute AND H in the Second Minute}) = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{12}$. Now in order to figure out the p.m.f. of X one simply need to count the number of times it assumes the values 0, 1, and 2 respectively. These counts yield the p.m.f. of X as

x	0	1	2
$p(x)$	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

the plot of which is given in the next page.



(b): The c.d.f. of X is given by $F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{6} & \text{if } 0 \leq x < 1 \\ \frac{5}{6} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$, which is plotted below.



(c): Mean $\mu = 0\times\frac{1}{6}+1\times\frac{2}{3}+2\times\frac{1}{6} = 1$. Standard Deviation $\sigma = \sqrt{\left(0^2 \times \frac{1}{6} + 1^2 \times \frac{2}{3} + 2^2 \times \frac{1}{6}\right) - 1} = \frac{1}{\sqrt{3}}$. Median $\xi_{0.5} = 1$ as $F(1) = \frac{5}{6} \geq 0.5$ and $F(1-) = \frac{1}{6} \leq 0.5$. IQR $= \xi_{0.75} - \xi_{0.25} = 1 - 1 = 0$. $\bigtriangledown$

3.3 Continuous R.V.

Definition 3.7: A r.v. X is said to be **continuous** if its c.d.f. $F(x)$ is continuous $\forall x \in \Re$.

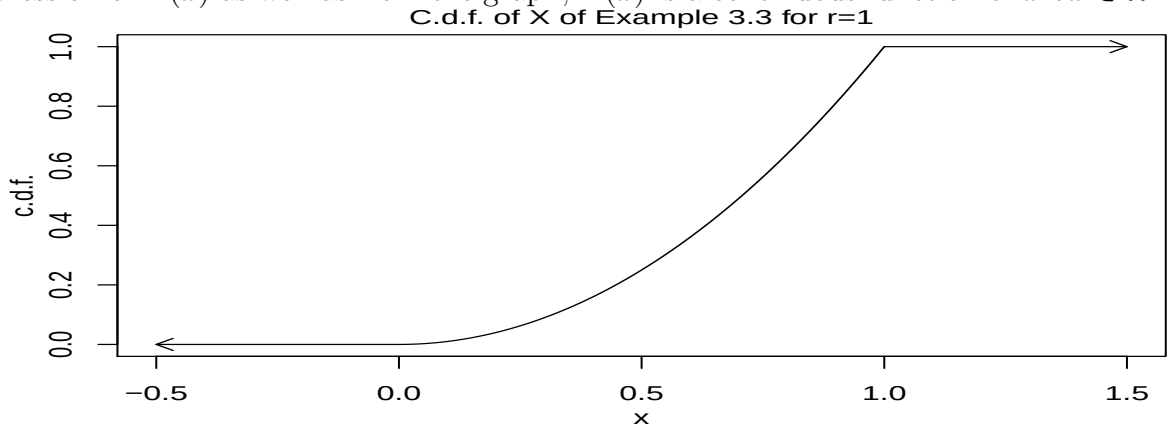
Note the difference between ***Definitions 2*** and ***7***. While a discrete r.v. is defined in terms of its sample space $\mathcal{X}$, a continuous r.v. is not defined as $\mathcal{X}$ being uncountable. What it means is since $P[X = x] = P[X \leq x] - P[X < x] = F(x) - F(x-)$ (see ***Property 4*** in Appendix A), and for a continuous $F(x)$, $F(x) = F(x-)\ \forall x \in \Re$, $P[X = x] = 0\ \forall x \in \Re$ for a continuous r.v. X. Or in other words alternatively but equivalently, a r.v.X may be defined to be continuous iff $P[X = x] = 0\ \forall x \in \Re$. Thus the notion of p.m.f. remains undefined for a continuous r.v. (as $\sum_{x\in\mathcal{X}} P[X = x]$ needs to equal 1 for a p.m.f.). The definition also goes on to show that a r.v. cannot simultaneously be discrete as well as continuous. This is because we have already seen that the c.d.f. of a discrete r.v. is necessarily a discontinuous step function and it is just argued that the notion of p.m.f. remains undefined for a continuous r.v..

We begin our discussion on continuous r.v. with a couple of examples, where we handle them in terms of their c.d.f., as it has already been seen in the discrete case how the c.d.f. may be used for probability calculations.

Example 3.3 (Continued): Consider the r.v. X denoting the distance from the bull's eye of a dirt thrown into a dartboard of radius r, where it is assumed that the dirt always lands somewhere on the dartboard. Now further assume that the dirt is "equally likely" to land anywhere on the dartboard. By this it is meant that the probability of the dirt landing within any region R of the dartboard is proportional to the area of R and does not depend on the exact location of R on the dartboard. Now with the Ω of this experiment equipped with this probability, let us figure out the c.d.f. of the r.v. X. Let $0 < x < r$. $F(x) = P[X \leq x]$ and the event $[X \leq x]$ can happen if and only if the dirt lands within the circular region of the board with the bull's eye in its center and radius x, which has an area of πx^2. Thus $F(x) \propto \pi x^2$. The constant of this proportionality is found to be $\frac{1}{\pi r^2}$ by observing that $F(r) = 1$ and $F(r) \propto \pi r^2$. Thus for $0 < x < r$ $F(x) = \frac{x^2}{r^2}$. Now obviously $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq r$. Thus the complete $F(x)$ may be specified as

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{x^2}{r^2} & \text{if } 0 < x < r \\ 1 & \text{if } x \geq r \end{cases}$$

and is plotted below for $r = 1$. Note both analytically from the expression of $F(x)$ as well as from the graph, $F(x)$ is a continuous function of x $\forall x \in \Re$. $\bigtriangledown$

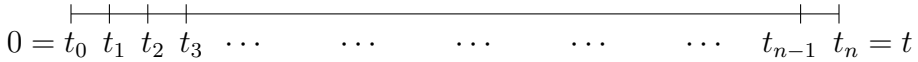


Example 3.9: Suppose we want to model the r.v. T denoting the number of hours a light bulb will last. In order to do so we make the following two postulates:

i. The probability that the light bulb will fail in any small time interval $(t, t+h]$ does not depend on t and equals $\lambda h + o(h)$ for some $\lambda > 0$, called the failure rate, and $o(h)$ is a function such that $\lim_{h \to 0} \frac{o(h)}{h} = 0$.

ii. If $(s, t]$ and $(u, v]$ are disjoint time intervals then the events of the light bulb failing in these two intervals are independent.

We shall find the distribution of T by figuring out $P[T > t]$ which is same as $1 - F(t)$, where $F(\cdot)$ is the c.d.f. of T. Fix a $t > 0$. Then the event $[T > t]$ says that the light bulb has not failed till time t. Divide the time interval $[0, t]$ in n equal and disjoint parts of length h as $[0, t] = [0 = t_0, t_1] \cup (t_1, t_2] \cup \cdots \cup (t_{n-1}, t_n = t]$ such that $h = \frac{t}{n}$ and for $i = 1, 2, \ldots, n$, $t_i - t_{i-1} = h$, as in the following diagram:



Now the event $[T > t]$ is same as the event that the light bulb has not failed in any of the intervals $(t_{i-1}, t_i]$ for $i = 1, 2, \ldots, n$, and this must be true for $\forall h > 0$. Therefore,

$$
\begin{aligned}
&P[T > t] \\
&= \lim_{n\to\infty} P\left[\cap_{i=1}^{n}\{\text{the light bulb has not failed in the time interval } (t_{i-1}, t_i]\}\right] \\
&\qquad\qquad (\text{as } n \to \infty \Leftrightarrow h \to 0) \\
&= \lim_{n\to\infty} \prod_{i=1}^{n} P\left[\text{the light bulb has not failed in the time interval } (t_{i-1}, t_{i-1}+h]\right] \\
&\qquad\qquad (\text{by the independence assumption of postulate } \mathbf{ii}) \\
&= \lim_{n\to\infty} \prod_{i=1}^{n} \left[1 - \lambda\frac{t}{n} + o\left(\frac{t}{n}\right)\right] \qquad (\text{since } h = \frac{t}{n} \text{ and by postulate } \mathbf{i}, \text{ probability of} \\
&\quad \text{failing in } (t_{i-1}, t_{i-1}+h] \text{ is } \lambda h + o(h), \text{ and therefore the probability of not failing is} \\
&\quad 1 - \lambda h + o(h), \text{ the sign of } o(h) \text{ being irrelevant}) \\
&= \lim_{n\to\infty} \left[1 - \lambda\frac{t}{n} + o\left(\frac{t}{n}\right)\right]^n \\
&= \lim_{n\to\infty} \exp\left\{n \log\left[1 - \lambda\frac{t}{n} + o\left(\frac{t}{n}\right)\right]\right\} \\
&= \lim_{n\to\infty} \exp\left\{n\left[\left(-\lambda\frac{t}{n} - \lambda^2\frac{t^2}{2n^2} - \lambda^3\frac{t^3}{3n^3} - \cdots\right) - \left(o\left(\frac{t}{n}\right)\frac{t}{n}\lambda + \frac{t}{n}o^2\left(\frac{t}{n}\right)\lambda + \right.\right.\right. \\
&\quad \left.\left.\left.\frac{t^2}{n^2}o\left(\frac{t}{n}\right)\lambda^2 + \cdots\right) + \left(o\left(\frac{t}{n}\right) + \frac{1}{2}o^2\left(\frac{t}{n}\right) + \frac{1}{3}o^3\left(\frac{t}{n}\right) + \cdots\right)\right]\right\} \\
&\quad (\text{as } \log(1-x) = -\sum_{k=1}^{\infty}\frac{x^k}{k} \text{ for } |x| < 1 \text{ and } \left|\lambda\frac{t}{n} + o\left(\frac{t}{n}\right)\right| < 1 \text{ for } n \text{ sufficiently large}) \\
&= \lim_{n\to\infty} \exp\left\{-\lambda t - \left(\lambda^2\frac{t^2}{2n} + \lambda^3\frac{t^3}{3n^2} + \cdots\right) + no\left(\frac{t}{n}\right)\left(1 + \lambda\frac{t}{n} + \lambda^2\frac{t^2}{n^2} + \cdots\right)\right. \\
&\quad \left. + no^2\left(\frac{t}{n}\right)\left(\frac{1}{2} + \lambda\frac{t}{n} + \cdots\right) + \cdots\right\} \\
&= e^{-\lambda t} \quad (\text{as for } k \geq 2, \lim_{n\to\infty} \lambda^k \frac{t^k}{kn^{k-1}} = 0, \\
&\qquad \text{and for } k \geq 1 \lim_{n\to\infty} no^k\left(\frac{t}{n}\right) = \lim_{n\to\infty} \frac{o(t/n)}{t/n}\frac{1}{t}o^{k-1}\left(\frac{t}{n}\right) = 0 \text{ by definition of } o(\cdot))
\end{aligned}
$$

Hence the c.d.f. of T is given by $F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1 - e^{-\lambda t} & \text{if } t > 0 \end{cases}$. Note that $F(t)$ is a continuous function of t. $\bigtriangledown$

Though one can compute probabilities and quantiles with the c.d.f., as mentioned earlier, it does not provide a very helpful graphical feel for the distribution, nor there is any obvious way to compute the moments with the c.d.f.. This calls for an entity which is analogous to p.m.f. as in the discrete case. This entity is called probability density function or p.d.f. defined as follows.

Definition 3.8: A function $f(x) : \Re \to \Re$ is called a **probability density function** if
a. it is non-negative *i.e.* $f(x) \geq 0\ \forall x \in \Re$, and

b. $\int_{-\infty}^{\infty} f(x)dx = 1$ *i.e.* the total area under $f(x)$ is 1.

It turns out that p.d.f. need not exist for an arbitrary continuous r.v.. It exists only for those continuous random variables whose c.d.f. $F(x)$ has a derivative "almost everywhere". Such random variables are called r.v. with a density and its p.d.f. is given by $f(x) = \frac{d}{dx}F(x)$. Here we shall deal with only those continuous random variables which have a density *i.e.* with a differentiable c.d.f.. Thus let us assume that the c.d.f. $F(x)$ is not only continuous but also has a derivative almost everywhere. Now

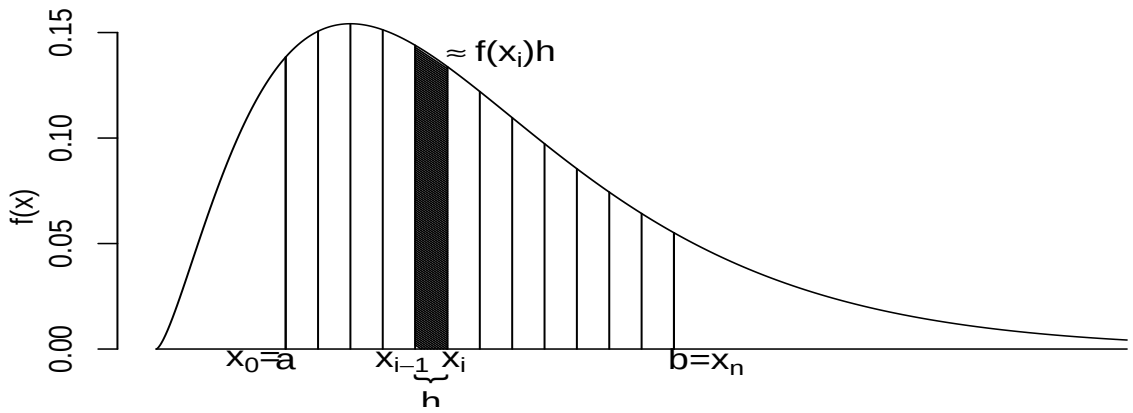
$$f(x) = \frac{d}{dx}F(x) = \lim_{h\to 0}\frac{F(x+h) - F(x)}{h} = \lim_{h\to 0}\frac{P[X \in (x, x \pm h]]}{h} \tag{2}$$

$(x, x \pm h]$ is used to denote the interval $(x, x+h]$ for $h > 0$, and $(x+h, x]$ for $h < 0$. Equation (2) gives the primary connection between p.d.f. and probability, from which it may also be seen why it is called a "density". According to (2), for small h, $P[X \in (x, x \pm h]] \approx hf(x)$ and thus $f(x)$ gives the probability that X takes a value in a small interval around x per unit length of the interval. Thus it is giving probability mass per unit length and hence the name "density".

Now let us examine how probability of X taking values in an interval of arbitrary length (not just in a small interval around a given point x) may be obtained using the p.d.f.. Of course for practical applications one would typically use the c.d.f. for such probability calculations, but then that raises the issue of how can one figure out the c.d.f. if only the p.d.f. is given. Both of these issues *viz.* probability calculation using p.d.f.and expressing the c.d.f. in terms of the p.d.f. are essentially one and the same. Since $f(x)$ is the derivative of $F(x)$, by the second fundamental theorem of integral calculus, $F(x)$ may be expressed as integral of $f(x)$ or $F(x) = \int_{-\infty}^{x} f(t)dt$ and thus $P[a < X \leq b]$ may be found as $\int_a^b f(x)dx$, and thus answering both the questions. However here we shall try to argue these out from scratch without depending on results from calculus which you might have done some years ago.

Let $f(x)$ be a p.d.f. as in Figure 3.2 below and we are interested in computing $P[a \leq X \leq b]$ using this p.d.f..

Figure 2: Probability Calculation Using p.d.f.



First divide the interval $[a, b]$ into n equal parts as $[a, b] = [a = x_0, x_1] \cup [x_1, x_2] \cup \cdots \cup [x_{n-1}, x_n = b]$ such that $h = \frac{b-a}{n}$ and for $i = 1, 2, \ldots, n$, $x_i - x_{i-1} = h$ as in Figure 2 above. For n sufficiently large, or equivalently for h sufficiently small, by (2), $P[x_{i-1} \leq X \leq x_i]$ may be approximated by $f(x_i)h$, as has been indicated with the shaded region in Figure 2. Thus an approximate value of $P[a \leq X \leq b]$ may be found by adding these for $i = 1, 2, \ldots, n$ as $\sum_{i=1}^{n} f(x_i)h$. However this is an approximation because of the discretization, and the exact

value can be obtained by simply considering the limit of $\sum_{i=1}^{n} f(x_i)h$ by letting $h \to 0$ or equivalently $n \to \infty$. Thus

$$P[a \leq X \leq b] = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\frac{b-a}{n} = \int_a^b f(x)dx \tag{3}$$

The last equality follows from the definition of definite integral. Thus we see that probabilities of intervals are calculated by integrating the density function, which is same as finding the area under the p.d.f. over this interval. Now one should be able to see the reason for the conditions **a** and **b** of ***Definition 8***. Non-negativity is required because probability is, and the total area under the p.d.f. is required to be 1 because $P(\Omega) = 1$. The intuitive reason behind this "area" business is as follows. Since $f(x)$ is a probability density (per unit length), in order to find the probability, one has to multiply it by the length of the interval leading to the notion of the area. This is also geometrically seen in Figure 2, where $P[x_{i-1} \leq X \leq x_i]$ is approximated by the area of a rectangle with height $f(x_i)$ and width h. Finally letting $a \to -\infty$ and substituting x for b and changing the symbol of the dummy variable of integration in (3) we get

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{4}$$

which gives the sought expression for c.d.f. in terms of the p.d.f. while the converse is given in (2). Before presenting some more examples pertaining to continuous random variables, we shall first settle the issue of computation of the moments and the quantiles.

Definition 3.9: For a positive integer k, the **k-th raw moment** of a r.v. X with p.d.f. $f(x)$ is given by $\int_{-\infty}^{\infty} x^k f(x)dx$ which is denoted by $E[X^k]$, and the **k-th central moment** is given by $\int_{-\infty}^{\infty}(x-\mu)^k f(x)dx$ which is denoted by $E[(X-\mu)^k]$, where $\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$, the first raw moment, is called the **Expectation** or **Mean** of X.

This definition may be understood if one understands why $\int_{-\infty}^{\infty} xf(x)dx$ is called the mean of a r.v. X with p.d.f. $f(x)$, as the rest of the definition follows from this via the law of unconscious statistician, as has been discussed in detail in §2.1. In order to find $E[X]$, just as in case of finding probabilities of intervals using the p.d.f., we shall first discretize the problem, then use ***Definition 5*** for expectation, and then pass on to the limit to find the exact expression. Thus first consider a r.v. X which takes values in a bounded interval $[a, b]$. Divide the interval $[a, b]$ into n equal parts as $[a, b] = [a = x_0, x_1] \cup [x_1, x_2] \cup \cdots \cup [x_{n-1}, x_n = b]$ such that $h = \frac{b-a}{n}$ and for $i = 1, 2, \ldots, n$, $x_i - x_{i-1} = h$. For n sufficiently large, or equivalently for h sufficiently small, by (2), $P[x_{i-1} \leq X \leq x_i]$ may be approximated by $f(x_i)h$. Thus by ***Definition 5***, an approximate value of $E[X]$ may be found by adding these for $i = 1, 2, \ldots, n$ as $\sum_{i=1}^{n} x_i f(x_i)h$. However this is an approximation because of the discretization, and the exact value can be obtained by simply considering the limit of $\sum_{i=1}^{n} x_i f(x_i)h$ by letting $h \to 0$ or equivalently $n \to \infty$. Thus $E[X] = \lim_{n\to\infty} \sum_{i=1}^{n} x_i f(x_i)\frac{b-a}{n} = \int_a^b xf(x)dx$. Now for an arbitrary r.v. X taking values in $\Re = (\infty, \infty)$, $E[X]$ is found by further considering $\lim_{\substack{a \to -\infty \\ b \to \infty}} \int_a^b xf(x)dx = \int_{-\infty}^{\infty} xf(x)dx$, and thus $E[X] = \int_{-\infty}^{\infty} xf(x)dx$.

Just as in the discrete case, mean $\mu = E[X]$, variance $\sigma^2 = E\left[(X-\mu)^2\right] = E\left[X^2\right] - \mu^2$, coefficient of skewness $\beta_1 = E\left[(X-\mu)^3\right]/\sigma^3$ and coefficient of kurtosis $\beta_2 = E\left[(X-\mu)^4\right]/\sigma^4$.

Chebyshev's inequality, relating the probability distribution to the first two moments - mean and variance, also remains valid in this case. Its proof is identical to the one given in the discrete case with just the summations replaced by integrals. Actually this last remark is true in a quite general sense *i.e.* typically any result involving p.m.f. has a continuous analogue involving p.d.f. which is proved in the same manner as the p.m.f. case with the summations replaced by integrals.

Though moments required a separate definition in the continuous case, the definition of quantile in general was given in ***Definition 6*** which is still valid in this continuous case. However the definition takes a slightly simpler form for the continuous case. **a** of ***Definition 6*** requires ξ_p, the p-th quantile to satisfy the inequality $F(\xi_p) \geq p$, while **b** requires $F(\xi_p-) \leq p$. But for a continuous X since its c.d.f. $F(\cdot)$ is continuous, $F(\xi_p) = F(\xi_p-)$ and thus **a** and **b** together imply that the p-th quantile ξ_p is such that $F(\xi_p) = p$, or ξ_p is the solution (or the set of solutions, as the case may be) of the equation $F(\xi_p) = p$. Now let us work out a few examples involving continuous r.v..

Example 3.10: The length (in inches) of the classified advertisements in a Saturday paper has the following p.d.f. $f(x) \propto \begin{cases} (x-1)^2(2-x) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$. Answer the following:

a. Find the proportionality constant and then graph the p.d.f..

b. What is the probability that a classified advertisement is more than 1.5 inches long?

c. What is the average length of a classified advertisement?

d. What is the modal length of a classified advertisement?

e. What is the median length of a classified advertisement?

f. What is the sign of its coefficient of skewness β_1 ?

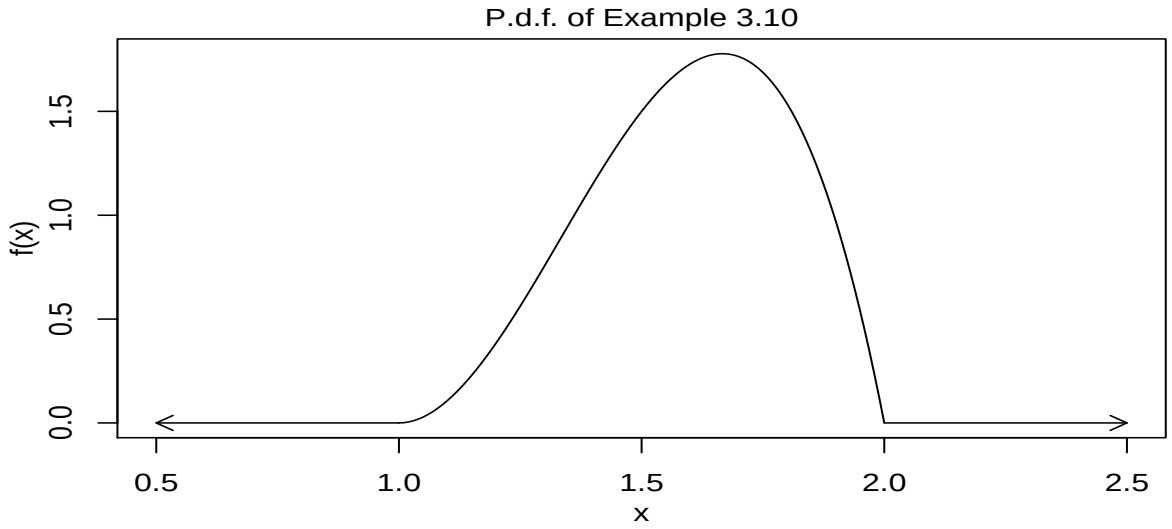
Solution (a): Let the constant of proportionality be c such that

$$f(x) = \begin{cases} c(x-1)^2(2-x) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases} .$$

Now c is determined from **b** of ***Definition 8*** requiring $\int_{-\infty}^{\infty} f(x) = 1$. Note that $f(x) \geq 0$ $\forall x \in \Re$.

$$\begin{aligned} & \int_1^2 (x-1)^2(2-x)dx \\ &= \int_1^2 (-x^3 + 4x^2 - 5x + 2)dx \\ &= -\frac{1}{4}x^4\Big|_{x=1}^{x=2} + \frac{4}{3}x^3\Big|_{x=1}^{x=2} - \frac{5}{2}x^2\Big|_{x=1}^{x=2} + 2x|_{x=1}^{x=2} \\ &= -\frac{15}{4} + \frac{28}{3} - \frac{15}{2} + 2 \\ &= \frac{-45 + 112 - 90 + 24}{12} \\ &= \frac{1}{12} \end{aligned}$$

Therefore c must equal 12. The graph of $f(x)$ is plotted below.



(b):

$$\begin{aligned} & P[X > 1.5] \\ &= 1 - P[X \leq 1.5] \\ &= 1 - 12\int_1^{1.5}(x-1)^2(2-x)dx \\ &= 1 - \left\{-3x^4\Big|_{x=1}^{x=1.5} + 16x^3\Big|_{x=1}^{x=1.5} - 30x^2\Big|_{x=1}^{x=1.5} + 24x\Big|_{x=1}^{x=1.5}\right\} \\ &= 1 - (-12.1875 + 38 - 37.5 + 12) \\ &= 1 - 0.3125 \\ &= 0.6875 \end{aligned}$$

(c):

$$\begin{aligned} & E[X] \\ &= 12\int_1^2(-x^4 + 4x^3 - 5x^2 + 2x)dx \\ &= -\frac{12}{5}x^5\Big|_{x=1}^{x=2} + 12x^4\Big|_{x=1}^{x=2} - 20x^3\Big|_{x=1}^{x=2} + 12x^2\Big|_{x=1}^{x=2} \\ &= -\frac{372}{5} + 180 - 140 + 36 \\ &= 1.6 \end{aligned}$$

(d): Mode of a continuous r.v. is that value around which intervals of same length has maximum probability compared to the other values. This is obviously that point where the density is maximum or the maxima of the p.d.f.. Note that maxima of $f(x)$ is same as the maxima of $\log f(x)$, which is found as follows.

$$\begin{aligned} & \frac{d}{dx}\log f(x) \\ &= \frac{d}{dx}2\log(x-1) + \log(2-x) \\ &= \frac{2}{x-1} - \frac{1}{2-x} \end{aligned}$$

and thus

$$\frac{d}{dx}\log f(x) = 0 \Rightarrow 4 - 2x - x + 1 = 0 \Rightarrow x = 1\frac{2}{3}.$$

Now note that

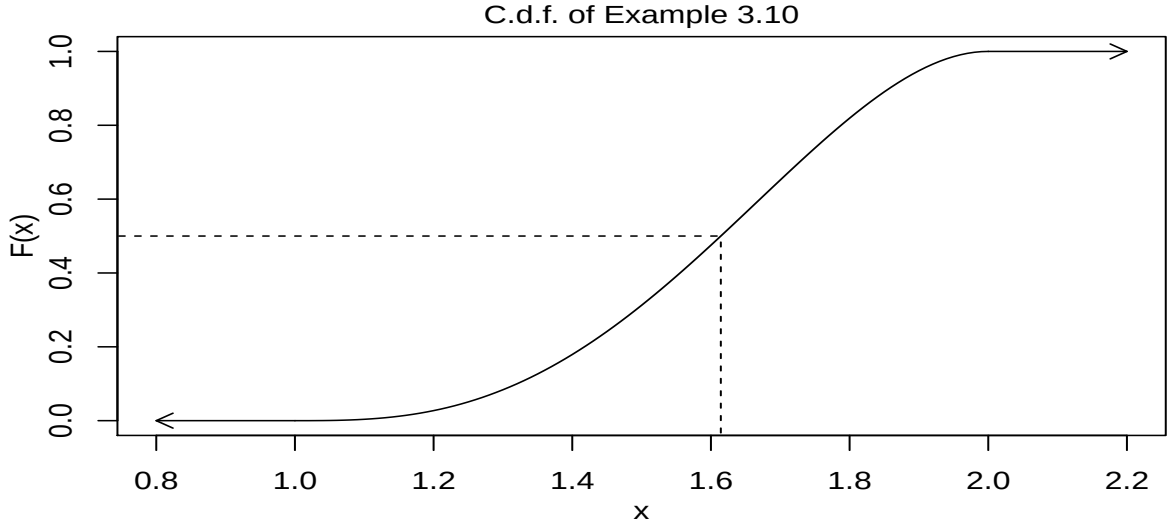
$$\frac{d^2}{dx^2}\log f(x) = -\frac{2}{(x-1)^2} - \frac{1}{(2-x)^2} < 0 \text{ at } x = 1\frac{2}{3}.$$

Therefore $1\frac{2}{3}$ is the maxima of $\log f(x)$ and thus the modal length is $1\frac{2}{3}$ inches.

(e): Any quantile computation requires the c.d.f. $F(x)$ which for $1 \leq x \leq 2$ is given by

$$F(x) = 12\int_1^x (t-1)^2(2-t)dt = -3x^4 + 16x^3 - 30x^2 + 24x - 7$$

which is plotted below, which also depicts the solution of the equation $F(\xi_{0.5}) = 0.5$ through the dotted lines.



Unfortunately the solution of the equation $F(\xi_{0.5}) = 0.5$ requires numerical methods which after applying the method of *regula falsi* yields $\xi_{0.5} = 1.614273$, which is correct up to 6 decimal places.

(f): From the graph of the p.d.f., it is fairly clear that it is slightly negatively skewed, which is also supported by the findings in **c**, **d** and **e** above, that shows that mean $<$ median $<$ mode which is a typical characteristic of negatively skewed distributions. Thus even without the horrendous computation of $E[(X-\mu)^3]$ it may be fairly confidently stated that the sign of β_1, the coefficient of skewness, would be negative. $\bigtriangledown$

Example 3.11: Company X assembles and sells PC's without any warranty for Rs.30,000 at a profit of Rs.5,000. But from the market pressure of competitors, X is now forced to sell an optional 1 year warranty, say W, for their PC's. The probability that a customer will buy W, say p_w, is unknown to X. From the past service and repair records however, X estimates that the chance of a system failure within a year of purchase is 0.01, and the cost of subsequent repairs has a continuous symmetric distribution with mean Rs.2000 and standard deviation Rs.500. It also seems reasonable to assume that the event of a system failure and the amount spent on repairs thereafter, is independent of a customer buying W.

Keeping the price of the PC's bought without W as before, for deciding upon the value of w, the price of W, the only criterion X has set forth is that, the probability of making at least Rs.5,000 profit from a sale of a PC must be at least 0.9999. Answer the following:

a. Express the probability distribution of the Profit, in thousands of Rs..

b. What is the maximum value p_w can have for which, the profit-goal is realized irrespective of the values of w?

c. What should be the minimum value of w, guarding against the worst possible scenario?

d. Give a strategy to bring down the value of w obtained in **c** above in the future, without any engineering improvement or compromise on the criterion of profit-goal.

Solution (a): Let P denote the profit and R denote the repair cost, both in thousands of Rs.. Now there are three possibilities that might arise in which cases the profits are potentially different. First a customer may not buy the warranty, which has a probability of $1 - p_w$, in which case the profit $P = 5$. The second case is where the customer buys the warranty and the PC does not fail within a year. In this case the profit $P = 5 + w$, because w amount (in thousands of Rs.) was paid by the customer for buying W over and above its price of Rs.30,000 that has a profit of Rs.5,000 included in it, but the company did not have to incur any repair cost as the PC did not fail within the warranty period. This case has a probability of $0.99p_w$ as it is assumed that the events of failure within a year and buying the warranty are independent with respective probabilities 0.01 and p_w. The third scenario is where the customer buys W, the PC fails within a year and as a result of which the company had to face a repair cost of R. In this case the profit $P = 5 + w - R$, and its probability is $0.01p_w$. Thus the probability distribution of P may be summarized as $P = \begin{cases} 5 & \text{with probability } 1 - p_w \\ 5 + w & \text{with probability } 0.99p_w \\ 5 + w - R & \text{with probability } 0.01p_w \end{cases}$. Note that P is neither pure discrete nor pure continuous as it has both the components.

(b): The profit-goal is $P[P \geq 5] \geq 0.9999$. But

$$\begin{aligned} & P[P \geq 5] \geq 0.9999 \\ \Leftrightarrow \quad & (1 - p_w) + 0.99p_w + 0.01p_w P[R \leq w] \geq 0.9999 \\ \Leftrightarrow \quad & 1 - 0.01p_w P[R > w] \geq 0.9999 \\ \Leftrightarrow \quad & P[R > w] \leq \frac{0.0001}{0.01p_w} = \frac{0.01}{p_w} \end{aligned}$$

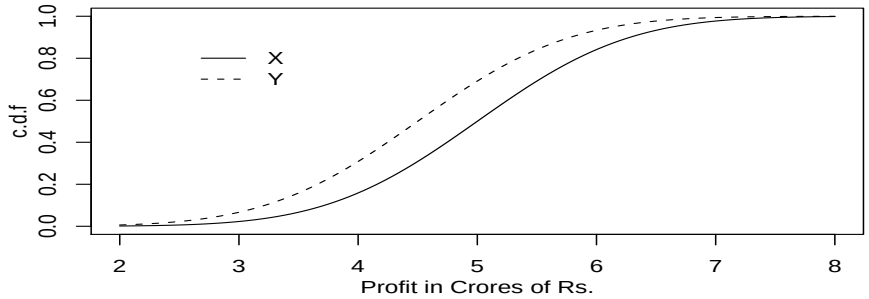
The maximum value $P[R > w]$ can take is 1, and thus no matter whatever be the value of w, the last inequality will always be satisfied as long as $p_w \leq 0.01$ making the ratio $\frac{0.01}{p_w} \geq 1$.

(c): The worst possible scenario is where everybody buys the warranty, in which case $p_w = 1$. In this situation as has just been shown in **b** above, the profit-goal will be realized as long as $P[R > w] \leq 0.01$ *i.e.* $w \geq \xi_{0.99}$, where $\xi_{0.99}$ is the 0.99-th quantile of R. Thus the minimum amount that needs to be charged as the price of the warranty should equal $\xi_{0.99}$. The only information we have about R is that it has a continuous symmetric distribution with mean Rs.2000 and standard deviation Rs.500, based on which we are to figure out $\xi_{0.99}$. The only way we can at least find a bound on $\xi_{0.99}$, based on the above information, is through

Chebyshev's inequality. Now note that Chebyshev's inequality gives probability bounds for symmetric intervals around the mean, while here we are interested in only the right-tail. But this can easily be done because of the additional information about the distribution of R being symmetric. That is by symmetry since $P[R > \xi_{0.99}] = P[R < \xi_{0.01}] = 0.01$, $P[\xi_{0.01} \leq R \leq \xi_{0.99}] = 0.98$ and Chebyshev's inequality says that $P[\mu - c\sigma \leq R \leq \mu + c\sigma] \geq 1 - \frac{1}{c^2}$. Therefore by equating $1 - \frac{1}{c^2} = 0.98$ and solving for c we get that $c \approx 7.0711$ and then by Chebyshev's inequality it follows that $P[2000 - 7.0711 \times 500 \leq R \leq 2000 + 7.0711 \times 500] = P[0 \leq R \leq 5535.53] \geq 0.98$ and thus by symmetry $\xi_{0.99} \leq 5535.53$. Thus the minimum value of w that guarantees the profit goal irrespective of the value of p_w and the exact distribution of R is Rs.5535.53, as long as $E[R] = 2000$ and $SD[R] = 500$.

(d): Rs.5535.53 as the price of an one-year warranty on a Rs.30,000 equipment is preposterous, eroding any competitive edge that X might be envisaging to create by introducing W. The reason for getting such absurdly high minimum value of w is two-fold. Instead of having just the mean and variance of R if we had a reasonable model for the distribution of R that would have given a much sharper estimate of $\xi_{0.99}$. Thus strategically it will be beneficial to have a probability model of R and not just its two moments. Additionally, the second way of reducing the value w would be to keep track of the number of customers buying the warranty, so that one has a reasonable estimate of p_w. This is because recall that in **c**, we are dealing with the worst-case scenario of $p_w = 1$. Thus the strategy would be a) model the distribution of R; and simultaneously b) keep collecting data on number of warranty sold for estimation of p_w. $\bigtriangledown$

Example 3.12: The c.d.f's of annual profits of market regions X and Y denoted by X and Y respectively are as follows:



Which region is more profitable and why?

Solution: Let $F_X(\cdot)$ and $F_Y(\cdot)$ respectively denote the c.d.f.'s of X and Y. Fix any x. Then from the above graph it is clear that $F_X(x) \leq F_Y(x)$ or $\forall x\ P[X \leq x] \leq P[Y \leq x]$, or in other words, probability of making a profit of x or less in market region Y is more likely than region X. Thus obviously X is more profitable. In general, for two random variables X and Y if $F_X(x) \leq F_Y(x)\ \forall x \in \Re$, then X is said to be **stochastically larger** than Y which is written as $X \overset{\text{st.}}{\geq} Y$. $\bigtriangledown$

3.4 Transformations

Quite often we have to deal with functions of a r.v.. For example if we have a model for the side X of a random square, and we are interested in its area, we have to work with the function $g(X) = X^2$. If the constant velocity V of a particle is random while covering a fixed distance S, then the time taken to cover this distance is the r.v. $g(V) = S/V$, which is a function of V. If the annual rate of return R of an investment compounded annually is random but once assumed a value remains fixed at that for n years, then the value of a fixed investment amount P after n years is the r.v. $g(R) = P(1+R)^n$, which is again a function of R.

First let us consider the discrete case *i.e.* let X be a discrete r.v. with p.m.f. $p_X(x)$, and suppose we are interested in a function $Y = g(X)$ of X. Note that since $\mathcal{X}$ is countable, $\mathcal{Y}$, the sample space of the r.v. Y is also necessarily so. Thus Y would also be a discrete r.v.. Now $p_Y(y)$, the p.m.f. of Y is easily figured out as follows. Fix a $y \in \mathcal{Y}$ and for finding $p_Y(y) = P[Y = y]$ look at the set $A = \{x \in \mathcal{X} : g(x) = y\}$, and then obviously $p_Y(y) = \sum_{x \in A} p_X(x)$.

Example 3.2 (Continued): Consider the r.v. X denoting the sum of the faces in a roll of two distinguishable fair die. From Figure 2, depicting its p.m.f., because of symmetry, it is clear that $\mu_X = E[X] = 7$. Now let us compute its variance. Obviously one should use (1) and the law of unconscious statistician to do so, but as an alternative let us attempt to figure this out from the definition of variance given by the formula $\sigma_X^2 = E[(X - \mu_X)^2]$. Thus in order to compute σ_X^2 from scratch, we need to figure out the distribution of the r.v. $Y = (X - \mu)^2$ and then take its expectation. Since X takes values in $\{2, 3, \ldots, 12\}$, the possible values that Y can take are 0, 1, 4, 9, 16 and 25. Among these values only $P[Y = 0] = P[X = 7] = \frac{1}{6}$. For every other value of Y, X can take two possible values and thus these probabilities need to be added in order to find $P[Y = y]$. Thus for instance $P[Y = 1] = P[\{X = 6\} \cup \{X = 8\}] = P[X = 6] + P[X = 8] = \frac{5}{18}$, $P[Y = 4] = P[\{X = 5\} \cup \{X = 9\}] = P[X = 5] + P[X = 9] = \frac{2}{9}$ etc. and continuing in this manner the p.m.f. of Y is found as

y	0	1	4	9	16	25
$p_Y(y)$	$\frac{1}{6}$	$\frac{5}{18}$	$\frac{2}{9}$	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{18}$

, which yields $\sigma_X^2 = \mu_Y = E[Y] = 0 \times \frac{1}{6} + 1 \times \frac{5}{18} + 4 \times \frac{2}{9} + 9 \times \frac{1}{6} + 16 \times \frac{1}{9} + 25 \times \frac{1}{18} = 5\frac{5}{6}$. ▽

Now let us turn our attention to the case where X has a p.d.f. $f_X(x)$. Let the c.d.f. of X be denoted by $F_X(x)$. We are interested in the distribution of $Y = g(X)$, a function of X. Unlike the discrete case, where the formula for $p_Y(y)$ can be written down for an arbitrary $g(\cdot)$, here care needs to be taken regarding the nature of $g(\cdot)$. Thus first consider the case where the function $g(\cdot)$ is one-to-one *i.e.* $g(x_1) = g(x_2) \Rightarrow x_1 = x_2$. If $g(\cdot)$ is one-to-one then it has an inverse in the sense that given a $y \in \mathcal{Y}$, the range of $g(\cdot)$, $\exists! x \in \mathcal{X} \ni g(x) = y$ and thus $g^{-1}(y) = x$. Also since $g(\cdot)$ is one-to-one, it is either strictly increasing or strictly decreasing and correspondingly so is $g^{-1}(\cdot)$. Now let the c.d.f. of Y be denoted by $F_Y(y)$.

$$\begin{aligned} & F_Y(y) \\ &= P[Y \leq y] \\ &= P[g(X) \leq y] \end{aligned}$$

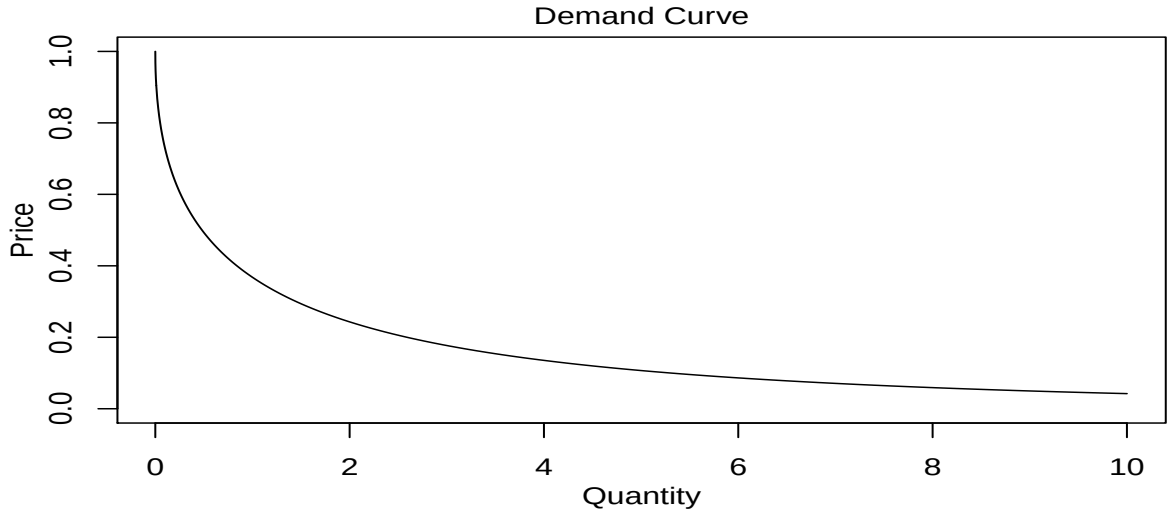
$$= \begin{cases} P\left[X \le g^{-1}(y)\right] & \text{if } g(\cdot) \text{ is increasing} \\ P\left[X \ge g^{-1}(y)\right] & \text{if } g(\cdot) \text{ is decreasing} \end{cases}$$

$$= \begin{cases} F_X\left(g^{-1}(y)\right) & \text{if } g(\cdot) \text{ is increasing} \\ 1 - F_X\left(g^{-1}(y)\right) & \text{if } g(\cdot) \text{ is decreasing} \end{cases}$$

Thus the p.d.f. of Y is given by

$$\begin{aligned} & f_Y(y) \\ &= \frac{d}{dy} F_Y(y) \\ &= \begin{cases} \frac{d}{dy} F_X\left(g^{-1}(y)\right) & \text{if } g(\cdot) \text{ is increasing} \\ \frac{d}{dy}\left[1 - F_X\left(g^{-1}(y)\right)\right] & \text{if } g(\cdot) \text{ is decreasing} \end{cases} \\ &= \begin{cases} f_X\left(g^{-1}(y)\right) \frac{d}{dy} g^{-1}(y) & \text{if } g(\cdot) \text{ is increasing} \\ -f_X\left(g^{-1}(y)\right) \frac{d}{dy} g^{-1}(y) & \text{if } g(\cdot) \text{ is decreasing} \end{cases} \\ &= \begin{cases} f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right| & \text{as for increasing } g(\cdot),\ g^{-1}(\cdot), \text{ is also increasing and hence} \\ & \frac{d}{dy} g^{-1}(y) > 0 \text{ implying } \left|\frac{d}{dy} g^{-1}(y)\right| = \frac{d}{dy} g^{-1}(y) \\ f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right| & \text{as for decreasing } g(\cdot),\ g^{-1}(\cdot), \text{ is also decreasing and hence} \\ & \frac{d}{dy} g^{-1}(y) < 0 \text{ implying } \left|\frac{d}{dy} g^{-1}(y)\right| = -\frac{d}{dy} g^{-1}(y) \end{cases} \\ &= f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right| \end{aligned} \tag{5}$$

Example 3.13: Suppose the demand curve of a product is given by the equation $P = e^{-\sqrt{Q}}$ where P denotes the Price and Q denotes the Quantity demanded as plotted in the following figure:

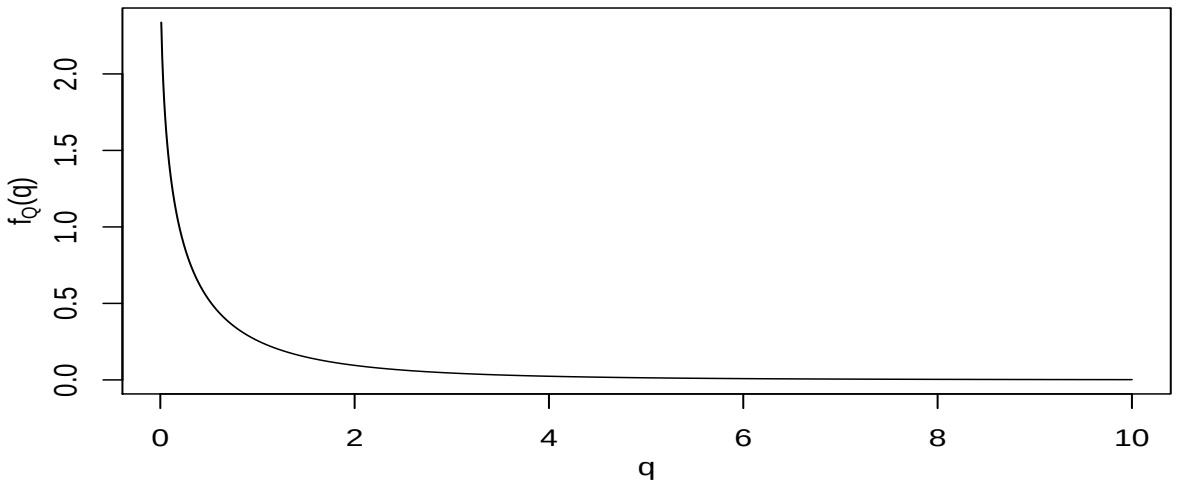


Suppose over a long horizon the Price of the product fluctuates on $[0, 1]$ according to the density given by $f_P(p) = \begin{cases} cp(1-p) & \text{if } 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}$. What is the distribution of demand over this horizon?

Solution: First note that

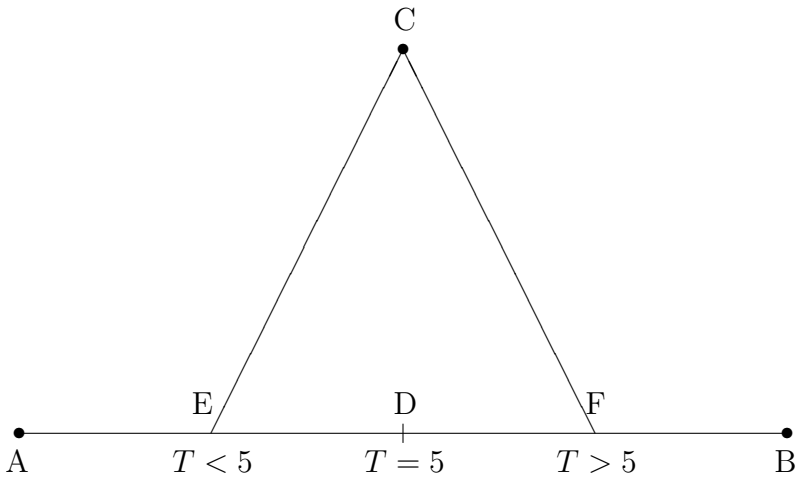
$$\begin{aligned}\frac{1}{c} &= \int_0^1 p(1-p)dp \\ &= \frac{1}{2}p^2\Big|_{p=0}^{p=1} - \frac{1}{3}p^3\Big|_{p=0}^{p=1} \\ &= \frac{1}{2} - \frac{1}{3} \\ &= \frac{1}{6}\end{aligned}$$

so that $f_P(p) = \begin{cases} 6p(1-p) & \text{if } 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}$. In order to find the p.d.f. $f_Q(q)$ of Q, expressing Q in terms of P we obtain that $Q = (-\log P)^2 = g(P)$, say. This implies that $g(P)$ is strictly decreasing and for $0 \le P \le 1$, $0 \le Q < \infty$. Now the inverse of $g(\cdot)$ has already been given as $P = g^{-1}(Q) = e^{-\sqrt{Q}}$ so that $\frac{d}{dQ}g^{-1}(Q) = -\frac{1}{2\sqrt{Q}}e^{-\sqrt{Q}}$ and therefore by (5) $f_Q(q) = \begin{cases} 3q^{-0.5}e^{-2\sqrt{q}}\left(1 - e^{-\sqrt{q}}\right) & \text{if } 0 \le q < \infty \\ 0 & \text{otherwise} \end{cases}$, the graph of which is as follows: $\bigtriangledown$



Example 3.14: While traveling eastwards at 60 km.p.h in a straight high-way from A to B, which are 10 km apart, a break-down van receives a distress call at a random point of time from a place C located 5 km north of the mid-point of A and B. Assuming that the van immediately takes a straight-line route to reach the distress point traveling in the same same speed, find the distribution and the expected amount of time it will take for the van to reach the distress point after receiving the call.

Solution: Let T denote the amount of time, in minutes, of receiving the distress call since the break-down van has left A. It will take the van 10 minutes to reach B and since the call is received at a random point of time, the p.d.f. of T is given by $f_T(t) = \begin{cases} \frac{1}{10} & \text{if } 0 \le t \le 10 \\ 0 & \text{otherwise} \end{cases}$. Now consider the location of the van and its distance from C at the time of receiving the distress call as depicted in the following diagram:



For $T < 5$ suppose the van is at point E. Then letting D denote the mid-way point on the high-way from A to B, the distance

$$\begin{aligned}
&\overline{\text{CE}} \\
&= \sqrt{\overline{\text{ED}}^2 + \overline{\text{CD}}^2} \\
&= \sqrt{(5-T)^2 + 25} \quad (\text{as } \overline{\text{ED}} = \overline{\text{AD}} - \overline{\text{AE}} = 5 - T) \\
&= \sqrt{T^2 - 10T + 50}
\end{aligned}$$

Similarly for $T > 5$ if the van is at point F,

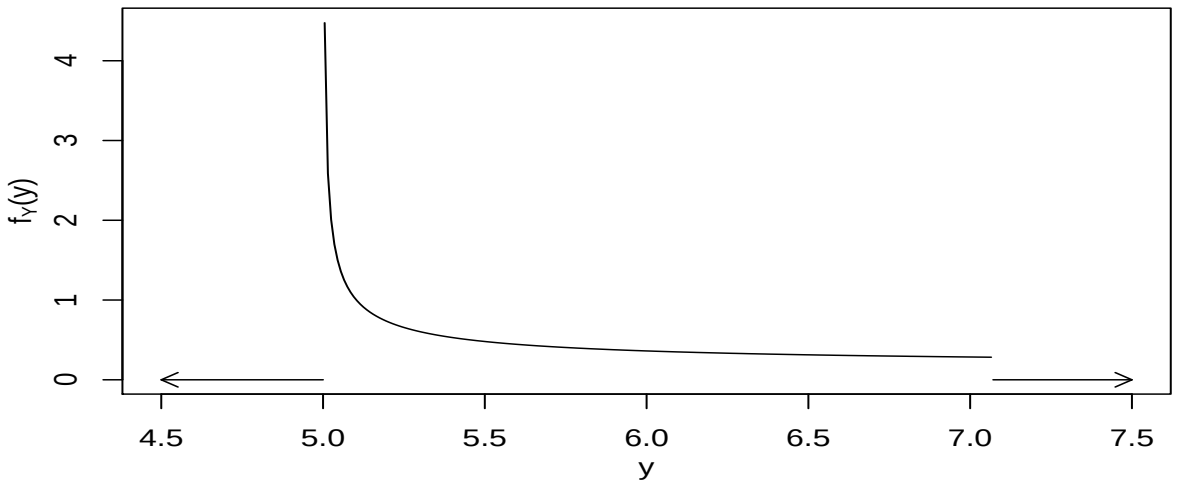
$$\begin{aligned}
&\overline{\text{CF}} \\
&= \sqrt{\overline{\text{FD}}^2 + \overline{\text{CD}}^2} \\
&= \sqrt{(T-5)^2 + 25} \quad (\text{as } \overline{\text{FD}} = \overline{\text{AF}} - \overline{\text{AD}} = T - 5) \\
&= \sqrt{T^2 - 10T + 50}
\end{aligned}$$

Thus distance of the van from C at the time of receiving the call is always $\sqrt{T^2 - 10T + 50}$ and thus the amount of time, in minutes it will take the van to reach C is $Y = \sqrt{T^2 - 10T + 50}$ $= g(T)$, say, as it is traveling at a speed of 60 km.p.h.. The maximum value that Y can take is attained at $T = 0$ and as well as at $T = 10$ and at these points $g(T) = \sqrt{50}$, while the minimum value is attained at $T = 5$ at which $g(T) = 5$. Thus the range of values that Y can assume is given by $[5, 5\sqrt{2}]$. Now note that $g(T)$ is not one-to-one on its domain $[0, 10]$ and thus we can no longer use the change of variable formula (5). However the c.d.f. route that was taken in deriving (5) may be used here for figuring out the p.d.f. of Y. Thus let $F_Y(y)$ denote the c.d.f. of Y and let $y \in [5, 5\sqrt{2}]$.

$$\begin{aligned}
&F_Y(y) \\
&= P[Y \le y] \\
&= P\left[\sqrt{T^2 - 10T + 50} \le y\right] \\
&= P\left[T^2 - 10T + 50 \le y^2\right] \qquad (\text{as } \sqrt{\cdot} \text{ is an increasing function}) \\
&= P\left[T^2 - 10T + (50 - y^2) \le 0\right]
\end{aligned}$$

$$
\begin{aligned}
&= P\left[\left(T-5-\sqrt{y^2-25}\right)\left(T-5+\sqrt{y^2-25}\right)\leq 0\right] \\
&= P\left[5-\sqrt{y^2-25}\leq T\leq 5+\sqrt{y^2-25}\right] \\
&= \frac{1}{5}\sqrt{y^2-25}
\end{aligned}
$$

Thus $f_Y(y)$, the p.d.f. of Y, is given by $\frac{d}{dy}F_Y(y)=\frac{y}{5\sqrt{y^2-25}}$ for $y\in[5,5\sqrt{2}]$, the graph of which is as follows:



Now in order to find the expected amount of time (in minutes) it would take the van to reach C, one needs to find the Expectation of Y having p.d.f. $f_Y(y)$, which requires a little bit of integration skill, but is worked out as follows:

$E[Y]$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} y f_Y(y)dy \\
&= \int_5^{5\sqrt{2}} \frac{y^2}{5\sqrt{y^2-25}}dy \\
&= 5\int_0^{\frac{\pi}{4}} \sec^3\theta d\theta \quad \text{(by substituting } y=5\sec\theta \text{ we get } \sqrt{y^2-25}=5\tan\theta,\ dy=5\sec\theta\tan\theta d\theta \\
&\qquad\qquad \text{and } y=5\Rightarrow\sec\theta=1\Rightarrow\theta=0 \ \&\ y=5\sqrt{2}\Rightarrow\sec\theta=\sqrt{2}\Rightarrow\theta=\frac{\pi}{4}) \\
&= \frac{5}{2}\left[\sec\theta\tan\theta+\log(\sec\theta+\tan\theta)\right]\Big|_{\theta=0}^{\theta=\frac{\pi}{4}} \quad \text{(this is because, integrating by parts we get,} \\
&\quad \int\sec^3\theta d\theta=\sec\theta\int\sec^2\theta d\theta-\int\left\{\frac{d}{d\theta}\sec\theta\int\sec^2\theta d\theta\right\}d\theta=\sec\theta\tan\theta-\int\sec\theta\tan^2\theta d\theta \\
&\quad =\sec\theta\tan\theta-\int\sec\theta\left(\sec^2\theta-1\right)d\theta=\sec\theta\tan\theta-\int\sec^3\theta d\theta+\int\sec\theta d\theta, \quad \text{implying} \\
&\quad \int\sec^3\theta d\theta=\frac{1}{2}\left[\sec\theta\tan\theta+\log\left(\sec\theta+\tan\theta\right)\right] \quad \text{as } \int\sec\theta d\theta=\log\left(\sec\theta+\tan\theta\right) \quad) \\
&= \frac{5}{2}\left[\sqrt{2}+\log\left(1+\sqrt{2}\right)\right] \\
&\approx 5.74 \text{ minutes}
\end{aligned}
$$

$\bigtriangledown$

The last example is an illustration of the method of handling transformations of continuous r.v. which are not one-to-one. As a second illustration consider the problem of deriving the p.d.f. of $Y = X^2$, $f_Y(y)$ say, for an arbitrary r.v. X having p.d.f. $f_X(x)$. The standard method to do this is to derive it through the c.d.f. route. Thus let $F_X(x)$ and $F_Y(y)$ respectively denote the c.d.f.'s of X and Y. Then

$$\begin{aligned} & F_Y(y) \\ &= P[Y \leq y] \\ &= P\left[X^2 \leq y\right] \\ &= P\left[-\sqrt{y} \leq X \leq \sqrt{y}\right] \\ &= F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right) \end{aligned}$$

and thus

$$\begin{aligned} & f_Y(y) \\ &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy}\left[F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right)\right] \\ &= \frac{1}{2\sqrt{y}} f_X\left(\sqrt{y}\right) + \frac{1}{2\sqrt{y}} f_X\left(-\sqrt{y}\right) \end{aligned}$$

The density of of X^2 as derived above is a special case of the situation where the transformation $Y = g(X)$ is not one-to-one. In general for many-to-one functions $g(\cdot)$, the method of deriving $f_Y(y)$, the p.d.f. of $Y = g(X)$ in terms of $f_X(x)$, the p.d.f. of X, is as follows. Suppose $\mathcal{X}$, the range of values of X, can be partitioned into finitely many disjoint regions $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_k$ (*i.e.* $\mathcal{X} = \cup_{i=1}^k \mathcal{X}_i$ and $\mathcal{X}_i \cap \mathcal{X}_j = \phi$ for $i \neq j$) in such a manner that $g : \mathcal{X}_i \to \mathcal{Y}$ *i.e.* $g(\cdot)$ restricted to $\mathcal{X}_i$, is one-to-one $\forall i = 1, 2, \ldots, k$, where $\mathcal{Y}$ is the range of values of Y. For $i = 1, 2, \ldots, k$ let $\mathcal{Y}_i = g(\mathcal{X}_i) = \{y \in \mathcal{Y} : y = g(x) \text{ for } x \in \mathcal{X}_i\}$ denote the range of $g(\cdot)$ restricted to $\mathcal{X}_i$. Note that though $\mathcal{X}_i \cap \mathcal{X}_j = \phi$ for $i \neq j$, $\mathcal{Y}_i \cap \mathcal{Y}_j$ need not be empty, and as a matter of fact since $g : \mathcal{X} \to \mathcal{Y}$ is many-to-one, some of the $\mathcal{Y}_i$'s will be necessarily overlapping. Now since $g : \mathcal{X}_i \to \mathcal{Y}_i$ is one-to-one, it has an inverse, say $g_i^{-1} : \mathcal{Y}_i \to \mathcal{X}_i$, and in terms of these k $g_i^{-1}(\cdot)$'s $f_Y(y)$ may be expressed as

$$f_Y(y) = \sum_{i=1}^{k} I[y \in \mathcal{Y}_i] f_X\left(g_i^{-1}(y)\right) \left|\frac{d}{dy} g_i^{-1}(y)\right| \tag{6}$$

where $I[y \in \mathcal{Y}_i]$ is the indicator function of the set $\mathcal{Y}_i$ *i.e.* $I[y \in \mathcal{Y}_i] = \begin{cases} 1 & \text{if } y \in \mathcal{Y}_i \\ 0 & \text{otherwise} \end{cases}$.

The intuitive idea behind (6) is a combination of the general solution in the discrete case and (5). For a $y \in \mathcal{Y}$, $f_Y(y)$, the density of Y at y is such that $f_Y(y)dy \approx P[y \leq Y \leq y + dy]$ for $dy \to 0$. For getting this probability we first look back at all those possible $x \in \mathcal{X}$ which would have yielded a $y = g(x)$ value in $[y, y + dy]$ and then add them up just as in the

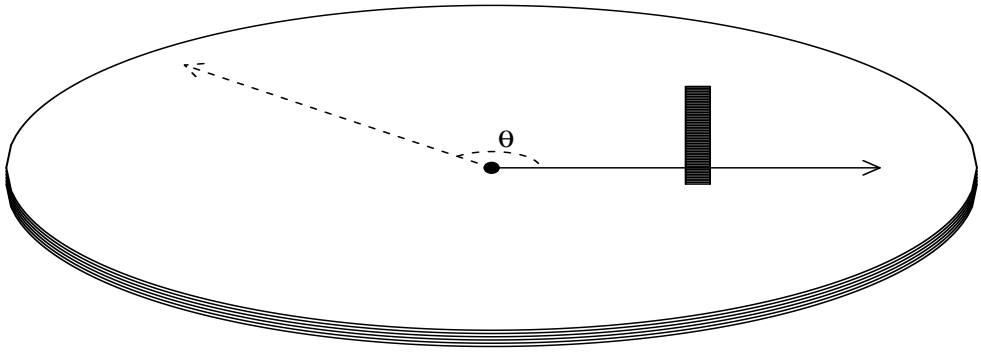
discrete case. Now for a finitely-many-to-one $g(\cdot)$, for a given y, there are only finitely many x's such that $g(x) = y$ and thus the sum mentioned in the last sentence must be a finite sum. Furthermore the $\mathcal{X}_i$'s have been chosen in such a manner that there is at most one x in each $\mathcal{X}_i$ such that $g(x) = y$, and by scanning through the k $\mathcal{X}_i$'s one can exhaust all those x's in $\mathcal{X}$ for which $g(x) = y$. This explains the sum and the $I[y \in \mathcal{Y}_i]$ part of (5). Now for an $x \in \mathcal{X}_i$ for which $g(x) = y$, its contribution towards $P[y \leq Y \leq y + dy]$ is given by $f_X\left(g_i^{-1}(y)\right)\left|\frac{d}{dy}g_i^{-1}(y)\right| dy$ according to (5). This completes the proof of (6).

For an appreciation of (6), again consider the problem of figuring out $f_Y(y)$, the p.d.f. of $Y = g(X) = X^2$. Here let $\mathcal{X} = \Re = (-\infty, \infty)$, the real line and thus $\mathcal{Y} = [0, \infty)$. Since $g(\cdot)$ is two-to-one, here $k = 2$. Let $\mathcal{X}_1 = (-\infty, 0)$ and $\mathcal{X}_2 = [0, \infty)$. Thus $\mathcal{Y} = \mathcal{Y}_1 = \mathcal{Y}_2 = [0, \infty)$. Now $g : \mathcal{X}_1 \to \mathcal{Y}_1$ has the inverse $g_1^{-1}(y) = -\sqrt{y}$ and $g : \mathcal{X}_2 \to \mathcal{Y}_2$ has the inverse $g_2^{-1}(y) = \sqrt{y}$. Thus since $I[y \in \mathcal{Y}] = 1\ \forall y \in \mathcal{Y}$, by (6),

$$f_Y(y) = f_X\left(-\sqrt{y}\right)\left|\frac{d}{dy}\left(-\sqrt{y}\right)\right| + f_X\left(\sqrt{y}\right)\left|\frac{d}{dy}\sqrt{y}\right| = \frac{1}{2\sqrt{y}}f_X\left(-\sqrt{y}\right) + \frac{1}{2\sqrt{y}}f_X\left(\sqrt{y}\right)$$

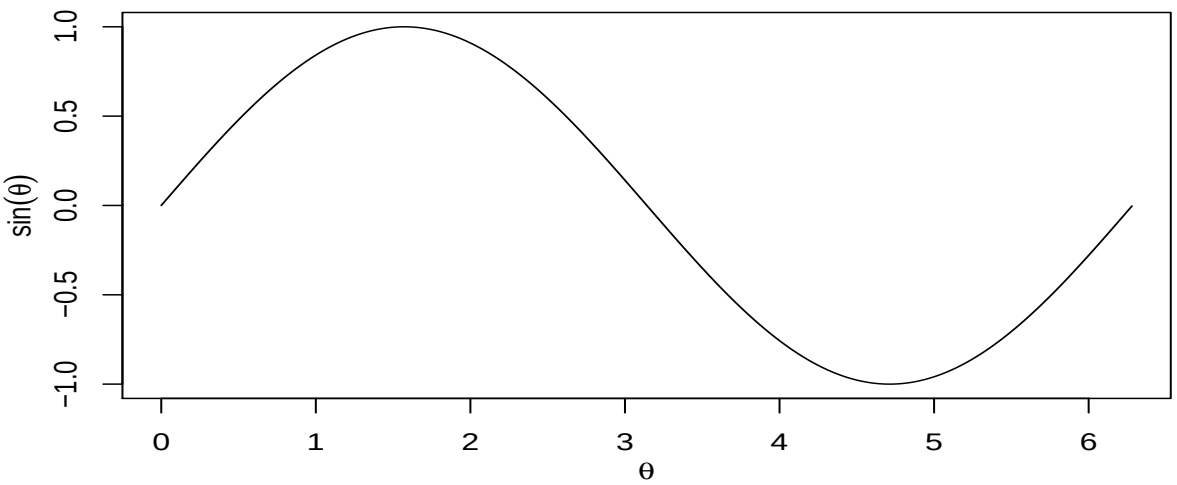
coinciding with the earlier result. We finish this section after providing an example that uses (6).

Example 3.15: A game of chance consists of a circular board with a needle fulcrumed in its center so that it can freely rotate counter-clockwise only one full circle. A stick hoisted in the middle of the board prevents the needle from more than one rotation, which also stops a clockwise rotation in the beginning of the game where the needle is kept touched to the stick as in the following diagram:



The game is as follows. A player strikes the needle using her middle finger and thumb (as for instance one hits the striker in the game of carom) and the needle rotates counter-clockwise and comes to a stop. Let θ denote the angle the needle makes in its final resting position with its initial position in the counter-clockwise direction as has been indicated with the dotted lines in the above diagram. Note that thus $0 < \theta < 2\pi$. Suppose the p.d.f. of θ is given by $f_\theta(\theta) = \begin{cases} \frac{1}{2\pi} & \text{if } 0 < \theta < 2\pi \\ 0 & \text{otherwise} \end{cases}$, and the pay-off of the game is Rs.$\sin\theta$. Find the distribution of the pay-off of the game.

Solution: We begin by examining the graph of $\sin\theta$ for $\theta \in (0, 2\pi)$ which is as follows:



Thus it is clear that in the given domain of θ, $Y = g(\theta) = \sin(\theta)$ is not one-to-one and thus we might consider using (6) for finding its p.d.f. $f_Y(y)$. We begin by disjoint partitioning of $\mathcal{X} = (0, 2\pi)$, the domain of θ, as $\mathcal{X}_1 = \left(0, \frac{\pi}{2}\right)$, $\mathcal{X}_2 = \left[\frac{\pi}{2}, \pi\right)$, $\mathcal{X}_3 = \left[\pi, \frac{3\pi}{2}\right)$ and $\mathcal{X}_4 = \left[\frac{3\pi}{2}, 2\pi\right)$ so that $g(\cdot)$ restricted to each of these $\mathcal{X}_i$'s is one-to-one $\forall i = 1, 2, 3, 4$. It is clear that $\mathcal{Y}_1 = \mathcal{Y}_2 = (0, 1]$ and $\mathcal{Y}_3 = \mathcal{Y}_4 = [-1, 0]$. $g(\cdot)$ restricted to each of the 4 $\mathcal{X}_i$'s is one-to-one with the respective inverses $g_1^{-1}(y) = \sin^{-1}(y)$, $g_2^{-1}(y) = \pi - \sin^{-1}(y)$, $g_3^{-1}(y) = \pi - \sin^{-1}(y)$ and $g_4^{-1}(y) = 2\pi + \sin^{-1}(y)$, where for $y \in [0, 1]$, $\sin^{-1}(y)$ is defined to be its principal value in $\left[0, \frac{\pi}{2}\right]$ and likewise for $y \in [-1, 0]$, $\sin^{-1}(y)$ is defined to be its principal value in $\left[-\frac{\pi}{2}, 0\right]$. Now note that for $y \in (0, 1]$, $I[y \in \mathcal{Y}_i]$ is 1 only for $i = 1, 2$ and likewise for $y \in [-1, 0)$, $I[y \in \mathcal{Y}_i]$ is 1 only for $i = 3, 4$. Thus by (6) we get that

$$
\begin{aligned}
& f_Y(y) \\
&= \begin{cases} \frac{1}{2\pi}\left[\left|\frac{d}{dy}\sin^{-1}(y)\right| + \left|\frac{d}{dy}\left(\pi - \sin^{-1}(y)\right)\right|\right] & \text{if } 0 \le y \le 1 \\ \frac{1}{2\pi}\left[\left|\frac{d}{dy}\left(\pi - \sin^{-1}(y)\right)\right| + \left|\frac{d}{dy}\left(2\pi + \sin^{-1}(y)\right)\right| +\right] & \text{if } -1 \le y < 0 \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{1}{\pi\sqrt{1-y^2}} & \text{if } -1 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

▽

3.5 Random Vectors

So far we have only talked about the situations where one measures a single quantity $X(\omega)$ on a given sample point ω. But often we shall be measuring multiple quantities, say $X_1, \ldots, X_p$, on a single ω. Since they are measurements on the same ω, though of different variables, we expect their values to be related. In this section we shall explore how such multiple (possibly related) variables may be handled. Of particular interest to us is the way one describes the distribution of values of these multiple variables simultaneously in a population.

This is because one of the major goals of applied statistical analysis is establishing and quantitatively modeling relationships between a set of variables in a population given a set of observations on these variables in a sample. This analysis would be meaningless unless we have a theoretical description of the quantities of interest summarizing this nature of relationship in the population, and the material in this section precisely covers this aspect.

For example one might measure both height and weight of an individual and might be interested in the theoretical relationship they might have between themselves in a population of individuals. For every month one might have a sales figure and the amount spent on advertising in monetary terms, and again might be interested in the theoretical relationship that might exist between these two variables in a hypothetical population of months in the past and the future. For every employee one can measure his/her motivation level for the job s/he is doing and his/her compensation level and the interest might rest in describing how these two variables jointly vary with each other in a population of employees. For a marketeer of soaps the interest might be in understanding how the preference of k available brands in the market are distributed across the different possible occupations of consumers in a population. An operations manager might be interested in the probability of a shipment reaching the client on time given the size of the shipment. We might be interested in the probability of a corporate bond defaulting its payment given its profitability ratio.

All the above examples involve two variables and in all of them we are essentially interested in seeing how the values of two variables are distributed simultaneously. For this we first begin with the formal definition of such entities.

Definition 3.10: Given a sample space Ω, a **random vector** $\boldsymbol{X}$ is a simultaneous consideration of p random variables $X_1(\omega), \ldots, X_p(\omega)$ with $\boldsymbol{X}(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_p(\omega) \end{pmatrix}$ being a $p \times 1$ vector.

Just as in §2 and §3 where the distributions of random variables were needed to be treated separately for discrete and continuous cases, here also their treatment will be different. Furthermore here there are additional complications of some of the X_i's being discrete and others being continuous leading to several different cases. To simplify matters we shall first confine ourselves to the case of $p = 2$ and discuss the way we define distributions when either both of them are discrete or both are continuous, and then taking cues from there later on generalize them for mixed cases and general p.

3.5.1 Discrete Case

Definition 3.11: A bivariate random vector $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is called **discrete** if $\boldsymbol{\mathcal{X}} \subseteq \Re^2$, the set of possible values $\boldsymbol{X}$ can take is countable.

Note that $\boldsymbol{\mathcal{X}}$ is countable if and only if the set of possible values that its two components X_1 and X_2 can take, say $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, are also countable. Before giving the formal

definition of the way the distribution of such a discrete bivariate random vector is specified, we first try to get a hang of the concept of a bivariate random vector through a couple of examples.

Example 3.16: In a toss of two distinguishable fair die, let X_1 denote the number on the first dice and X_2 denote the minimum of the two. We want to study the distribution of these two variables simultaneously. For this we first need to obtain the sample space Ω of this chance experiment and then look at the pair of values that $(X_1, X_2)'$ can take for each $\omega \in \Omega$, each one of which has a probability of $\frac{1}{36}$, in order to come up with their so called joint distribution. $\Omega = \{(1,1), \ldots, (6,6)\}$ and for each $\omega \in \Omega$, X_1 takes a value in $\{1, \ldots, 6\}$. However when $X_1(\omega)$ takes the value $x_1 \in \{1, \ldots, 6\}$, X_2, the minimum of the two cannot exceed x_1 and must take a value in $\{1, \ldots, x_1\}$. Of these, $P((X_1, X_2)' = (x_1, x_2)') = \frac{1}{36}$ whenever $x_2 < x_1$ because in that case there is a unique ω *viz.* $\omega = (x1, x_2)$ for which $(X_1, X_2)'(\omega) = (x_1, x_2)'$. However the event $[(X_1, X_2)'(\omega) = (x_1, x_1)'] = \{\omega \in \Omega : (X_1, X_2)'(\omega) = (x_1, x_1)'\} = \{(x_1, x_1), (x_1, x_1 + 1), \ldots, (x_1, 6)\}$ has $7 - x_1$ ω's in it and thus $P((X_1, X_2)' = (x_1, x_1)') = \frac{7-x_1}{36}$. These joint probabilities or the probabilities of the joint event $P(X_1 = x_1 \cap X_2 = x_2)$ are typically summarized in a tabular form as follows:

$x_2 \rightarrow$ $x_1 \downarrow$	1	2	3	4	5	6
1	$\frac{6}{36}$	0	0	0	0	0
2	$\frac{1}{36}$	$\frac{5}{36}$	0	0	0	0
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{4}{36}$	0	0	0
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{3}{36}$	0	0
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	0
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

$\triangledown$

Example 3.17: Consider the experiment of randomly distributing 3 distinguishable balls into 3 distinguishable cells. In this experiment $|\Omega| = 3^3 = 27$ with $P(\{\omega\}) = \frac{1}{27}$ $\forall \omega \in \Omega$, and for each $\omega \in \Omega$ suppose we are concerned with two variables: $X_1(\omega)$= Number of empty cells, and $X_2(\omega)$= Number of balls in cell 1. Then clearly the possible values that X_1 can take are 0, 1 and 2 and the possible values that X_2 can take are 0, 1, 2 and 3. Now we shall know everything about the joint behavior of $(X_1, X_2)'$ if we can figure out the probabilities of each of the possible $3 \times 4 = 12$ pairs of values that $(X_1, X_2)'$ can take. These joint probabilities are summarized in the following table followed by their explanations.

$x_2 \rightarrow$ $x_1 \downarrow$	0	1	2	3
0	0	$\frac{6}{27}$	0	0
1	$\frac{6}{27}$	$\frac{6}{27}$	$\frac{6}{27}$	0
2	$\frac{2}{27}$	0	0	$\frac{1}{27}$

X_1 can take the value 0 only together with $X_2 = 1$, because in this case each cell must contain exactly one ball each and there are $3! = 6$ ways of doing this. $[X_1 = 1 \cap X_2 = 0]$ can happen if and only if the empty cell is the cell 1. Then the 3 balls have to be distributed in the other 2 cells so that none of them are empty. As such 3 balls can be distributed in 2 cells in $2^3 = 8$ ways but out of that there are 2 cases where all the 3 balls go into one of the

cells leaving the other one empty. Thus $[X_1 = 1 \cap X_2 = 0]$ can happen in $8 - 2 = 6$ ways. The number of ways that $[X_1 = 1 \cap X_2 = 1]$ can happen is argued as follows. First choose one of the 3 balls in $\begin{pmatrix} 3 \\ 1 \end{pmatrix} = 3$ ways to go to cell 1, and then choose one of the remaining 2 cells in $\begin{pmatrix} 3 \\ 1 \end{pmatrix} = 2$ ways to contain the remaining 2 balls. For counting the number of ways the event $[X_1 = 1 \cap X_2 = 2]$ ca happen, again first choose 2 balls in $\begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$ ways to go to cell 1, and then choose one of the remaining 2 cells in $\begin{pmatrix} 3 \\ 1 \end{pmatrix} = 2$ ways to contain the other remaining ball. If $X_1 = 2$, then one of the cells contains all the 3 balls and the remaining 2 are empty. Thus the event $[X_1 = 2]$ can happen together with either $X_2 = 0$ or $X_2 = 3$. $[X_1 = 2 \cap X_2 = 0]$ can happen in one of the 2 cases where either one of the remaining 2 cells contain all the 3 balls, and $[X_1 = 2 \cap X_2 = 3]$ can happen for the single case of cell 1 containing all the 3 balls. $\bigtriangledown$

Above two examples illustrate the way one handles the case where both X_1 and X_2 are discrete. In this case their **joint distribution** is expressed in terms of the **joint probability mass function** or the joint p.m.f., the definition of which is as follows.

Definition 3.12: A function $p : \mathcal{X}_1 \times \mathcal{X}_2 \to \Re$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are countable sets, is called a **joint p.m.f.** if

a. $\forall x_1 \in \mathcal{X}_1$ and $\forall x_2 \in \mathcal{X}_2$, $p(x_1, x_2) \geq 0$, and

b. $\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2) = 1$

The joint p.m.f. $p(x_1, x_2)$ of a random vector $\boldsymbol{X} = (X_1, X_2)'$ gives the joint probabilities $P[X_1 = x_1 \cap X_2 = x_2]$ $\forall x_1 \in \mathcal{X}_1$ and $\forall x_2 \in \mathcal{X}_2$. The joint probability tables worked out in **Examples 16** and **17** above thus are nothing but the joint p.m.f. of the respective random vectors $(X_1, X_2)'$ in those examples. The entire distributional property of the random vector $\boldsymbol{X}$ is contained in its joint p.m.f. for the case when both of its components X_1 and X_2 are discrete. For instance given the joint p.m.f. one can easily figure out the distributions or p.m.f.'s of each of these components. Thus if $p_i(x_i)$ denotes the p.m.f. of X_i for $i = 1, 2$ they can be easily figured out from the joint p.m.f. $p(x_1, x_2)$ as follows:

$$
\begin{aligned}
& p_1(x_1) \\
&= P[X_1 = x_1] \\
&= P\left[\cup_{x_2 \in \mathcal{X}_2} \{X_1 = x_1 \cap X_2 = x_2\}\right] \quad \text{(as the event } [X_1 = x_1] = \cup_{x_2 \in \mathcal{X}_2} [X_1 = x_1 \cap X_2 = x_2]\text{)} \\
&= \sum_{x2 \in \mathcal{X}_2} P\left[X_1 = x_1 \cap X_2 = x_2\right] \quad \text{(as } [X_1 = x_1 \cap X_2 = x_2] \cap [X_1 = x1 \cap X_2 = x_2'] = \phi \text{ for } x_2 \neq x_2'\text{)} \\
&= \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)
\end{aligned}
$$

and similarly $p_2(x_2) = \sum_{x_1 \in \mathcal{X}_1} p(x_1, x_2)$. Since these distributions are found as the row and column sums of a joint p.m.f. table, they are written at its margin and are thus called **marginal distributions**.

Example 3.16 (Continued): The marginal p.m.f. of X_1 is given by $p_1(x_1) = \frac{1}{6}\ \forall x_1 \in \{1,2,3,4,5,6\}$, as it should be, because X_1 simply denotes the outcome of rolling the first dice, and it agrees with the row totals of the joint p.m.f. table. The marginal p.m.f. of X_2 is found from the column sums yielding its marginal p.m.f. as

x_2	1	2	3	4	5	6
$p_2(x_2)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

$\bigtriangledown$

Example 3.17 (Continued): The marginal p.m.f. of X_1 and X_2 are found respectively as row and column totals as follows:

$x_2 \rightarrow$ $x_1 \downarrow$	0	1	2	3	$p_1(x_1)$
0	0	$\frac{6}{27}$	0	0	$\frac{6}{27}$
1	$\frac{6}{27}$	$\frac{6}{27}$	$\frac{6}{27}$	0	$\frac{18}{27}$
2	$\frac{2}{27}$	0	0	$\frac{1}{27}$	$\frac{3}{27}$
$p_2(x_2)$	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$	1

$\bigtriangledown$

One of the major reasons for studying the joint distribution or the distribution of both the components simultaneously is that it enables one to study the dependence structure, if any, between the two components X_1 and X_2 of a random vector $\boldsymbol{X}$. The joint distribution contains this association structure between X_1 and X_2 in sort of an implicit form, which is explicitly brought out by introducing the notion of **conditional distributions.** There are two sets of conditional distributions: conditional distributions of $X_1|X_2$ and the conditional distributions of $X_2|X_1$. Note the plural in "sets of conditional distributions". This is because for every $x_2 \in \mathcal{X}_2$ there is a conditional distribution of $X_1|X_2 = x_2$ and thus there are $|\mathcal{X}_2|$ many conditional distributions of $X_1|X_2$ and similarly there are $|\mathcal{X}_1|$ many conditional distributions of $X_2|X_1$. Thus fix an $x_2 \in \mathcal{X}_2$ and let us see what we mean by the conditional distribution of $X_1|X_2 = x_2$. For $x_1 \in \mathcal{X}_1$, the **conditional p.m.f.** of $X_1|X_2 = x_2$ gives $P[X_1 = x_1|X_2 = x_2]$, which is denoted by $p_{1|2}(x_1|x_2)$, and is as follows:

$$\begin{aligned} & p_{1|2}(x_1|x_2) \\ &= P[X_1 = x_1|X_2 = x_2] \\ &= \frac{P[X_1 = x_1 \cap X_2 = x_2]}{P[X_2 = x_2]} \\ &= \frac{p(x_1, x_2)}{p_2(x_2)}. \end{aligned}$$

Similarly for a fixed $x_1 \in \mathcal{X}_1$, for $x_2 \in \mathcal{X}_2$, the conditional p.m.f. of $X_2|X_1 = x_1$, denoted by $p_{2|1}(x_2|x_1)$, is given by $p_{2|1}(x_2|x_1) = p(x_1, x_2)/p_1(x_1)$. Thus the conditional p.m.f. is just the ratio of the joint p.d.f. to the marginal p.d.f. of the conditioning variable.

Example 3.16 (Continued): There are 6 conditional p.m.f's $p_{1|2}(x_1|x_2)$ and likewise there are 6 $p_{2|1}(x_2|x_1)$'s. With the marginals $p_2(x_2)$ and $p_1(x_1)$ already derived above, a straightforward division yields

$p_{1\|2}(x_1 \downarrow \|x_2 \rightarrow)$	1	2	3	4	5	6
1	6/11	0	0	0	0	0
2	1/11	5/9	0	0	0	0
3	1/11	1/9	4/7	0	0	0
4	1/11	1/9	1/7	3/5	0	0
5	1/11	1/9	1/7	1/5	2/3	0
6	1/11	1/9	1/7	1/5	1/3	1

$p_{2\|1}(x_2 \rightarrow \|x_1 \downarrow)$	1	2	3	4	5	6
1	1	0	0	0	0	0
2	1/6	5/6	0	0	0	0
3	1/6	1/6	4/6	0	0	0
4	1/6	1/6	1/6	3/6	0	0
5	1/6	1/6	1/6	1/6	2/6	0
6	1/6	1/6	1/6	1/6	1/6	1/6

$\bigtriangledown$

Example 3.17 (Continued): Similarly straight forward division gives the 4 $p_{1|2}(x_1|x_2)$'s and 3 $p_{2|1}(x_2|x_1)$'s as:

$p_{1\|2}(x_1 \downarrow \|x_2 \rightarrow)$	0	1	2	3
0	0	1/2	0	0
1	3/4	1/2	1	0
2	1/4	0	0	1

$p_{2\|1}(x_2 \rightarrow \|x_1 \downarrow)$	0	1	2	3
0	0	1	0	0
1	1/3	1/3	1/3	0
2	2/3	0	0	1/3

$\bigtriangledown$

With the conditional distributions defined as above we are now in a position to define independence of two discrete r.v.'s. Recall that two events A and B were defined to be independent if $P(A|B) = P(A)$. We have a similar definition in case of random variables.

Definition 3.13: Two discrete random variables X_1 and X_2 are said to be **statistically** or **stochastically independent** if, $p_{1|2}(x_1|x_2)$, the conditional p.m.f. of $X_1|X_2 = x_2$ does not depend on x_2.

At this juncture it would be apt to point out that while doing probability/statistics, independence always means statistical/stochastic independence, and thus we shall drop the qualifying adverb statistical/stochastic. We shall encounter situations where independence would mean physical or logical independence. We shall be careful in those situations to distinguish the two but as shall be seen, these two notions of independence will be used interchangeably but mathematically independence will always be treated as it has been defined in ***Definition 13*** or its equivalent forms.

While ***Definition 3.13*** is intuitively very appealing, the same *viz.* independence of two discrete X_1 and X_2, may also be expressed in various alternative but equivalent ways, many a times which are much easier to check or apply (as the case may be - "checking" for independence or "applying" after assuming physical or logical independence). The first question is, if $p_{1|2}(x_1|x_2)$ does not depend on x_2, then what do these $|\mathcal{X}_2|$ many conditional

distributions equal to across all the $x_2 \in \mathcal{X}_2$? Since $p_{1|2}(x_1|x_2)$ does not depend on x_2, let $p_{1|2}(x_1|x_2) = q(x_1)$, say. Then

$$
\begin{aligned}
& p_{1|2}(x_1|x_2) = q(x_1) \ \ \forall x_1 \in \mathcal{X}_1 \ \& x_2 \in \mathcal{X}_2 \\
\Rightarrow \ & \frac{p(x_1, x_2)}{p_2(x_2)} = q(x_1) \ \ \forall x_1 \in \mathcal{X}_1 \ \& x_2 \in \mathcal{X}_2 \ \text{(by definition of the conditional p.m.f. } p_{1|2}(x_1|x_2)) \\
\Rightarrow \ & p(x_1, x_2) = p_2(x_2) q(x_1) \ \ \forall x_1 \in \mathcal{X}_1 \ \& x_2 \in \mathcal{X}_2 \\
\Rightarrow \ & \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2) = q(x_1) \sum_{x_2 \in \mathcal{X}_2} p_2(x_2) \ \ \forall x_1 \in \mathcal{X}_1 \\
\Rightarrow \ & p_1(x_1) = q(x_1) \ \ \forall x_1 \in \mathcal{X}_1 \ \text{(since } p_1(x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2) \text{ and } \sum_{x_2 \in \mathcal{X}_2} p_2(x_2) = 1 \text{ as it is the} \\
& \qquad \text{marginal p.m.f. of } X_2)
\end{aligned}
$$

This shows that if X_1 and X_2 are independent, then the all the $|\mathcal{X}_2|$ many conditional p.m.f.'s $p_{1|2}(x_1|x_2)$ of $X_1|X_2$ coincide with $p_1(x_1)$, the marginal p.m.f. of X_1. Conversely, if all the $|\mathcal{X}_2|$ many conditional p.m.f.'s of $X_1|X_2$ coincide with the marginal p.m.f. of X_1, then obviously $p_{1|2}(x_1|x_2)$ does not depend on x_2, and thus by definition, X_1 and X_2 must be independent. Therefore we see that independence is equivalent to the condition of all the conditional p.m.f.'s of $X_1|X_2$ being identical to the marginal p.m.f. of X_1. Now if that is the case,

$$p_{1|2}(x_1|x_2) = p_1(x_1) \Leftrightarrow \frac{p(x_1, x_2)}{p_2(x_2)} = p_1(x_1) \Leftrightarrow p(x_1, x_2) = p_1(x_1) p_2(x_2)$$

showing that independence of two discrete r.v. is equivalent to the condition of the joint p.m.f. being the product of the two marginal p.m.f.'s. This result is analogous to the corresponding result for the independence of two events A and B, which states that two events are independent if and only if $P(A \cap B) = P(A)P(B)$. Just as in the case of independence of events, though we had defined independence as $P(A|B) = P(A)$, later after establishing the equivalence of independence with $P(A \cap B) = P(A)P(B)$, we observed that this implies $P(B|A) = P(B)$, as it should do logically; here also the same remark holds true. Thus independence of two discrete r.v. X_1 and X_2 is equivalent to the condition that $p_{2|1}(x_2|x_1)$, the conditional p.m.f. of $X_2|X_1 = x_1$, does not depend on x_1 and all these $|\mathcal{X}_1|$ many conditional p.m.f. coincide with $p_2(x_2)$, the marginal p.m.f. of X_2.

X_1 and X_2 of **Examples 3.16** and **3.17** are not independent. This is because in **Example 16** for instance, $p(3,5) = P(X_1 = 3 \& X_2 = 5) = 0 \neq \frac{1}{6} \cdot \frac{1}{12} = P(X_1 = 3)P(X_2 = 5) = p_1(3)p_2(5)$. Likewise in **Example 17** $p(1,3) = P(X_1 = 1 \& X_2 = 3) = 0 \neq \frac{2}{3} \cdot \frac{1}{27} = P(X_1 = 1)P(X_2 = 3) = p_1(1)p_2(3)$. Thus if the equality $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ is violated even for one pair (x_1, x_2) then X_1 and X_2 are not independent. This is because for X_1 and X_2 to be independent, $p(x_1, x_2)$ must equal $p_1(x_1)p_2(x_2)$ $\forall x_1 \in \mathcal{X}_1$ and $\forall x_2 \in \mathcal{X}_2$.

Example 3.18: Consider the experiment of randomly permuting the 4 letters a, b, c and d. For this experiment $|\Omega| = 4! = 24$. For a given $\omega \in \Omega$ define $X_1(\omega) = \begin{cases} 1 & \text{if } a \text{ precedes } b \text{ in } \omega \\ 0 & \text{otherwise} \end{cases}$ and $X_2(\omega) = \begin{cases} 1 & \text{if } c \text{ precedes } d \text{ in } \omega \\ 0 & \text{otherwise} \end{cases}$. Then the event $[X_1 = 1 \& X_2 = 1]$ can happen in

6 ways. This is because for $X_1 = 1$ to happen, there are 3 positions where a can be placed *viz.* 1, 2 and 3. If a is at position 1, then b can be placed in either positions 2, 3, or 4, and in each of these cases the positions of c and d get fixed (b at 2 forces c to be at 3 and d to be at 4, b at 3 forces c to be at 2 and d to be at 4, and b at 4 forces c to be at 2 and d to be at 3) leading to 3 possibilities. If a is at 2, b can be at either 3 or 4, and again the positions of c and d get automatically fixed (b at 3 forces c at 1 and d at 4, and b at 4 forces c at 1 and d at 3), this time leading to 2 possibilities. Finally if a is at 3, then b has to be at 4, c has to be at 1 and d has to be at 2, leading to the single possibility. This reasoning yields that $\{\omega : X_1(\omega) = 1 \& X_2(\omega) = 1\} = \{abcd, acbd, acdb, cabd, cadb, cdab\}$ and thus $P[X_1 = 1 \& X_2 = 1] = \frac{6}{24} = \frac{1}{4}$. In the preceding argument by interchanging the roles of c and d, a and b, and both a and b and c and d we respectively find $P[X_1 = 1 \& X_2 = 0]$, $P[X_1 = 0 \& X_2 = 1]$ and $P[X_1 = 0 \& X_2 = 0]$ all equal $\frac{1}{4}$. Thus the joint p.m.f. of $(X_1, X_2)'$ and the two marginals p.m.f.'s are given by

$x_2 \rightarrow$ $x_1 \downarrow$	0	1	$p_1(x_1)$
0	1/4	1/4	1/2
1	1/4	1/4	1/2
$p_2(x_2)$	1/2	1/2	1

From the above table it is clear that $\forall\ x_1 \in \mathcal{X}_1 = \{0, 1\}$ and $x_2 \in \mathcal{X}_2 = \{0, 1\}$, $p(x_1, x_2) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = p_1(x_1)p_2(x_2)$. Therefore in this example X_1 and X_2 are independent. $\bigtriangledown$

Just as in case of events, here also independence is used both ways. That is, there are situations where given a joint p.m.f. one would need to check whether X_1 and X_2 are independent as in **Example 18**. Conversely there will be situations where at the outset independence would be assumed from the physical or logical structure of the problem, which is then exploited for figuring out the joint p.m.f.. The most common application of this later type is that of so-called "independent trials" which we have already been using without any explicit reference to the joint distribution. An example should help clarify the point.

Example 3.19: Suppose we have a biased dice with the following p.m.f. for the outcome X of rolling it once:

x	1	2	3	4	5	6
$p_X(x)$	0.1	0.2	0.2	0.2	0.2	0.1

Now suppose the dice is rolled twice and let X_i denote the outcome of the i-th roll, $i = 1, 2$. We are interested in the distribution of the sum $Y = X_1 + X_2$ just as in **Example 2**. However in **Example 2** since the dice was assumed to be unbiased, we did not have any problem in assigning equal probabilities to the 36 possible ω's of this experiment, and did not really have to go through the joint distribution route. Here however we have to first determine the joint distribution of $(X_1, X_2)'$, which in turn will be used to figure out the distribution of the sum Y. Towards this end, since the two rolls are physically independent it may be quite reasonable to assume that X_1 and X_2 are also statistically independent[7]. Furthermore the marginal p.m.f. of both X_1 and X_2 are same as $p_X(\cdot)$. Thus for $x_1 \in \mathcal{X}_1 = \{1, 2, 3, 4, 5, 6\}$

[7]Note the care in bothering to state the nature of independence *viz.* physical vis-a-vis stochastic, as mentioned in the paragraph following ***Definition 13***.

and $x_2 \in \mathcal{X}_2 = \{1,2,3,4,5,6\}$, the joint p.m.f. $p(x_1,x_2) = p_1(x_1)p_2(x_2) = p_X(x_1)p_X(x_2)$, which may be expressed in a tabular form as follows:

$x_2 \rightarrow$ $x_1 \downarrow$	1	2	3	4	5	6
1	0.01	0.02	0.02	0.02	0.02	0.01
2	0.02	0.04	0.04	0.04	0.04	0.02
3	0.02	0.04	0.04	0.04	0.04	0.02
4	0.02	0.04	0.04	0.04	0.04	0.02
5	0.02	0.04	0.04	0.04	0.04	0.02
6	0.01	0.02	0.02	0.02	0.02	0.01

Now for the p.m.f. of $Y = X_1 + X_2$ we have to look at the value of Y for each of the 36 possibilities of $(x_1,x_2) \in \{1,2,3,4,5,6\}^2$ and then add the $p(X_1,x_2)$ values for those (x_1,x_2)'s yielding the same value of Y. This gives the p.m.f. of Y as

y	2	3	4	5	6	7	8	9	10	11	12
$p_Y(y)$	0.01	0.04	0.08	0.12	0.16	0.18	0.16	0.12	0.08	0.04	0.01

$\bigtriangledown$

While conditional distributions depict the association between X_1 and X_2 quite accurately, the information content in them is very large. This leads us to seek some key summary measures which can capture this degree of association between two r.v.. This is analogous to defining mean as a measure of location of a distribution despite having the entire distribution at one's disposal. We start by defining a measure of association called **covariance**.

Definition 3.14: **Covariance** of two random variables X_1 and X_2 is given by $E\left[(X_1 - E[X_1])(X_2 - E[X_2])\right]$ and is denoted by $\text{Cov}(X_1,X_2)$.

In order to understand the motivation behind this definition, let us study the sign of the product $(X_1 - E[X_1])(X_2 - E[X_2])$. If X_2 tends to take higher (lower) values for high (low) values of X_1, then that implies that if $X_1 - E[X_1] > 0$ ($X_1 - E[X_1] < 0$), with a high probability $X_2 - E[X_2]$ will also tend to be positive (negative). In such a situation, X_1 and X_2 have an increasing relationship and $(X_1 - E[X_1])(X_2 - E[X_2])$ will have a positive sign with high probability leading to a positive $\text{Cov}(X_1,X_2)$. On the other hand if X_1 and X_2 have a decreasing relationship *i.e.* X_2 tends to take lower (higher) values for higher (lower) values of X_1, then $X_1 - E[X_1] > 0$ ($X_1 - E[X_1] < 0$) $\Rightarrow X_2 - E[X_2] < 0$ ($X_2 - E[X_2] > 0$) with a high probability, making $\text{Cov}(X_1,X_2)$ negative.

Though the sign of the covariance in a nut-shell depicts whether X_1 and X_2 have an increasing or decreasing relationship, caution must be exercised in this interpretation of covariance. If the relationship is highly curvy-linear, an increasing or decreasing relationship is meaningless and in such situations, so is covariance. Thus covariance as a measure of degree of association only makes sense when the relationship between X_1 and X_2 is approximately linear so that one can assign a clear-cut meaning to a relationship being increasing or decreasing. Therefore the sign of covariance is best interpreted as the *sign of linear association* between two random variables X_1 and X_2.

In the last sentence, instead of "sign", it would have been nice if we could have said that covariance is a measure of "degree" of linear association. But unfortunately we cannot do that because the raw numerical value of the covariance is mired with other incidental nuisance factors. To see this first observe that if either X_1 or X_2 has a large variability that will inflate the covariance when this variability factor should have nothing to do with a measure of association between two variables. That is for example since[8] $\text{Cov}(cX_1,X_2)=c\text{Cov}(X_1,X_2)$, though the degree of linear association between cX_1 and X_2 is same as that between X_1 and X_2, for $c > 1$ the former pair would have a larger covariance, rendering just the raw numerical value of covariance (without any further adjustment) rather useless as a measure of degree of linear association. Second, a measure of degree of association, linear or otherwise, should be a pure number that does not depend on the unit of measurement. For appreciating this point, consider the problem of measuring association between heights and weights of individuals. The degree of association is that between the concept of height and weight, and not between say foot-pound or meter-kilogram *i.e.* the degree of association should not depend on whether the height is measured in foot or meter or the weight is measured in pound or kilogram. However the raw covariance value in the foot-pound case is different from the meter-kilogram one, again rendering just the raw covariance value (without any further adjustment) useless as a measure of degree of association.

Once we recognize the problems, the solution now becomes rather obvious. We can retain covariance as the basic measure, provided it is appropriately scaled making it unit free and insensitive to the variabilities in X_1 and X_2. But formula-wise it will be a different quantity requiring a separate name. This quantity is called the **correlation coefficient** as defined below.

Definition 3.15: **Correlation coefficient** of two random variables X_1 and X_2 is given by $\frac{\text{Cov}(X_1,X_2)}{\sqrt{V[X_1]V[X_2]}}$ and is denoted by ρ_{X_1,X_2} or simply ρ when dealing with just two variables.

The correlation coefficient ρ is a pure number, free of the units as well as the variability of the original variables. Also note that it is essentially a covariance based measure of association but is free of the pitfalls of just the raw covariance. To see that it is basically a covariance, but being measured in a standardized scale free of unit and variability, for $i = 1, 2$ define $Z_i = \frac{X_i-\mu_i}{\sigma_i}$ where μ_i and σ_i respectively are the mean and standard deviation of X_i. Now since $E[Z_i] = 0$ and $V[Z_i] = 1$,

$$\begin{aligned}
&\rho_{Z_1,Z_2} \\
&= \text{Cov}(Z_1, Z_2) \\
&= E[Z_1Z_2] \\
&= E\left[\frac{X_1-\mu_1}{\sigma_1}\frac{X_2-\mu_2}{\sigma_2}\right] \\
&= \frac{E[(X_1-E[X_1])(X_2-E[X_2])]}{\sqrt{V[X_1]V[X_2]}}
\end{aligned}$$

[8]See Appendix B, where the properties of moments, including that of covariance, have been assembled together.

$$= \frac{\text{Cov}(X_1, X_2)}{\sqrt{V[X_1]V[X_2]}}$$
$$= \rho_{X_1,X_2}$$

Thus the correlation coefficient between X_1 and X_2 is nothing but the covariance between Z_1 and Z_2, where Z_1 and Z_2 are such that they are unit free (since $X_i - \mu_i$ and σ_i have the same unit) with unit variance.

Since correlation coefficient is a covariance based measure and the kind of association that is meaningfully captured by the covariance is linear, now we can say that the correlation coefficient may be interpreted as the *degree of linear association*. Since the sign of correlation coefficient is inherited from the covariance, its sign indicates whether the relationship is increasing or decreasing, and furthermore unlike the raw covariance, its numerical value is very nicely interpretable as it has been shown in Appendix B that $-1 \leq \rho \leq 1$, with equality (± 1) if and only if $X_2 = a + bX_1$ for some constants a and b *i.e.* when there is an exact linear relationship between X_1 and X_2. Larger the absolute value of ρ more linear is the association between them. The exact interpretation of the numerical value of ρ will be deferred till the Applied Statistics chapters. At this point it is advisable to read Appendix B assimilating the properties of covariance and ρ before proceeding any further.

If the nature of association between X_1 and X_2 is linear then the single number correlation coefficient is a useful quantity for measuring this degree of association. But in general how can one summarize the association between X_1 and X_2? If the relationship is non-linear, on the surface it appears that a single number may not be able to do the job[9], and instead we seek a function that would be useful in depicting the relationship. Ultimately the entire story of the nature of relationship is contained in the conditional distributions and here we are seeking how to summarize this vast information. The most common and intuitively appealing way of summarizing a distribution is to quote its mean or expected value. When we do the same with the conditional distributions it is called **regression**, which is formally defined as follows.

Definition 3.16: The function $g(x_2) = E[X_1|X_2 = x_2] = \sum_{x_1 \in \mathcal{X}_1} x_1 p_{1|2}(x_1|x_2)$ *i.e.* the conditional mean of $X_1|X_2$ is called the **regression of X_1 on X_2**.

First note that the mean is being computed using the conditional distribution, and hence it is the conditional mean. Second point to note is that the conditional distribution of $X_1|X_2$ and thus the mean as well, depends on the conditioning value x_2 of X_2, and thus $E[X_1|X_2 = x_2]$ is a function of x_2 and as such $E[X_1|X_2]$ is a r.v. since X_2 is. As stated earlier the best thing to do for depicting the association would be carrying around all the $|\mathcal{X}_2|$ many conditional distributions of $X_1|X_2$. But since this information may be overwhelming, the next best thing to do would be at least carrying the simplest summary measure *viz.* the mean, of all these conditional distributions. Since there are $|\mathcal{X}_2|$ many of these conditional means $E[X_1|X_2]$, one for each $x_2 \in \mathcal{X}_2$, their totality can be viewed as a function of x_2, and is named regression of X_1 on X_2.

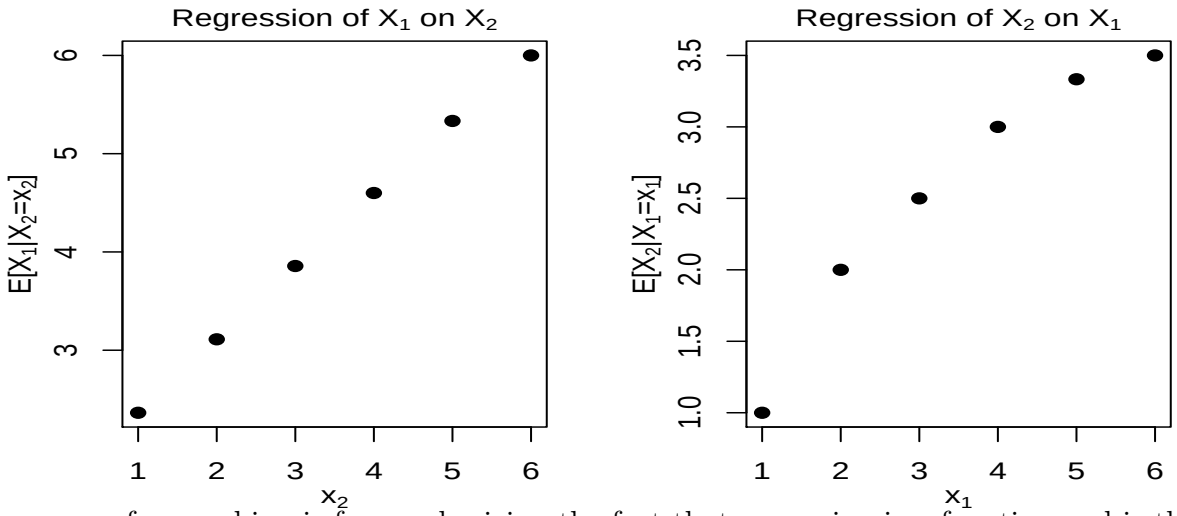
Example 3.16 (Continued): Based on the conditional distributions, the two regressions,

[9]Strictly speaking this is not true, as shall be seen later in the Applied Statistics chapters.

X_1 on X_2 and X_2 on X_1 is given as follows:

x_1/x_2	1	2	3	4	5	6
$E[X_1\|X_2]$	26/11	28/9	27/7	23/5	16/3	6
$E[X_2\|X_1]$	1	2	5/2	3	10/3	7/2

which are also plotted in the following two graphs:



The reason for graphing is for emphasizing the fact that regression is a function and is thus best interpreted in terms of its graphs. From these two graphs it also appears that the regression of X_1 on X_2 is approximately linear while that of X_2 on X_1 is a curvy-linear concave function of x_1. However on the domain of x_1, though curvy-linear, $E[X_2|X_1 = x_1]$ is an increasing function of x_1, and thus computation of the correlation would make some sense here and we shall expect it to be positive.

In order to find ρ, the first thing we need is $\mathrm{Cov}(X_1, X_2)$. For this we shall use the short-cut formula $\mathrm{Cov}(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2]$ given in Appendix B.

$$\begin{aligned} & E[X_1X_2] \\ &= \sum_{x_1}\sum_{x_2} x_1x_2p(x_1,x_2) \\ &= \frac{1}{36}\left[\{6+(2+3+4+5+6)\}+\{4\times5+2(3+4+5+6)\}+\{9\times4+3(4+5+6)\}\right. \\ &\qquad \left.+\{16\times3+4(5+6)\}+\{25\times2+30\}+36\right] \\ &= \frac{371}{36} \end{aligned}$$

Based on the marginal p.m.f. of X_1, we find that $E[X_1] = 7/2$ and $E[X_1^2] = 91/6$; and from $p_2(x_2)$, we find that $E[X_2] = 91/36$ and $E[X_2^2] = 301/36$. Thus we get that

$$\mathrm{Cov}(X_1,X_2) = \frac{371}{36} - \frac{7}{2}\cdot\frac{91}{36} = \frac{105}{72},\ \ V[X_1] = \frac{91}{6} - \frac{49}{4} = \frac{35}{12},\ \text{ and }\ V[X_2] = \frac{301}{36} - \frac{91^2}{36^2} = \frac{2555}{36^2},$$

$$\text{so that } \rho = \frac{\frac{105}{72}\times 36}{\sqrt{\frac{35\times2555}{12}}} \approx 0.6082 \qquad \bigtriangledown$$

Example 3.17 (Continued): Here let us first find the correlation coefficient ρ.

$$\begin{aligned} & E[X_1X_2] \\ &= \sum_{x_1}\sum_{x_2} x_1x_2p(x_1,x_2) \\ &= \frac{1}{27}[6+12+6] \\ &= \frac{8}{9} \end{aligned}$$

$$E[X_1] = \frac{8}{9}, \;\; E[X_1^2] = \frac{10}{9} \qquad E[X_2] = 1, \;\; \text{and} \;\; E[X_2^2] = \frac{5}{3}$$

Thus since $\text{Cov}(X_1, X_2) = \frac{8}{9} - 1 \cdot \frac{8}{9} = 0$, $\rho = 0$. However note that we have already established that in this example X_1 and X_2 are *not independent*. Thus here is an example of two random variables X_1 and X_2 which are not independent despite their correlation coefficient (covariance) being 0. (See ***Property C5*** of Appendix B.)

Based on the conditional distributions $p_{1|2}(x_1|x_2)$ and $p_{2|1}(x_2|x_1)$ we find the regression of X_1 on X_2 and X_2 on X_1 as follows:

x_2	0	1	2	3
$E[X_1\|X_2]$	5/4	1/2	1	2

x_1	0	1	2
$E[X_2\|X_1]$	1	1	1

Note that based on the two regression functions it may be concluded that there is no obvious relationship between X_1 and X_2. While the regression of X_1 on X_2 shows an initially decreasing and then increasing relationship, that of X_2 on X_1 is flat. Thus though X_1 and X_2 are not independent, since their relationship is not even approximately linear, correlation coefficient, which is 0 anyway, is not very useful in summarizing their relationship. $\bigtriangledown$

We finish this sub-section after working out a couple of problems involving discrete joint distribution.

Example 3.20: The joint p.m.f. $p(x, y)$ of Y, the number of completed projects in a given year, and age X, of engineers, working for a software firm is as follows:

$y \rightarrow$ $x \downarrow$	0	1	2
21	0.05	0.03	0.02
22	0.08	0.22	0.10
23	0.05	0.18	0.07
24	0.06	0.12	0.02

Answer the following:

a. What is the probability that an engineer has finished at least one project?

b. What is the probability that an engineer is 22 years or younger?

c. What is the probability that an engineer is 22 years or younger and has finished at least one project?

d. What is the probability that an engineer who is 22 years or younger has finished at least one project?

e. What is the most likely number of completed projects by engineers who are 22 years or younger?

f. What is the probability that an engineer who has finished at least one project is 22 years or younger?

g. What is the average Age of the engineers finishing at least one project?

h. Give the marginal c.d.f. of Y.

i. Are X and Y independent?

j. Find the correlation coefficient between X and Y.

k. Find the regression of Y on X and use it to determine the most productive age of the engineers.

Solution (a): $P(Y \geq 1) = \sum_{x=21}^{24} \sum_{y=1}^{2} P(X = x, Y = y) = 1 - \sum_{x=21}^{24} P(X = x, Y = 0) = 1 - 0.24 = 0.76$.
(b): $P(X \leq 22) = \sum_{x=21}^{22} \sum_{y=0}^{2} P(X = x, Y = y) = 0.5$.
(c): $P(X \leq 22 \cap Y \geq 1) = \sum_{x=21}^{22} \sum_{y=1}^{2} P(X = x, Y = y) = 0.03 + 0.02 + 0.22 + 0.10 = 0.37$.
(d): $P(Y \geq 1 | X \leq 22) = \frac{P(X \leq 22 \cap Y \geq 1)}{P(X \leq 22)} = \frac{0.37}{0.5} = 0.74$.
(e): For answering this question, we need to look at the conditional distribution of $Y|X \leq 22$. That is we need to find $P(Y = y|X \leq 22)$ for $y = 0, 1, 2$. $P(Y = 0|X \leq 22) = \frac{P(Y=0 \cap X \leq 22)}{P(X \leq 22)} = \frac{0.05+0.08}{0.5} = 0.26$. After calculating $P(Y = 1|X \leq 22)$ and $P(Y = 2|X \leq 22)$ in a similar fashion we find

y	0	1	2
$P(Y = y\|X \leq 22)$	0.26	0.50	0.24

as the conditional p.m.f. of $Y|X \leq 22$, form which it is now obvious that the most likely number of projects completed by engineers who are 22 years or younger is 1.
(f): $P(X \leq 22|Y \geq 1) = \frac{P(X \leq 22 \cap Y \geq 1)}{P(Y \geq 1)} = \frac{0.37}{0.76} \approx 0.4868$.
(g): For answering this, like **e**, we need to find the conditional distribution of $X|Y \geq 1$. $P(X = 21|Y \geq 1) = \frac{(X=21 \cap Y \geq 1)}{P(Y \geq 1)} = \frac{0.03+0.02}{0.76} \approx 0.0658$. Proceeding in a similar manner we find that the conditional p.m.f. of $X|Y \geq 1$ is given by

x	21	22	23	24
$P(X = x\|Y \geq 1)$	0.0658	0.4211	0.3289	0.1842

which yields $E[X|Y \geq 1] = 21 \times 0.0658 + 22 \times 0.4211 + 23 \times 0.3289 + 24 \times 0.1842 = 22.6315$.
(h): $p_Y(y)$, the marginal p.m.f. of Y, is found as the column sums of the joint p.m.f. table, which is given by

y	0	1	2
$p_Y(y)$	0.24	0.55	0.21

and thus the marginal c.d.f. of Y is given by $F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 0.24 & \text{if } 0 \leq y < 1 \\ 0.79 & \text{if } 1 \leq y < 2 \\ 1 & \text{if } y \geq 2 \end{cases}$.
(i): $P(X = 21, Y = 0) = 0.05 \neq 0.1 \times 0.24 = P(X = 21)P(Y = 0)$ for instance, and thus X and Y are not independent.

(j): By property ***Property R3*** of correlation coefficient since ρ is invariant under linear transformation, to facilitate computation we shall work with $Z = X - 21$ instead of X *i.e.* we shall find $\rho_{Z,Y}$ which is same as $\rho_{X,Y}$. $E[ZY] = 0.22+0.36+0.36+0.2+0.28+0.12 = 1.54$. Since the marginal distribution of Y has already been found in part **h**, we now need to find the marginal p.m.f. $p_Z(z)$ of Z which is as follows:

z	0	1	2	3
$p_Z(z)$	0.1	0.4	0.3	0.2

.

This yields $E[Z] = 1.6$ and $E[Z^2] = 3.5$ so that $V[Z] = 0.94$. Likewise using the marginal p.m.f $p_Y(y)$ of Y in **h** we get, $E[Y] = 0.97$ and $E[Y^2] = 1.39$ so that $V[Y] = 0.9409$. $\text{Cov}(Z,Y) = E[ZY] - E[Z]E[Y] = 1.54 - 1.6 \times 0.97 = -0.012$ and thus $\rho_{X,Y} = \rho_{Z,Y} = \frac{-0.012}{\sqrt{0.94 \times 0.9409}} \approx -0.0128$.

(k): The conditional distribution of $Y|X$ is given by

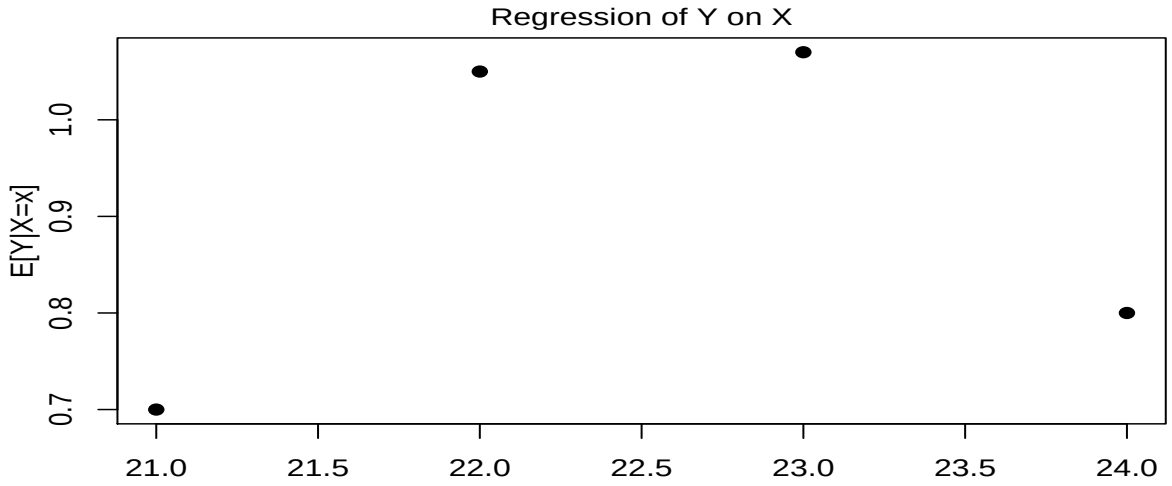
$p_{Y\|X}(y \to \|X = x \downarrow)$	0	1	2
21	0.50	0.30	0.20
22	0.20	0.55	0.25
23	$0.1\dot{6}$	0.60	$0.2\dot{3}$
24	0.30	0.60	0.10

and thus the regression of $Y|X$ or $E[Y|X = x]$ is given by

x	21	22	23	24
$E[Y\|X = x]$	0.70	1.05	$1.0\dot{6}$	0.8

.

the graph of which is as follows:



Thus the most productive age of the engineers is 23. ▽

Example 3.21: Let X denote the sugar content (in gm.) and Y denote the preference score (larger the better) of customers for a 100 gm. cup of a certain brand of ice-cream. The joint p.m.f. $p(x,y)$ of (X,Y), estimated after a market survey, is as follows:

$y \to$ $x \downarrow$	1	2	3
5	0.15	0.12	0.03
7	0.08	0.12	0.20
10	0.06	0.12	0.12

Answer the following:

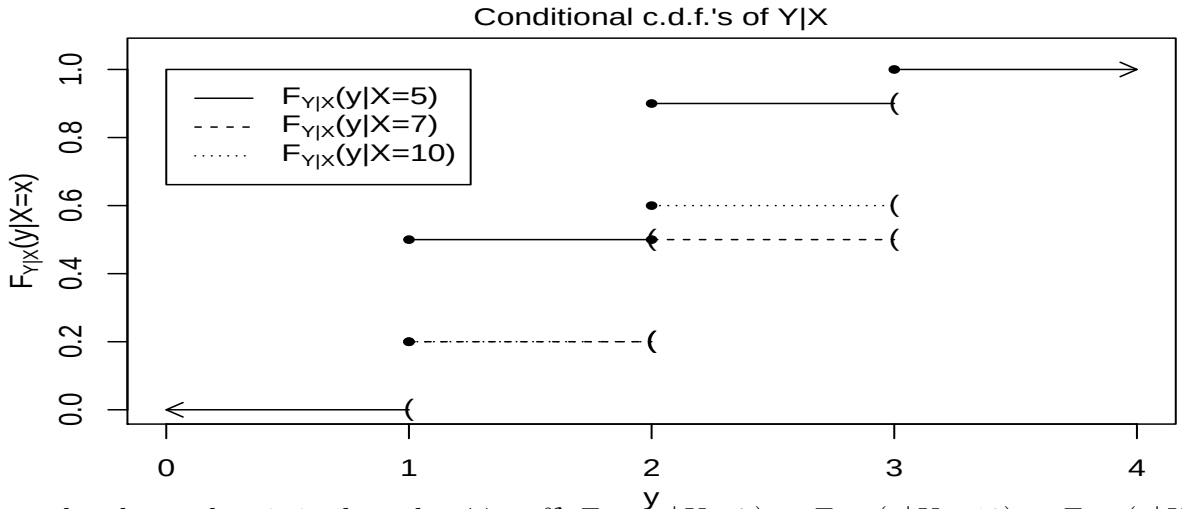
a. What is the probability of a customer giving a score of at least 2 to ice-creams containing at least 7 gm. of sugar?

b. Find the conditional distributions of Y given X and based on their stochastic ranking recommend the optimal level of sugar content for the ice-cream.

Solution (a): $P(Y \geq 2|X \geq 7) = \frac{P(Y \geq 2 \cap X \geq 7)}{P(X \geq 7)} = \frac{0.12+0.20+0.12+0.12}{0.4+0.3} = 0.8$.
(b): The conditional p.m.f.'s of $Y|X = x$ is given by

$p_{Y\|X}(y \rightarrow \|X = x \downarrow)$	1	2	3
5	0.5	0.4	0.1
7	0.2	0.3	0.5
10	0.2	0.4	0.4

and the conditional c.d.f.'s $F_{Y|X}(y|X = x)$ based on the above conditional p.m.f.'s are as follows:



From the above plots it is clear that $\forall y \in \Re\ F_{Y|X}(y|X=7) \leq F_{Y|X}(y|X=10) \leq F_{Y|X}(y|X=5)$. Thus we are in a (rare) situation here where there is a clear-cut stochastic ranking of the conditional distributions of $Y|X$. $Y|X = 7 \overset{\text{st.}}{\geq} Y|X = 10 \overset{\text{st.}}{\geq} Y|X = 5$ and thus the customer preference scores, though random, are clearly stochastically the largest when $X = 7$. Therefore the optimal sugar content should be 7. Note that the stochastic ordering $Y|X = 7 \overset{\text{st.}}{\geq} Y|X = 10 \overset{\text{st.}}{\geq} Y|X = 5$ necessarily implies that $E[Y|X = 7] \geq E[Y|X = 10] \geq E[Y|X = 5]$ which can also be independently and directly verified through the computation of the regression function $E[Y|X = 5] = 1.6$, $E[Y|X = 7] = 2.3$ and $E[Y|X = 10] = 2.2$. However the other way round need not be necessarily true *i.e.* in general $E[Y|X = x_1] > E[Y|X = x_2] \not\Rightarrow Y|X = x_1 \overset{\text{st.}}{\geq} Y|X = x_2$. Thus in a rare situation such as in this example, where the conditional distributions can be ordered, decisions can be taken based on the totality of the conditional distributions without resorting to regression. In general however such an ordering may not be possible which forces one to use the next best thing *viz.* regression to take such decisions. $\bigtriangledown$

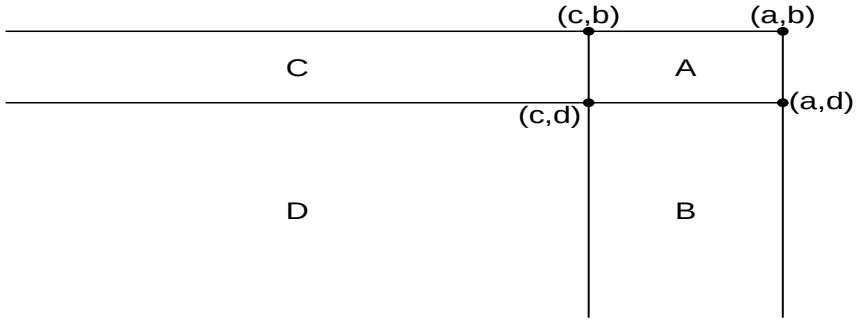
3.5.2 Continuous Case

Just as in the case of the univariate random variables, here also the bivariate continuous random vector is defined in terms of the continuity of the bivariate c.d.f.. However unlike the univariate case where the c.d.f. was defined during the discussion of discrete case itself, here we are yet to define the bivariate c.d.f..

Definition 3.17: For a bivariate random vector $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, its **c.d.f.** is given by $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$.

Definition 3.18: A bivariate random vector $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is said to be **continuous** if its c.d.f. $F(x_1, x_2)$, a function of two variables, is a continuous function of (x_1, x_2).

The utility of c.d.f. in probability computation was fairly convincing in the univariate case, where it was demonstrated how to compute probabilities of intervals using the c.d.f.. In two dimension the analogue of intervals are rectangles. The way probabilities of rectangles in a plane are determined by a bivariate c.d.f. may be understood with the help of the following diagram:



That is, suppose we are interested in the probability of the random vector taking values in the rectangle A in the above diagram with the points (a, b) and (c, d) respectively as its north-east and south-west corners with $c < a$ and $d < b$ *i.e.* $A = [\{c < X_1 \leq a\} \cap \{d < X_2 \leq b\}]$. Now $F(a, b)$ gives the probability of the union of rectangles $A \cup B \cup C \cup D$, where B, C and D are respectively used to denote the (potentially infinite) rectangles $[\{c < X_1 \leq a\} \cap \{X_2 \leq d\}]$, $[\{X_1 \leq c\} \cap \{d < X_2 \leq b\}]$ and $[\{X_1 \leq c\} \cap \{X_2 \leq d\}]$. Note that $F(c, b) = P(C \cup D)$, $F(a, d) = P(B \cup D)$ and $F(c, d) = P(D)$. Thus

$$\begin{aligned} & P(A) \\ &= P(A \cup B \cup C \cup D) - P(B \cup C \cup D) \\ &= F(a, b) - \{P(C \cup D) + P(B \cup D) - P(D)\} \\ &= F(a, b) - F(c, b) - F(a, d) + F(c, d) \end{aligned} \tag{7}$$

Thus since the probability of any rectangle may be computed using the c.d.f. so will be the probability of any set in the plane which can be approximated by rectangles.

Now let us study the consequence of $F(\cdot, \cdot)$ being continuous over the plane $\Re^2$. Fix any $(x_1, x_2) \in \Re^2$ and consider a small rectangle around it given by $[\{x_1 < X_1 < x_1 + dx_1\} \cap \{x_2 <$

$X_2 < x_2 + dx_2\}]$. If dx_1 and dx_2 are positive, by (7)

$$\begin{aligned} & P[\{x_1 < X_1 \le x_1 + dx_1\} \cap \{x_2 < X_2 \le x_2 + dx_2\}] \\ &= F(x_1 + dx_1, x_2 + dx_2) - F(x_1, x_2 + dx_2) - F(x_1 + dx_1, x_2) + F(x_1, x_2) \end{aligned} \tag{8}$$

and thus

$$\lim_{\substack{dx_1 \to 0 \\ dx_2 \to 0}} P[\{x_1 < X_1 \le x_1 + dx_1\} \cap \{x_2 < X_2 \le x_2 + dx_2\}] = P[X_1 = x_1, X_2 = x_2] = 0$$

As a matter of fact, not just a single point like (x_1, x_2) , using (7) and a little bit of mathematical analysis one can show that probability of any "one-dimensional" subset (*e.g.* $[X_1 = x_1]$, $[X_2 = x_2]$, $[aX_1 + bX_2 = c]$, $[aX_1^2 \pm bX_2^2 = c]$ etc. of which the first three are straight lines and the last one is a conic section) of $\Re^2$ will also be 0 if $F(x_1, x_2)$ is continuous.

Just as in the univariate case, for a smooth handle on the continuous random vectors, here also it will be much more convenient if we can cook up a notion of probability density analogous to joint p.m.f. of §5.1. Taking a lead from the univariate case, where the p.d.f. was interpreted as the limit of the probability per unit length, as the length goes to 0; here let us study what happens to the limit of the probability per unit *area* (in $\Re^2$) as the area goes to 0. The probability content of a rectangle around a point $(x_1, x_2) \in \Re^2$ of area $dx_1 dx_2$ is given in (8). Diving this probability by the area of the rectangle and letting the area go to 0 we find that

$$\begin{aligned} & \lim_{\substack{dx_1 \to 0 \\ dx_2 \to 0}} \frac{P[\{x_1 < X_1 \le x_1 + dx_1\} \cap \{x_2 < X_2 \le x_2 + dx_2\}]}{dx_1 dx_2} \\ &= \lim_{dx_2 \to 0} \frac{1}{dx_2} \lim_{dx_1 \to 0} \frac{F(x_1 + dx_1, x_2 + dx_2) - F(x_1, x_2 + dx_2)}{dx_1} \\ &\qquad\qquad - \lim_{dx_2 \to 0} \frac{1}{dx_2} \lim_{dx_1 \to 0} \frac{F(x_1 + dx_1, x_2) - F(x_1, x_2)}{dx_1} \\ &= \lim_{dx_2 \to 0} \frac{1}{dx_2} \frac{\partial}{\partial x_1} F(x_1, x_2 + dx_2) - \lim_{dx_2 \to 0} \frac{1}{dx_2} \frac{\partial}{\partial x_1} F(x_1, x_2) \\ &= \lim_{dx_2 \to 0} \frac{\frac{\partial}{\partial x_1} F(x_1, x_2 + dx_2) - \frac{\partial}{\partial x_1} F(x_1, x_2)}{dx_2} \\ &= \frac{\partial}{\partial x_2} \frac{\partial}{\partial x_1} F(x_1, x_2). \end{aligned}$$

Pairing the first term with the third and the second with the fourth of the r.h.s. of (8), and then letting dx_2 go to 0 first and then dx_1, and following exactly the same argument as above we again get

$$\lim_{\substack{dx_1 \to 0 \\ dx_2 \to 0}} \frac{P[\{x_1 < X_1 \le x_1 + dx_1\} \cap \{x_2 < X_2 \le x_2 + dx_2\}]}{dx_1 dx_2} = \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} F(x_1, x_2).$$

Thus if the repeated partial derivatives of $F(x_1, x_2)$ exist, then the quantity $f(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F(x_1, x_2) = \frac{\partial^2}{\partial x_2 \partial x_1} F(x_1, x_2)$ may be interpreted as the **joint probability density function** of (X_1, X_2) in the sense that for $dx_1, dx_2 \to 0$

$$P[\{x_1 < X_1 \leq x_1 + dx_1\} \cap \{x_2 < X_2 \leq x_2 + dx_2\}] \approx f(x_1, x_2) dx_1 dx_2. \tag{9}$$

With the above interpretation we formally define the **joint p.d.f.** of a bivariate random vector $\boldsymbol{X}$ as follows.

Definition 19: A function $f : \Re^2 \to \Re$ is called a **joint p.d.f.** if
a. $\forall (x_1, x_2) \in \Re^2$, $f(x_1, x_2) \geq 0$, and
b. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1$

For an arbitrary subset $A \subseteq \Re^2$ of the plane, $P[(X_1, X_2) \in A]$ may be found by first dividing A into small rectangles of area $dx_1 dx_2$, approximating the probability content of such a rectangle around a point $(x_1, x_2) \in A$ by $f(x_1, x_2) dx_1 dx_2$, then adding the probabilities of such rectangles for getting an approximate value of $P[(X_1, X_2) \in A]$ and finally the exact value is obtained by letting $dx_1, dx_2 \to 0$. This process yields nothing but the double integral of $f(\cdot, \cdot)$ on A and thus for $A \subseteq \Re^2$,

$$P[(X_1, X_2) \in A] = \int_A \int f(x_1, x_2) dx_1 dx_2.$$

Before working with some examples, we first define all the other concepts associated with the continuous random vectors. This is because the basic ideas of all these concepts have already been introduced in §5.1 in the context of discrete random vectors, where the emphasis was lay-ed on understanding the concepts rather than the technicalities and thus the notions were introduced using numerical examples. Here the corresponding concepts are exactly the same as in the discrete case (and thus the reader is expected to be already aware of what they really mean) except that their technical definitions differ (in terms p.d.f.'s in place of p.m.f.'s and thus replacing the summations by the integrals - exactly as in the univariate case). Since it is just a matter of technicality (of concepts already introduced) rather than introduction of new entities altogether, I prefer to define them first and defer working out a few examples till we get the holistic view.

Thus let us first define the **marginal distributions** in terms of the **marginal c.d.f.'s** and **marginal p.d.f.'s**. For $i = 1, 2$, the marginal c.d.f. and p.d.f. of X_i shall be denoted by $F_i(\cdot)$ and $f_i(\cdot)$ respectively.

$$F_1(x_1) = P(X_1 \leq x_1) = P(X_1 \leq x_1, X_2 < \infty) = \lim_{x_2 \to \infty} F(x_1, x_2) = F(x_1, \infty).$$

$$F_2(x_2) = P(X_2 \leq x_2) = P(X_1 < \infty, X_2 \leq x_2) = \lim_{x_1 \to \infty} F(x_1, x_2) = F(\infty, x_2).$$

Alternatively,

$$F_1(x_1) = P(X_1 \leq x_1) = P(X_1 \leq x_1, X_2 < \infty) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(t, x_2) dx_2 dt,$$

and therefore

$$f_1(x_1) = \frac{d}{dx_1} F_1(x_1) = \frac{d}{dx_1} \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(t, x_2) dx_2 dt = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2.$$

Similarly

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1.$$

Unlike the discrete case, defining **conditional distributions** of $X_2|X_1 = x_1$ ($X_1|X_2 = x_2$) here is a little tricky task as $P[X_1 = x_1] = 0$ ($P[X_2 = x_2] = 0$). However the **conditional p.d.f.** of $X_2|X_1 = x_1$ ($X_1|X_2 = x_2$) may be defined with the help of the limiting arguments as follows. Let $F_{2|1}(x_2|x_1)$ denote the **conditional c.d.f.** of $X_2|X_1 = x_1$. Then

$$\begin{aligned}
& F_{2|1}(x_2|x_1) \\
&= \lim_{dx_1 \to 0} P(X_2 \leq x_2 | x_1 < X_1 \leq x_1 + dx_1) \\
&= \lim_{dx_1 \to 0} \frac{\int_{-\infty}^{x_2} \int_{x_1}^{x_1 + dx_1} f(s,t) ds dt}{\int_{-\infty}^{\infty} \int_{x_1}^{x_1 + dx_1} f(s,t) ds dt} \\
&= \lim_{dx_1 \to 0} \frac{\frac{1}{dx_1} \int_{-\infty}^{x_2} \int_{x_1}^{x_1 + dx_1} f(s,t) ds dt}{\frac{1}{dx_1} \int_{x_1}^{x_1 + dx_1} f_1(s) ds} \\
&= \frac{1}{f_1(x_1)} \int_{-\infty}^{x_2} f(x_1, t) dt
\end{aligned}$$

and therefore the conditional p.d.f. $f_{2|1}(x_2|x_1)$ of $X_2|X_1 = x_1$ is given by

$$\frac{d}{dx_2} F_{2|1}(x_2|x_1) = \frac{1}{f_1(x_1)} \frac{d}{dx_2} \int_{-\infty}^{x_2} f(x_1, t) dt = \frac{f(x_1, x_2)}{f_1(x_1)}$$

Since $f_1(x_1)$ appears in the denominator and there is no guarantee that $f_1(x_1) > 0$, a little bit of additional caution needs to be exercised in defining the conditional densities $X_2|X_1 = x_1$ and $X_1|X_2 = x_2$, which are as follows. Define

$$f_{2|1}(x_2|x_1) = \begin{cases} \frac{f(x_1,x_2)}{f_1(x_1)} & \text{if } 0 < f_1(x_1) < \infty \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_{1|2}(x_1|x_2) = \begin{cases} \frac{f(x_1,x_2)}{f_2(x_2)} & \text{if } 0 < f_2(x_2) < \infty \\ 0 & \text{otherwise} \end{cases}$$

X_1 and X_2 are said to be independent if their marginal densities coincide with the respective conditional densities *i.e.*

$$f_1(x_1) = f_{1|2}(x_1|x_2) \ \ \forall (x_1, x_2) \in \Re^2 \quad \text{or} \quad f_2(x_2) = f_{2|1}(x_2|x_1) \ \ \forall (x_1, x_2) \in \Re^2$$

As has already been discussed in detail in the discrete case (albeit w.r.t. p.m.f.'s) this is equivalent to the condition that

$$f(x_1, x_2) = f_1(x_1) f_2(x_2).$$

Covariance, correlation coefficient and the two regression functions are defined exactly as in §3.5.1, except that the required Expectations are now found by integrating w.r.t. the appropriate density functions instead of summing as before. Thus

$$\begin{aligned}
&\mathrm{Cov}(X_1, X_2) \\
&= E[(X_1 - E[X_1])(X_2 - E[X_2])] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_1 - \int_{-\infty}^{\infty} t f_1(t)dt\right)\left(x_2 - \int_{-\infty}^{\infty} t f_2(t)dt\right) f(x_1, x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f(x_1, x_2)dx_1dx_2 - \int_{-\infty}^{\infty} x_1 f_1(x_1)dx_1 \int_{-\infty}^{\infty} x_2 f_2(x_2)dx_2
\end{aligned}$$

$E[X_2|X_1 = x_1]$, the regression of X_2 on X_1 is given by

$$E[X_2|X_1 = x_1] = \int_{-\infty}^{\infty} x_2 f_{2|1}(x_2|x_1)dx_2$$

and similarly the regression of X_1 on X_2, $E[X_1|X_2 = x_2]$, is given by

$$E[X_1|X_2 = x_2] = \int_{-\infty}^{\infty} x_1 f_{1|2}(x_1|x_2)dx_1.$$

Example 3.3 (Continued): Consider the dart throwing example where the dart is "equally likely" to land anywhere on a circular dartboard of radius r. The notion of "equally likely" was quantified in §3 by saying that the probability of landing in any region is proportional to the area of the region. While that approach was alright, a more crisp way of defining this notion of "equally likely" in the context of such a continuous bivariate random vector would be to say that the joint distribution of the random vector (X, Y), where X and Y respectively denote the abscissa and ordinate of the point where the dirt has landed w.r.t. a pair of orthogonal axes with the center of the dartboard as the origin, is *uniform* over its natural domain.

A distribution over a domain is called uniform, if the density is constant over that domain and 0 elsewhere. Here the domain is given by $\{(x, y) : x^2 + y^2 \leq r^2\}$. If the joint density $f(x, y)$ is a constant c over this domain, the value of the constant needs to be such that $c\int_{\{(x,y):x^2+y^2\leq r^2\}}\int dxdy = 1$ according to requirement **b** of ***Definition 19***. But $c\int_{\{(x,y):x^2+y^2\leq r^2\}}\int dxdy$ simply represents the volume of a right circular cylinder of height c with base $\{(x, y) : x^2 + y^2 \leq r^2\}$, which is same as $\pi r^2 c$, and thus c has to be $\frac{1}{\pi r^2}$. Thus the joint p.d.f. of (X, Y) is given by

$$f(x, y) = \begin{cases} \frac{1}{\pi r^2} & \text{if } x^2 + y^2 \leq r^2 \\ 0 & \text{otherwise} \end{cases}$$

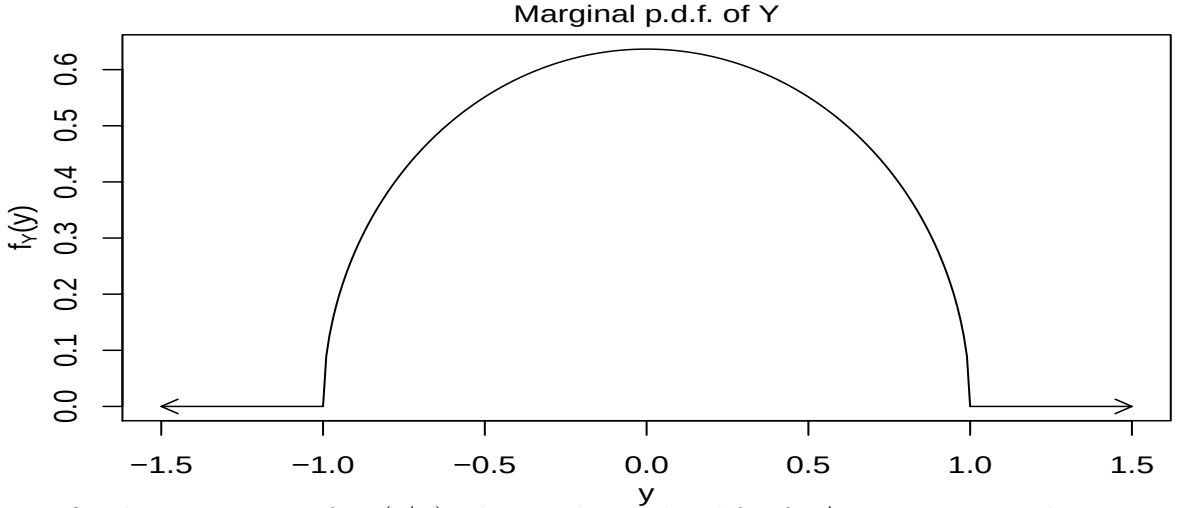
Let $f_X(x)$ denote the marginal p.d.f. of X. Then $f_X(x)$ will take positive values only for $-r \leq x \leq r$ which is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy = \frac{1}{\pi r^2}\int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} dy = \frac{2}{\pi r^2}\sqrt{r^2 - x^2}.$$

The last but one equality follows from the fact that, for a fixed $-r \leq x \leq r$, $f(x,y)$ is the constant $\frac{1}{\pi r^2}$ only for $-\sqrt{r^2-x^2} \leq y < \sqrt{r^2-x^2}$ and 0 otherwise. To check that $f_X(x) = \begin{cases} \frac{2}{\pi r^2}\sqrt{r^2-x^2} & \text{if } -r \leq x \leq r \\ 0 & \text{otherwise} \end{cases}$ is a legitimate p.d.f. first note that it is non-negative and

$$\begin{aligned}
&\int_{-\infty}^{\infty} f_X(x)dx \\
&= \frac{2}{\pi r^2}\int_{-r}^{r}\sqrt{r^2-x^2}dx \\
&= \frac{2}{\pi}\int_{-\pi/2}^{\pi/2}\cos^2\theta d\theta \ \text{ (by substituting } x = r\sin\theta) \\
&= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2}(\cos 2\theta + 1)d\theta \\
&= \frac{1}{\pi}\left(\frac{1}{2}\sin 2\theta\Big|_{\theta=-\pi/2}^{\theta=\pi/2} + \theta\Big|_{\theta=-\pi/2}^{\theta=\pi/2}\right) \\
&= \frac{1}{\pi}\left\{\frac{1}{2}(\sin\pi - \sin(-\pi)) + \left(\frac{\pi}{2}+\frac{\pi}{2}\right)\right\} \\
&= 1
\end{aligned}$$

By symmetry the marginal p.d.f. of Y is given by $f_Y(y) = \begin{cases} \frac{2}{\pi r^2}\sqrt{r^2-y^2} & \text{if } -r \leq y \leq r \\ 0 & \text{otherwise} \end{cases}$ and is plotted below for $r = 1$.



For a fixed $-r \leq x \leq r$, $f_{Y|X}(y|x)$, the conditional p.d.f. of $Y|X = x$ is given by

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{2\sqrt{r^2-x^2}} & \text{if } -\sqrt{r^2-x^2} \leq y \leq \sqrt{r^2-x^2} \\ 0 & \text{otherwise} \end{cases}$$

Note that in $f_{Y|X}(y|x)$, x is considered to be fixed and it is to be viewed as a function of y as it depicts the distribution of Y for a fixed $X = x$. Here it is a constant over the range $-\sqrt{r^2-x^2} \leq y \leq \sqrt{r^2-x^2}$ (and the constant is such that $\int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} f_{Y|X}(y|x)dy = 1$) and thus as explained above, $Y|X = x$ is Uniform over the domain $[-\sqrt{r^2-x^2}, \sqrt{r^2-x^2}]$. By symmetry it is easy to see that $X|Y = y$ is Uniform over the domain $[-\sqrt{r^2-y^2}, \sqrt{r^2-y^2}]$.

Since both $f_X(x)$ and $f_Y(y)$ are symmetric about 0, $E[X] = E[Y] = 0$. Furthermore

$$\begin{aligned} & E[XY] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \\ &= \frac{1}{\pi r^2}\int_{-r}^{r}\left\{y\int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} xdx\right\}dy \\ &= 0 \end{aligned}$$

Thus $\text{Cov}(X,Y) = \rho_{X,Y} = 0$. However note that $f(x,y) \neq f_X(x)f_Y(y)$ and thus serving another example where X and Y are not independent, but still having 0 correlation.

Since $Y|X = x$ is Uniform over $[-\sqrt{r^2-x^2}, \sqrt{r^2-x^2}]$, $E[Y|X = x] = 0$ and so is $E[X|Y = y]$ by symmetry. Thus here the regression of both Y on X as well as X on Y are both constants identically equaling 0. The intuitive reason behind this is as follows. The only job the knowledge of $X = x$ does for Y is fixing its range, because $\forall x \in [-r, r]$ $Y|X = x$ is Uniform. But since this range is symmetric about 0, the mean of the conditional distribution remains unchanged for changing values of $X = x$. Actually in applications this is typically what is expected in case of 0 correlation. When the correlation coefficient is 0, and there is no non-linearity among the association between two variables, one would expect the regression line to be flat. $\bigtriangledown$

3.6 Generating Functions

So far the distributions of random variables have been expressed in a straight forward manner through p.m.f./p.d.f. or c.d.f.. While for visualizing a distribution and moment computations the p.m.f./p.d.f., and for probability and quantile calculations the c.d.f., are indispensable, there are alternative ways of expressing a distribution. These are like capsules which packages the distribution in a unique manner that is used for some specialized purposes. The original distribution in terms of its p.m.f./p.d.f. can be recovered back from these capsules if one desires so, but the main purpose of these capsules are not depicting a distribution in an alternative form (though this can be one of the interpretations of these capsules *viz.* a signature of a distribution), but using them for some special purposes like probability computation for a sum of random variables, moment calculation, proving theoretical results using the capsule as a unique signature of a distribution etc.. We shall discuss three such different packaging of distributions, each having its own unique usages.

3.6.1 Probability Generating Functions

Consider a discrete non-negative integer valued random variable X with $\mathcal{X} = \{0, 1, 2, \ldots\}$ and p.m.f. $p_n = P[X = n]$ for $n = 0, 1, 2, \ldots$.

Definition 3.20: The function $g(t) = \sum_{n=0}^{\infty} p_n t^n$ defined for $t \in [-1, 1]$ is called the **probability generating function**, or p.g.f. for short, of the non-negative integer valued random variable X.

First note that since $\{p_n\}_{n=0}^{\infty}$ is a p.m.f., the p.g.f. $g(t)$ defined above indeed exists $\forall t \in [-1,1]$, as for any $t \in [-1,1]$ $|g(t)| \leq \sum_{n=0}^{\infty} p_n |t|^n \leq \sum_{n=0}^{\infty} p_n = 1$. Next observe that by the law of unconscious statistician, $g(t) = E\left[t^X\right]$. The p.g.f. $g(t)$ actually packages the p.m.f. $\{p_n\}_{n=0}^{\infty}$ in a capsule from which the original p.m.f. can be recovered in the following manner:

$g(0) = p_0 + p_1 0^1 + p_2 0^2 + \cdots = p_0$
$g'(0) = \frac{d}{dt} g(t)\Big|_{t=0} = p_1 + 2p_2 0^1 + 3p_3 0^2 + \cdots = p_1$
$g''(0) = \frac{d^2}{dt^2} g(t)\Big|_{t=0} = 2p_2 + 3.2.p_3 0^1 + 4.3.p_4 0^2 + \cdots = 2p_2$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$
$g^{(n)}(0) = \frac{d^n}{dt^n} g(t)\Big|_{t=0} = n(n-1)\ldots 2.1.p_n + (n+1)n\ldots 2p_{n+1}0^1 + (n+2)(n+1)\ldots 3p_{n+2}0^2$
$+(n+3)(n+2)\ldots 4p_{n+3}0^3 + \cdots = n!p_n$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$

Thus given the p.g.f. $g(t)$ of a non-negative integer valued random variable X, the p.m.f. p_n or $P[X=n]$ can be easily recovered as $p_n = \frac{1}{n!} g^{(n)}(0)$. But that's not the only thing, the moments of the random variable can also be easily computed using the p.g.f. as follows:

$g(1) = p_0 + p_1 + p_2 + \cdots = \sum_{n=0}^{\infty} p_n = 1$
$g'(1) = \frac{d}{dt} g(t)\Big|_{t=1} = 0p_0 + p_1 + 2p_2 + 3p_3 + \cdots = \sum_{n=0}^{\infty} np_n = E[X]$
$g''(1) = \frac{d^2}{dt^2} g(t)\Big|_{t=1} = 0.(0-1).p_0 + 1.(1-1).p_1 + 2.(2-1).p_2 + 3.(3-1).p_3 + \cdots$
$= \sum_{n=0}^{\infty} n(n-1)p_n = E[X(X-1)]$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$
$g^{(n)}(1) = \frac{d^n}{dt^n} g(t)\Big|_{t=1} = 0(0-1)\ldots\left(0-\overline{n-1}\right)p_0 + \cdots + i(i-1)\ldots(i-i)\ldots\left(i-\overline{n-1}\right)p_i$
$+\cdots+(n-1)\left(\overline{n-1}-1\right)\ldots\left(\overline{n-1}-\overline{n-1}\right)p_{n-1} + n(n-1)\ldots\left(n-\overline{n-2}\right)\left(n-\overline{n-1}\right)p_n$
$+(n+1)\left(\overline{n+1}-1\right)\ldots\left(\overline{n+1}-\overline{n-1}\right)p_{n+1} + (n+2)\left(\overline{n+2}-1\right)\ldots\left(\overline{n+2}-\overline{n-1}\right)p_{n+2}$
$+(n+3)\left(\overline{n+3}-1\right)\ldots\left(\overline{n+3}-\overline{n-1}\right)p_{n+2} + \cdots$
$= \sum_{k=0}^{\infty} k(k-1)\ldots\left(k-\overline{n-1}\right)p_n = E\left[X(X-1)\ldots\left(X-\overline{n-1}\right)\right]$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$
$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$

Thus we see that given the p.g.f. $g(t)$ of a (non-negative integer valued) random variable X, $E\left[X(X-1)\ldots\left(X-\overline{n-1}\right)\right]$ is easily found as $g^{(n)}(1)$. Thus for instance, as seen above, $E[X] = g'(1)$. Since $g''(1) = E[X(X-1)]$, $E[X^2] = g''(1) + g'(1)$ so that $V[X] = E[X^2] - (E[X])^2 = g''(1) + g'(1) - (g'(1))^2$. Hence given the p.g.f. it is fairly convenient to extract the raw and central moments of the distribution.

Example 3.4 (Continued): Consider the r.v. X denoting the number of Tails till the first Head appears, in the experiment where a coin is tossed till a Head shows up, introduced in page 5. Recall that this X has p.m.f. $p(x) = q^x p$ for $x = 0, 1, 2, \ldots$. The first two

moments of X were also derived in pp.10-11 using tricky infinite sum calculation. However all these information can be capsuled in its p.g.f. in one go, so that one can then later extract whatever one needs about this r.v. - be it its p.m.f. or moments - from this capsule simply by differentiation. The p.g.f. of X is given by

$$g(t) = E\left[t^X\right] = \sum_{x=0}^{\infty} t^x p(x) = \sum_{x=0}^{\infty} t^x q^x p = p\sum_{x=0}^{\infty}(tq)^x = \frac{p}{1-qt}$$

so that $g'(t) = \frac{pq}{(1-qt)^2}$, $g''(t) = \frac{2pq^2}{(1-qt)^3}$, $g^{(3)}(t) = \frac{3!pq^3}{(1-qt)^4}$ etc. $g^{(n)}(t) = \frac{n!pq^n}{(1-qt)^{n+1}}$. Thus $P[X=n]$ is easily seen to be $g^{(n)}(0)/n! = q^n p$ as it should be. $E\left[X(X-1)\ldots\left(X-\overline{n-1}\right)\right]$ is given by $g^{(n)}(1) = n!\left(\frac{q}{p}\right)^n$ so that $E[X] = \frac{q}{p}$ and $V[X] = g''(1) + g'(1) - (g'(1))^2 = 2\frac{q^2}{p^2} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q^2}{p^2} + \frac{q}{p} = \frac{q^2+pq}{p^2} = \frac{q(q+p)}{p^2} = \frac{q}{p^2}$ as were also found by direct computation in page 11. $\bigtriangledown$

Example 3.22: So far we have mainly dealt with well behaved random variables with finite moments. As an exception, consider a random variable X which takes the value n with probability $\frac{1}{n(n+1)}$ for $n = 1, 2, \ldots$. Let us first check that $p(n) = \frac{1}{n(n+1)}$ for $n = 1, 2, \ldots$ is a legitimate p.m.f.. For this, first note that for $n = 1, 2, \ldots$ $\frac{1}{n(n+1)} > 0$. Thus the only other thing that remains to be shown is that $\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1$ which is proved as follows:

$$\begin{aligned}
&\sum_{n=1}^{\infty} \frac{1}{n(n+1)} \\
&= \lim_{N\to\infty} \sum_{n=1}^{N} \left(\frac{1}{n} - \frac{1}{n+1}\right) \\
&= \lim_{N\to\infty} \left\{\left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \left(\frac{1}{N-1} - \frac{1}{N}\right) + \left(\frac{1}{N} - \frac{1}{N+1}\right)\right\} \\
&= \lim_{N\to\infty} \left\{1 - \frac{1}{N+1}\right\} \\
&= 1
\end{aligned}$$

Note that $E[X]$ does not exist owing to the fact that

$$\sum_{n=1}^{\infty} np(n) = \sum_{n=1}^{\infty} \frac{n}{n(n+1)} = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty.$$

Now the p.g.f. of X is given by

$$g(t) = \sum_{n=1}^{\infty} \frac{t^n}{n(n+1)} = \sum_{n=1}^{\infty} \left(\frac{t^n}{n} - \frac{t^n}{n+1}\right) = \sum_{n=1}^{\infty} \frac{t^n}{n} - \sum_{n=1}^{\infty} \frac{t^n}{n+1}.$$

Now

$$\sum_{n=0}^{\infty} t^n = \frac{1}{1-t} \Rightarrow \int \sum_{n=0}^{\infty} t^n dt = \sum_{n=0}^{\infty} \int t^n dt = \sum_{n=0}^{\infty} \frac{t^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{t^n}{n} = \int \frac{1}{1-t} dt + c = -\log(1-t) + c$$

for some constant c. Since for $t = 0$ both $\sum_{n=1}^{\infty} \frac{t^n}{n}$ and $-\log(1-t)$ equals 0, $c = 0$, yielding $\sum_{n=1}^{\infty} \frac{t^n}{n} = -\log(1-t)$. Similarly

$$\begin{aligned}
&\sum_{n=1}^{\infty} t^n = \frac{t}{1-t} \Rightarrow \\
&\qquad \int \sum_{n=1}^{\infty} t^n dt = \sum_{n=1}^{\infty} \int t^n dt = \sum_{n=1}^{\infty} \frac{t^{n+1}}{n+1} = \int \frac{t}{1-t} dt + c = \int \left\{ \frac{1}{1-t} - 1 \right\} dt + c \\
&= \quad -\log(1-t) - t + c
\end{aligned}$$

for some constant c. Since for $t = 0$ both $\sum_{n=1}^{\infty} \frac{t^{n+1}}{n+1}$ and $-\log(1-t) - t$ equals 0, $c = 0$, yielding $\sum_{n=1}^{\infty} \frac{t^{n+1}}{n+1} = -\log(1-t) - t$ and thus $\sum_{n=1}^{\infty} \frac{t^n}{n+1} = \frac{1}{t} \sum_{n=1}^{\infty} \frac{t^{n+1}}{n+1} = -\frac{1}{t} \log(1-t) - 1$. Therefore the p.g.f. of X is given by

$$g(t) = \sum_{n=1}^{\infty} \frac{t^n}{n} - \sum_{n=1}^{\infty} \frac{t^n}{n+1} = -\log(1-t) + \frac{1}{t} \log(1-t) + 1 = 1 + \frac{1-t}{t} \log(1-t).$$

In what follows, limiting arguments will be used to evaluate $g(0)$, $g'(1)$ etc. and Lhospital's Rule[10] will be used repeatedly for evaluating these limits. As a first check, by Lhospital's Rule, $\lim_{t \to 0} \frac{\log(1-t)}{t} = \lim_{t \to 0} \frac{-1/(1-t)}{1} = -1$ (since both $\lim_{t \to 0} \log(1-t)$ and $\lim_{t \to 0} t$ equals 0 and the derivatives of $\log(1-t)$ and t equals $-\frac{1}{1-t}$ and 1 respectively[11]), and thus $g(0) = 1 + (1-0)(-1) = 0 = P[X = 0]$. Simple substitution of $t = 1$ yields $g(1) = 1$.

$$\begin{aligned}
&g'(t) \\
&= \quad \frac{d}{dt} \left\{ 1 + \frac{1}{t}(1-t) \log(1-t) \right\} \\
&= \quad -\frac{1}{t^2}(1-t) \log(1-t) - \frac{1}{t} \log(1-t) - \frac{1}{t} \\
&= \quad -\frac{1}{t} \left[1 + \log(1-t) \left\{ 1 + \frac{1-t}{t} \right\} \right] \\
&= \quad -\frac{1}{t^2} \left[t + \log(1-t) \right]
\end{aligned}$$

$$g'(0) = \lim_{t \to 0} \left\{ -\frac{1}{t^2} \left[t + \log(1-t) \right] \right\} = -\lim_{t \to 0} \frac{1 - \frac{1}{1-t}}{2t} = -\lim_{t \to 0} \frac{-1/(1-t)^2}{2} = \frac{1}{2} = P[X = 1]$$

Similarly repeatedly using Lhospital's rule and tedious algebra it can be shown that $g''(0) = \frac{2}{2.3} = 2P[X = 2]$, $g^{(3)}(0) = \frac{3!}{3.4} = 3!P[X = 3]$ etc.. Now what about the moments? We have

[10] Lhospital's Rule states that if

i. $f(x)$ and $g(x)$ are real valued functions with continuous derivatives with $g'(x) \neq 0$

ii. both $\lim_{x \to a} f(x)$ and $\lim_{x \to a} g(x)$ are 0, and

iii. $\lim_{x \to a} \frac{f'(x)}{g'(x)} = L$

then $\lim_{x \to a} \frac{f(x)}{g(x)} = L$.

[11] From now on checking the conditions for the applicability of Lhospital's Rule will be left to the reader and will not be explicitly worked out as in this case.

seen above that $E[X]$ does not exist. Thus if one attempts to evaluate $\lim_{t\to 1} g'(t)$ using the expression for $g'(t)$ derived above, it is easy to check that this limit does not exist confirming the same result of non-existence of $E[X]$. $\bigtriangledown$

Above we saw how to use the p.g.f. for obtaining the p.m.f. and moments of (non-negative integer valued) random variables. We shall end this sub-section after illustrating its use in another domain of application, namely in the treatment of sum of independent and identically distributed (called i.i.d. for short) random variables, which crops up every now and then in statistical applications. Thus suppose $X_1, \ldots, X_n$ be i.i.d. with p.g.f. $g(t)$. The question is then what is the expression of $g_n(t)$, the p.g.f. of $S_n = X_1 + \cdots + X_n$ in terms of $g(t)$? This can be found as follows:

$$g_n(t) = E\left[t^{S_n}\right] = E\left[t^{X_1+\cdots+X_n}\right] = E\left[t^{X_1}\right]\ldots E\left[t^{X_n}\right] = \underbrace{g(t)\ldots g(t)}_{n\text{-times}} = [g(t)]^n$$

The third equality follows from the independence of $X_1, \ldots, X_n$ and the next equality follows from the fact that all the X_i's have identical distribution and hence the same p.g.f. $g(t)$. The power of this result can be well appreciated from the remaining examples in this sub-section. However since these examples require the following theorem, which is useful in its own right, we shall present this theorem, called **Negative Binomial Theorem** first. However even before that we first need to introduce the notation of negative binomial coefficients.

For a positive integer n, the binomial coefficient $\left(\begin{array}{c} n \\ k \end{array}\right)$ is given by $\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{(n-k)!k!} = \frac{1}{k!}n.(n-1)\ldots(n-k+1)$. Now for any real number a and non-negative integer k define $\left(\begin{array}{c} a \\ k \end{array}\right)$ as $\left(\begin{array}{c} a \\ k \end{array}\right) = \frac{1}{k!}a(a-1)\ldots(a-k+1)$. Thus for negative integer $-n$ the negative binomial coefficient is given by

$$\begin{aligned}
&\left(\begin{array}{c} -n \\ k \end{array}\right) \\
&= \frac{1}{k!}(-n)(-n-1)\ldots(-n-k+1) \\
&= (-1)^k\frac{1}{k!}n.(n+1)\ldots(n+k-1) \\
&= (-1)^k\frac{(n+k-1).(n+k-2)\ldots(n+1).n.(n-1)\ldots 1}{k!(n-1)!} \\
&= (-1)^k\left(\begin{array}{c} n+k-1 \\ k \end{array}\right)
\end{aligned}$$

Negative Binomial Theorem: For $|t| < 1$ and positive integer n,

$$(1-t)^{-n} = \sum_{k=0}^{\infty}\left(\begin{array}{c} -n \\ k \end{array}\right)(-1)^k t^k = \sum_{k=0}^{\infty}\left(\begin{array}{c} n+k-1 \\ k \end{array}\right)t^k.$$

Proof: With the last equality being same as that of the definition of the negative binomial coefficient, we shall prove that $(1-t)^{-n} = \sum_{k=0}^{\infty}\left(\begin{array}{c} n+k-1 \\ k \end{array}\right)t^k$. We shall prove it by

induction. For $n = 1$ it simply states that $\frac{1}{1-t} = \sum_{k=0}^{\infty} t^k$, which is nothing but the infinite geometric series formula for $|t| < 1$. Now assume that the equality is true for some $n \geq 1$.

$$
\begin{aligned}
& (1-t)^{-(n+1)} \\
&= \frac{1}{n}\frac{d}{dt}(1-t)^{-n} \\
&= \frac{1}{n}\frac{d}{dt}\sum_{k=0}^{\infty}\binom{n+k-1}{k}t^k && \text{(by induction hypothesis)} \\
&= \frac{1}{n}\sum_{k=1}^{\infty}\frac{(n+k-1)!}{k!(n-1)!}\frac{d}{dt}t^k && \text{(interchange allowed for a power series)} \\
&= \sum_{k=1}^{\infty}\frac{k}{n}\frac{(n+k-1)!}{k!(n-1)!}t^{k-1} \\
&= \sum_{k=1}^{\infty}\frac{(n+k-1)!}{(k-1)!n!}t^{k-1} \\
&= \sum_{\ell=0}^{\infty}\frac{(n+\ell)!}{\ell!n!}t^{\ell} && \text{(by substituting } \ell = k-1) \\
&= \sum_{\ell=0}^{\infty}\binom{\overline{n+1}+\ell-1}{\ell}t^{\ell} && \triangledown
\end{aligned}
$$

Example 3.23: Suppose a student is taking 5 courses in a semester, with each course being marked with an integer score between 0 and 100, and we are interested in finding the number of ways in which the student can secure a total marks of 300 in these 5 courses in that given semester. Although it is a combinatorics problem, none of the counting methods we have learned in the previous chapter can come to our rescue for solving this problem. We shall convert this to a probability problem instead from which we shall get this count.

For $i = 1, 2, \ldots, 5$ let the marks scored in the i-th course be denoted by X_i. We make (the very unrealistic) assumption (but it does not matter) that the student is equally likely to score any marks between 0 to 100 in all the 5 courses *i.e.* $\forall i = 1, 2, \ldots, 5$, $P[X_i = j] = \frac{1}{101}$ $\forall j = 0, 1, \ldots, 99, 100$. We also further assume that the X_i's are independent. Then the total score the student gets is given by $S = X_1 + X_2 + \cdots + X_5$ and we are interested in counting the number ways in which $S = 300$. Now the total number of possible 5-tuples in the 5 courses that the student can score is given by 101^5 and thus if we can first compute $P[S = 300]$, multiplying it by 101^5 will give the number of ways in which the student can score a total of 300. The probability distribution of S is determined by using its p.g.f..

Since $P[X_i = j] = \frac{1}{101}$ $\forall j = 0, 1, \ldots, 99, 100$, the p.g.f. of X_i is given by $\frac{1}{101}[1+t+\cdots+t^{100}] = \frac{1}{101}\frac{1-t^{101}}{1-t}$ by the geometric series formula, and thus the p.g.f. of $S = X_1 + X_2 + \cdots + X_5$ is given by $\frac{1}{101^5}\frac{\left(1-t^{101}\right)^5}{(1-t)^5}$. Hence $P[S = 300]$ can be found by figuring out the coefficient of t^{300} in the expansion of $\frac{1}{101^5}\frac{\left(1-t^{101}\right)^5}{(1-t)^5}$ and thus the number of ways $S = 300$ is nothing but the

coefficient of t^{300} in the expansion of $\frac{\left(1-t^{101}\right)^5}{(1-t)^5}$. Now

$$\frac{(1-t^{101})^5}{(1-t)^5} = \sum_{k=0}^{5} \left(\begin{array}{c} 5 \\ k \end{array} \right) (-1)^k t^{101k} \sum_{\ell=0}^{\infty} \left(\begin{array}{c} 4+\ell \\ \ell \end{array} \right) t^{\ell}$$

There are only 6 terms in the first (finite) summation *viz.* t^0, t^{101}, t^{202}, t^{303}, t^{404} and t^{505} and of these the last three terms do not contribute to the coefficient of t^{300} in the final expansion, as the second (infinite) sum only has non-negative powers of t in it. Thus in order to find the coefficient of t^{300} in the final expansion, we just need to multiply the coefficients of t^0, t^{101} and t^{202} in the first sum with the respective coefficients of t^{300}, t^{199} and t^{98} in the second sum, and then add them up. This gives

$$\begin{aligned} & \left(\begin{array}{c} 5 \\ 0 \end{array} \right) \left(\begin{array}{c} 304 \\ 300 \end{array} \right) - \left(\begin{array}{c} 5 \\ 1 \end{array} \right) \left(\begin{array}{c} 203 \\ 199 \end{array} \right) + \left(\begin{array}{c} 5 \\ 2 \end{array} \right) \left(\begin{array}{c} 102 \\ 98 \end{array} \right) \\ = & \frac{1}{24} \left[304 \times 303 \times 302 \times 301 - 5 \times 203 \times 202 \times 201 \times 200 + 10 \times 102 \times 101 \times 100 \times 99 \right] \\ = & 47,952,376 \end{aligned}$$

ways of scoring a total of 300. ▽

Example 3.24: This is a generalization of the above example in which an m-faced dice with faces marked with integers $1, 2, \ldots, m$ is rolled n times and we are interested in the distribution of the sum of the faces S. Assuming the dice is fair, and letting X_i denote the outcome of the i-th roll, the p.g.f. of X_i is given by $\frac{1}{m}\left[t + t^2 + \cdots t^m\right] = \frac{1}{m} t(1-t^m)(1-t)^{-1}$, so that the p.g.f. of the sum is given by $\frac{1}{m^n} t^n (1-t^m)^n (1-t)^{-n}$. For a given $k \in \{n, n+1, \ldots, mn\}$, $P[S=k]$ is given by the coefficient of t^k in the expansion of $\frac{1}{m^n} t^n (1-t^m)^n (1-t)^{-n}$, which is found as follows.

$$t^n (1-t^m)^n (1-t)^{-n} = t^n \sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) (-1)^i t^{m \times i} \sum_{j=0}^{\infty} \left(\begin{array}{c} n+j-1 \\ j \end{array} \right) t^j$$

Thus the coefficient of t^k in this expression is given by

$$\sum_{i=0}^{n} (-1)^i \left(\begin{array}{c} n \\ i \end{array} \right) \left(\begin{array}{c} k-m \times i-1 \\ k-n-m \times i \end{array} \right) = \sum_{i=0}^{n} (-1)^i \left(\begin{array}{c} n \\ i \end{array} \right) \left(\begin{array}{c} k-m \times i-1 \\ n-1 \end{array} \right).$$

As in the above example it is clear that all i's from 0 to n are not going to contribute in the above sum towards the coefficient of t^k, and in particular, i should be such that $n+m \times i \leq k$. However the expression above is theoretically correct *i.e.* there is no harm in writing the sum for $i = 0$ to n, as for large i's violating the condition *i.e.* for i's such that $n+m \times i > k$, $k-m \times i-1 < n-1$ and thus by the definition of binomial coefficients $\left(\begin{array}{c} k-m \times i-1 \\ n-1 \end{array} \right)$ for such i's will be 0, rendering the expression correct.

A couple of numerical illustrations should make the the above formula clear. Suppose a regular 6-faced dice is thrown 10 times and we are interested in the probability of obtaining a sum

of 30. The number of ways the sum can be 30 is given by $\sum_{i=0}^{10}(-1)^i\binom{10}{i}\binom{30-6i-1}{9}$. $30-6i-1 \geq 9 \Leftrightarrow i \leq 3$ and thus for $i \geq 4$ $\binom{30-6i-1}{9} = 0$. Thus the coefficient of t^{30} is given by

$$\binom{10}{0}\binom{29}{9} - \binom{10}{1}\binom{23}{9} + \binom{10}{2}\binom{17}{9} - \binom{10}{3}\binom{11}{9}$$

which after some arithmetic is found to be 2,930,455 so that the probability of the sum being 30 is given by $2,930,455/6^{10} = 0.04846$.

Now suppose again a regular 6-faced dice is thrown 5 times. Obviously the sum cannot exceed 30 in this case. But what does the formula say about the probability of getting the sum to be a number bigger than 30? Let's find out for the simplest case, $k = 31$. $31-6i-1 \geq 4 \Leftrightarrow i \leq 4$. Thus the coefficient of t^{31} is given by

$$\begin{aligned}
&\sum_{i=0}^{4}(-1)^i\binom{5}{i}\binom{30-6i}{4} \\
&= \binom{5}{0}\binom{30}{4} - \binom{5}{1}\binom{24}{4} + \binom{5}{2}\binom{18}{4} - \binom{5}{3}\binom{12}{4} + \binom{5}{4}\binom{6}{4} \\
&= 27405 - 53130 + 30600 - 4950 + 75 \\
&= 0
\end{aligned}$$

▽

Example 3.25: For $i = 1, 2, \ldots, n$ suppose X_i's are i.i.d. with $P[X_i = 1] = p$ and $P[X_i = 0] = 1 - p = q$ (say). We are interested in figuring out the distribution of the sum $S = \sum_{i=1}^{n} X_i$. You can think of it as a generalization of **Example 1** where we were concerned with the distribution of the number of Heads in 3 tosses of a coin with $P(H) = 0.6$. Instead of 3 tosses, now the coin is being tossed n times and instead of $P(H) = 0.6$ it is now an arbitrary real number $p \in [0, 1]$. Code the result of the i-th toss as 1 if it results in a Head and 0 otherwise. Then S simply represents the number of Heads in n tosses of this coin. Among all different treatments of this problem of figuring out the distribution of S the one with p.g.f. is at least algebraically the most straight forward one. The p.g.f. of X_i is given by $P[X_i = 0]t^0 + P[X_i = 1]t = q + pt$, as $P[X_i = j] = 0\ \forall j \geq 2$. Therefore the p.g.f. of S is given by $g(t) = (q + pt)^n = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}t^k$ by the binomial theorem. Hence by directly reading out the coefficient of t^k in the binomial expansion of the p.g.f. of S it can be found that $P[S = k] = \binom{n}{k}p^k q^{n-k}\ \forall k = 0, 1, \ldots, n$. This is the famous **Binomial** random variable, which is useful in modeling various natural phenomena, and will be discussed in detail in the next chapter. For now let us just compute its mean and variance using the p.g.f. approach.

$$g'(t) = \frac{d}{dt}(q + pt)^n = np(q + pt)^{n-1} \text{ and } g''(t) = \frac{d}{dt}np(q + pt)^{n-1} = n(n-1)p^2(q + pt)^{n-2}$$

Therefore its mean is given by $E[X] = g'(1) = np$ (since $p + q = 1$) and its variance is given by $V[X] = g''(1) + g'(1) - (g'(1))^2 = n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p) = npq$. $\bigtriangledown$

3.6.2 Moment Generating Functions

During our discussion in the previous section on p.g.f., apart from parenthetically mentioning a few times, the fact that p.g.f. is only defined for non-negative integer valued random variables, was not overly emphasized. The time has now come to draw attention to this fact. While the quantity t^X is a well-defined real number $\forall t \in [-1, 1]$ for a non-negative integer valued random variable, for an arbitrary random variable X, t^X may not be a real number and even may be undefined. For instance with X taking negative values, t^X is an imaginary number for $t < 0$ and is undefined at $t = 0$ (the criticality of being able to define a generating function at 0 should be clear to the reader from the foregoing discussion on p.g.f.). This gives rise to the need of defining a generating function for an arbitrary random variable X in a slightly different manner than p.g.f..

For an integer valued random variable, since its range of possible values are known before hand, it makes sense to seek for a function, from which its p.m.f. can be retrieved back. For an arbitrary random variable, this property of being able to directly retrieve the p.m.f. or the p.d.f., as the case may be, from a generating function, is too much to expect. Other than capsuling the p.m.f., two other major uses of p.g.f. are easy moment derivation and handling of i.i.d. random variables. Thus while attempting to define a generating function for an arbitrary random variable, these two features are desired while keeping the quantity a well-defined real number at the same time. t^X may be ill-defined for an arbitrary random variable for $t < 0$, but similar properties can be retained if one just forces t to remain positive. This is accomplished if one considers e^{tX} instead.

Definition 3.21: The function $M(t) = E\left[e^{tX}\right]$ (provided that the expectation exists) is called the **moment generating function**, or m.g.f. for short, of the random variable X.

Note that for a non-negative integer valued random variable X, if its p.g.f. is given by $g(s) = E\left[s^X\right]$, its m.g.f. $M(t)$ is easily obtained by substituting $s = e^t$ in its p.g.f. *i.e.* $M(t) = g(e^t)$. Now the question that naturally arises is, "why is it called moment generating function?" In order to answer the question observe that

$$M(t) = E\left[e^{tX}\right] = E\left[1 + tX + \frac{(tx)^2}{2!} + \frac{(tX)^3}{3!} + \cdots\right] = 1 + tE[X] + \frac{1}{2!}t^2E\left[X^2\right] + \frac{1}{3!}t^3E\left[X^3\right] + \cdots$$

so that

$$M(0) = 1, M'(0) = E[X], M''(0) = E\left[X^2\right], M^{(3)}(0) = E\left[X^3\right], \ldots, M^{(n)}(0) = E\left[X^n\right], \ldots$$

Thus the moments of a random variable, if they exist, can be easily obtained by differentiating its m.g.f. and hence the name moment generating function. The m.g.f. does not exist if the moments do not exist. In order to see this let us revisit **Example 22** introduced in the previous subsection.

Example 3.22 (Continued): We have already seen that the random variable X with its p.m.f. $p(n) = \frac{1}{n(n+1)}$ for $n = 1, 2, \ldots$ does not have any finite moment of any order. Also its p.g.f. is derived as $g(s) = 1 + \{(1-s)/s\}\log(1-s)$ for $s \in [-1, 1]$. Now as mentioned above, the m.g.f. $M(t)$ of a non-negative integer valued random variable can be easily obtained by substituting $s = e^t$. But for $t \geq 0$ $\log(1 - e^t)$ is un-defined and thus the m.g.f. of this r.v. does not exist. $\bigtriangledown$

Example 3.4 (Continued): The p.g.f. of the random variable X with p.m.f. $p(x) = q^x p$ for $x = 0, 1, 2, \ldots$ was derived to be $g(s) = \frac{p}{1-qs}$ in the previous subsection. Thus by substituting $s = e^t$ in this p.g.f. its m.g.f. is found to be $M(t) = \frac{p}{1-qe^t}$. As an illustration for deriving moments using the m.g.f. first note that $M(0) = 1$,

$$M'(0) = \frac{d}{dt}M(t)\Big|_{t=0} = \frac{d}{dt}\frac{p}{1-qe^t}\Big|_{t=0} = \frac{pqe^t}{(1-qe^t)^2}\Big|_{t=0} = \frac{pq}{p^2} = \frac{q}{p} = E[X],$$

$$M''(0) = \frac{d^2}{dt^2}M(t)\Big|_{t=0} = \frac{d}{dt}\frac{pqe^t}{(1-qe^t)^2}\Big|_{t=0} = \frac{(1-qe^t)\,pqe^t + 2pq^2e^{2t}}{(1-qe^t)^3}\Big|_{t=0} = \frac{q}{p} + 2\frac{q^2}{p^2}$$

so that

$$V[X] = E\left[X^2\right] - (E[X])^2 = M''(0) - (M'(0))^2 = \frac{q}{p} + 2\frac{q^2}{p^2} - \frac{q^2}{p^2} = \frac{q}{p^2} \qquad \bigtriangledown$$

As mentioned in the beginning of this subsection, the main reason for which one has to go beyond p.g.f. and define m.g.f. is its applicability being not limited to non-negative integer valued random variables only. Whenever the appropriate moments exist, the m.g.f. can be computed and utilized for moment computation of any random variable. Thus in the next example we illustrate the usage of m.g.f. for a continuous random variable.

Example 3.9 (Continued): In this example, we had derived the c.d.f. of how long a light-bulb might last, based on some physical postulates. It was derived that the c.d.f. $F(x)$ of the random variable T denoting the number of hours that a light-bulb will last is given by $F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$ for some parameter λ giving the rate of failure. Now from this it is easy to derive the p.d.f. of T which is given by $f(x) = \frac{d}{dx}F(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if} x \leq 0 \end{cases}$. Obviously one can compute the moments directly from definition using this p.d.f.. That is for example,

$$\begin{aligned} E[X] & \\ &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \lambda \int_0^{\infty} x e^{-\lambda x} dx \\ &= \frac{1}{\lambda}\int_0^{\infty} u e^{-u} du \qquad \text{(by substituting } u = \lambda x) \\ &= \frac{1}{\lambda}\left[-ue^{-u}\Big|_{u=0}^{\infty} + \int_0^{\infty} e^{-u} du\right] \qquad \text{(after integrating by parts)} \end{aligned}$$

$$= \frac{1}{\lambda} \qquad (\text{as } \lim_{u\to 0} ue^{-u} = 0, \lim_{u\to\infty} ue^{-u} = 0 \text{ and } \int_0^\infty e^{-u}du = -e^{-u}\Big|_{u=0}^{\infty} = 1),$$

and

$$\begin{aligned}
&E[X^2] \\
&= \int_{-\infty}^{\infty} x^2 f(x)dx \\
&= \lambda \int_0^\infty x^2 e^{-\lambda x}dx \\
&= \frac{1}{\lambda^2}\int_0^\infty u^2 e^{-u}du \qquad (\text{by substituting } u = \lambda x) \\
&= \frac{1}{\lambda^2}\left[-u^2e^{-u}\Big|_{u=0}^{\infty} + 2\int_0^\infty ue^{-u}du\right] \qquad (\text{after integrating by parts}) \\
&= \frac{2}{\lambda^2} \quad (\text{as } \lim_{u\to 0} u^2e^{-u} = 0, \lim_{u\to\infty} u^2e^{-u} = 0 \text{ and we have just shown that } \int_0^\infty ue^{-u}du = 1),
\end{aligned}$$

so that $V[X] = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$. However for each moment instead of trying to evaluate the integrals again and again, it is much more convenient to perform an integration once and for all, package it in a capsule called m.g.f. and then extract the moments from it as and when required by simple differentiation. Thus we first calculate the m.g.f. of X as

$$\begin{aligned}
&M(t) \\
&= E\left[e^{tX}\right] \\
&= \int_{-\infty}^{\infty} e^{tx} f(x)dx \qquad (\text{by the law of unconscious statistician}) \\
&= \lambda\int_0^\infty e^{-(\lambda-t)x}dx \\
&= \frac{\lambda}{\lambda-t}\int_0^\infty e^{-u}du \qquad (\text{by substituting } u = (\lambda-t)x) \\
&= \frac{\lambda}{\lambda-t}.
\end{aligned}$$

Next $E[X]$ and $E[X^2]$ are obtained as

$$E[X] = M'(0) = \frac{d}{dt}M(t)\Big|_{t=0} = \frac{d}{dt}\frac{\lambda}{\lambda-t}\Big|_{t=0} = \frac{\lambda}{(\lambda-t)^2}\Big|_{t=0} = \frac{1}{\lambda},$$

$$E\left[X^2\right] = M''(0) = \frac{d^2}{dt^2}M(t)\Big|_{t=0} = \frac{d}{dt}\frac{\lambda}{(\lambda-t)^2}\Big|_{t=0} = \frac{2\lambda}{(\lambda-t)^3}\Big|_{t=0} = \frac{2}{\lambda^2}$$

so that again $V[X] = E[X^2] - (E[X])^2 = M''(0) - (M'(0))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$. $\bigtriangledown$

Ease of moment calculation is not the only usage of m.g.f.. One of its major utilization lies in the characterization of a distribution. By this it is meant that each distribution's m.g.f. is its unique signature, which turns out to be extremely useful in proving certain theoretical results. Note that p.g.f. also possess this uniqueness property proof of which lies in the way

one can retrieve the p.m.f. from the p.g.f.. The proof of the uniqueness property of the m.g.f. is beyond the scope of these lecture notes and thus we simply state the theorem below for future reference without attempting to prove it.

Uniqueness Property of m.g.f.: If $M_1(t)$ and $M_2(t)$ respectively are the m.g.f.'s of two distribution functions $F_1(x)$ and $F_2(x)$ such that $M_1(t) \equiv M_2(t)\ \forall t \in \Re$, then $F_1(x) \equiv F_2(x)$ $\forall x \in \Re$.

Thus if we can recognize the m.g.f. of a certain random variable to belong to a particular family of distributions, that can be used to ascertain the distribution of that random variable. Before providing an example illustrating this usage of m.g.f., it would be useful to examine, like p.g.f., how m.g.f. also can be used to handle the sum of i.i.d. random variables. Thus if $X_1, X_2, \ldots, X_n$ are i.i.d. with each having m.g.f. $M(t)$, the m.g.f. of their sum $S = X_1 + X_2 + \cdots + X_n$ is given by

$$M_S(t) = E\left[e^{tS}\right] = E\left[e^{t(X_1+\cdots+X_n)}\right] = E\left[e^{tX_1}\right]\ldots E\left[e^{tX_n}\right] = \underbrace{M(t)\ldots M(t)}_{n-\text{times}} = [M(t)]^n$$

The third equality follows from the independence and the next equality follows from the identical distribution assumption of the X_i's. We finish this sub-section after providing a couple of examples illustrating these last two properties.

Example 3.26: Consider a r.v. X having p.d.f. $f(x) = \begin{cases} \frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$ for some positive integer n and $\lambda > 0$. That this is a legitimate p.d.f. can be proven by repeatedly integrating by parts as follows:

$$\begin{aligned}
&\int_0^\infty x^{n-1}e^{-\lambda x}dx \\
&= \frac{1}{\lambda^n}\int_0^\infty u^{n-1}e^{-u}du && \text{(by substituting } u = \lambda x) \\
&= \frac{1}{\lambda^n}\left[-u^{n-1}e^{-u}\Big|_{u=0}^\infty + (n-1)\int_0^\infty u^{n-2}e^{-u}du\right] && \text{(after integrating by parts)} \\
&= \frac{1}{\lambda^n}\left[-u^{n-2}e^{-u}\Big|_{u=0}^\infty + (n-1)(n-2)\int_0^\infty u^{n-3}e^{-u}du\right] && \text{(since } \lim_{u\to 0} u^{n-1}e^{-u} = \lim_{u\to\infty} u^{n-1}e^{-u} = 0 \\
& && \text{and then again integrating } \int_0^\infty u^{n-2}e^{-u}du \text{ by parts)} \\
&= \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
&= \frac{1}{\lambda^n}\left[e^{-u}\Big|_{u=0}^\infty + (n-1)(n-2)\ldots 1\int_0^\infty e^{-u}du\right] \\
&= \frac{(n-1)!}{\lambda^n}
\end{aligned}$$

so that $\int_0^\infty f(x)dx = 1$. Now in order to find its m.g.f. we again use this integral result just proven.

$$\begin{aligned}
&M(t) \\
&\quad = E\left[e^{tX}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda^n}{(n-1)!}\int_0^\infty e^{tx}x^{n-1}e^{-\lambda x}dx \qquad \text{(by the law of unconscious statistician)}\\
&= \frac{\lambda^n}{(n-1)!}\int_0^\infty x^{n-1}e^{-(\lambda-t)x}dx\\
&= \frac{\lambda^n}{(n-1)!(\lambda-t)^n}\int_0^\infty u^{n-1}e^{-u}du \quad \text{(by substituting } u=(\lambda-t)x)\\
&= \frac{\lambda^n(n-1)!}{(n-1)!(\lambda-t)^n} \qquad \text{(as we have just proven that } \int_0^\infty u^{n-1}e^{-u}du=(n-1)!)\\
&= \left(\frac{\lambda}{\lambda-t}\right)^n
\end{aligned}
$$

Now let us get back to **Example 9 (Continued)**. There we had shown that the random variable T has m.g.f. $\frac{\lambda}{\lambda-t}$. Now consider n i.i.d. copies $T_1,\ldots,T_n$ of T and their sum $S=T_1+\cdots+T_n$. What is the p.d.f. of S? Since m.g.f. of the sum n i.i.d. random variables is the m.g.f. of this random variable raised to the power n, it is clear that the m.g.f. of S is $\left(\frac{\lambda}{\lambda-t}\right)^n$ which is same as that of X we started with. Thus by the Uniqueness property of m.g.f., the p.d.f. of S ($f_S(s)$ say) must equal $f_S(s)=\begin{cases}\frac{\lambda^n}{(n-1)!}s^{n-1}e^{-\lambda s} & \text{if } s\geq 0\\ 0 & \text{if } s<0\end{cases}$. $\bigtriangledown$

Example 3.27: Consider a discrete random variable X with p.m.f. $p(x)=e^{-\lambda}\frac{\lambda^x}{x!}$ for $x=0,1,2,\ldots$ for some parameter $\lambda>0$. That this is a legitimate p.m.f. follows immediately from the fact that $e^\lambda=\sum_{x=0}^\infty\frac{\lambda^x}{x!}$. This random variable is said to follow a **Poisson** distribution with parameter λ and is denoted by $X\sim\text{Poisson}(\lambda)$. (Like the Binomial r.v. of **Example 25**, this distribution, which is one of the most important discrete probability models, will be studied in detail in the next chapter.) The m.g.f. of a Poisson(λ) random variable is given by

$$
M(t)=E\left[e^{tX}\right]=e^{-\lambda}\sum_{x=0}^\infty e^{tx}\frac{\lambda^x}{x!}=e^{-\lambda}\sum_{x=0}^\infty\frac{(\lambda e^t)^x}{x!}=e^{-\lambda}e^{\lambda e^t}=\exp\left\{\lambda\left(e^t-1\right)\right\}.
$$

Now consider n independent Poisson random variables with $X_i\sim\text{Poisson}(\lambda_i)$ for $i=1,2,\ldots,n$. Note that X_i's are independent but not identically distributed as we are allowing the parameter of the distribution λ_i to change with i. We are interested in figuring out the distribution of $S=X_1+\cdots+X_n$. As a motivating background in business application, imagine you own n retail electronic stores in a city and the daily number of 25" color television sets sold by the i-th store has a Poisson distribution (it fits fairly well empirically as well as some theoretical argument can also be put forth in favor of this model) with parameter λ_i, and you are interested in the distribution of the total number of 25" color television sets sold by all these n stores in the city on a given day. The easiest way to figure it out is by using the m.g.f. which is done as follows. M.g.f. of S is given by

$$
\begin{aligned}
&M_S(t)\\
&= E\left[e^{tS}\right]\\
&= E\left[e^{t(X_1+\cdots+X_n)}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[e^{tX_1}\right]\ldots E\left[e^{tX_n}\right] && \text{(by independence of } X_1,\ldots X_n\text{)}\\
&= \exp\left\{\lambda_1\left(e^t-1\right)\right\}\ldots\exp\left\{\lambda_n\left(e^t-1\right)\right\} && \text{(since m.g.f. of Poisson}(\lambda_i)\text{ is } \exp\left\{\lambda_i\left(e^t-1\right)\right\}\text{)}\\
&= \exp\left\{\left(\lambda_1+\cdots+\lambda_n\right)\left(e^t-1\right)\right\}\\
&= \exp\left\{\lambda\left(e^t-1\right)\right\} && \text{(say, where } \lambda=\sum_{i=1}^{n}\lambda_i\text{)}
\end{aligned}
$$

This is clearly identified as the m.g.f. of a Poisson(λ) random variable. Thus $S \sim$ Poisson(λ) and we have proved an important result which says that if for $i = 1, 2, \ldots, n$, $X_i \sim$ Poisson(λ_i) and $X_1, \ldots X_n$ are independent then $X_1 + \cdots + X_n \sim$ Poisson($\lambda_1 + \cdots + \lambda_n$). $\bigtriangledown$

3.6.3 Characteristic Function

While m.g.f. is very useful for characterizing a distribution, moment calculation and handling i.i.d. sum, one major problem with it is that it may not exist (as in **Example 3.22**). Of course if the moments do not exist[12] m.g.f. may not be used for computation of moments but its two other uses mentioned above gets handicapped as well in such cases. This calls for another capsule which will always exist for any r.v.. Note that the p.g.f. serves the purpose for non-negative integer valued random variables, which always exists, but here we are looking for a similar tool for an arbitrary r.v..

Towards this end, instead of requiring to find $E\left[e^{tX}\right]$ if one seeks $E\left[e^{itX}\right]$, where $i = \sqrt{-1}$, then the problem is solved. This is because, by Euler's identity $e^{itX} = \cos(tX) + i\sin(tX)$ and thus $\left|e^{itX}\right| = 1\ \forall t \in \Re$ and hence $\int_{-\infty}^{\infty}\left|e^{itX}\right| f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1 < \infty$ leading to the conclusion that $E\left[e^{itX}\right]$ always exists for any r.v. X.

Definition 3.22: The function $\phi(t) = E\left[e^{itX}\right]$, where $i = \sqrt{-1}$, is called the **characteristic function**, or c.f. for short, of the random variable X.

As seen above c.f. of a r.v. always exists. Interestingly, when the moments exist these can also be computed from the c.f. just as in case of the m.g.f.. To see this, just as we had expanded the m.g.f., expanding the c.f. we get

$$\phi(t) = E\left[e^{itX}\right] = E\left[1 + itX + \frac{(itX)^2}{2!} + \frac{(itX)^3}{3!} + \cdots\right] = 1 + itE[X] + \frac{i^2}{2!}t^2\left[X^2\right] + \frac{i^3}{3!}t^3\left[X^3\right] + \cdots$$

so that

$$\phi(0) = 1, \frac{1}{i}\phi'(0) = E[X], \frac{1}{i^2}\phi''(0) = E\left[X^2\right], \ldots, \frac{1}{i^n}\phi^{(n)}(0) = E\left[X^n\right], \ldots$$

[12] We have never quite formally broached this issue and at this point of time it may be worth while examining this. As usual we shall present it for the p.d.f. case and for the p.m.f. case the equivalent results are found by replacing the p.d.f. term $f(x)dx$ with the p.m.f. $p(x)$ and the integral with summation. The function $g(X)$ of a r.v. X with p.d.f. $f(x)$ is said to have a finite expectation or its mean said to exist if and only if $\int_{-\infty}^{\infty}|g(x)|f(x)dx < \infty$, in which case $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.

As the name suggests one major usage of c.f. is to characterize a distribution. By that we mean that the characteristic function of a distribution is unique and one can identify a distribution from its c.f. just as in case of m.g.f. (in case it exists). As a matter of fact given the c.f. $\phi(t)$ of a r.v. its m.g.f. $M(t) = \phi(-it)$ *i.e.* by substituting $-it$ for s in the expression of the c.f. $\phi(s)$ one obtains the m.g.f. and thus it is no surprise that c.f. inherits all the properties of an m.g.f.. In the case of c.f. however the characterization is a little more crisp. By that we mean one can get an explicit formula for recovering the c.d.f. $F(\cdot)$ given its c.f. $\phi(t)$, just as one has a direct formula of recovering the p.m.f. from the p.g.f. of a non-negative integer valued random variable. The result which enables one to do so in case of the c.f. is called the inversion theorem of c.f. which is just stated below without a proof.

Inversion Theorem for Characteristic Functions: If $\phi(t)$ is the characteristic function of a c.d.f. $F(\cdot)$ and $F(\cdot)$ is continuous in the interval $(x-c, x+c)$ then

$$F(x+c) - F(x-c) = \lim_{T\to\infty} \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ct}{t} e^{-itx} \phi(t) dt$$

Just as in case of the p.g.f. and m.g.f., the c.f. also gives one an easy handle on i.i.d. sums. If $X_1, \ldots, X_n$ are i.i.d. with c.f. $\phi(t)$ the c.f. of the sum $S = X_1 + \cdots + X_n$ is given by

$$\phi_S(t) = E\left[e^{itS}\right] = E\left[e^{it(X_1+\cdots+X_n)}\right] = E\left[e^{itX_1}\right] \ldots E\left[e^{itX_n}\right] = \underbrace{\phi(t)\ldots\phi(t)}_{n-\text{times}} = [\phi(t)]^n$$

At this juncture it should be mentioned that the major use of c.f. is for proving theoretical results. From the application point of view it is enough to be familiar with the notions of p.g.f. and m.g.f.. However for proving theoretical results in general cases, one encounters the pathology of non-existence of m.g.f. and in such situations the tool that is used is the c.f.. We shall prove one such result, which is extremely important from application point of view as well as from motivation perspective of the most famous statistical distribution called the **Normal** distribution, in the next chapter, which requires some preliminaries which is best introduced in this sub-section on characteristic function. Thus we shall give one definition and state one theorem without proof, which is very useful for proving theoretical results, and as mentioned above we shall prove one such theoretical result called the ***Central Limit Theorem*** using this theorem in the next chapter.

Definition 3.23: A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ with X_n having c.d.f. $F_n(\cdot)$ is said to **converge in distribution** or **converge in law** or **converge weakly** to a random variable X with c.d.f. $F(\cdot)$ if $\{F_n(x)\}$ as a sequence of real numbers converges to $F(x)$ for every continuity point x of $F(\cdot)$. In such a situation the convergence is denoted by $X_n \xrightarrow{\mathcal{D}} X$ or $X_n \xrightarrow{\mathcal{L}} X$ or$X_n \xrightarrow{w} X$ or $F_n \Rightarrow F$.

Weak Convergence Theorem: Let $\phi_n(t)$ denote the c.f. of the c.d.f. $F_n(\cdot)$ and $\phi(t)$ denote the c.f. of the c.d.f. $F(\cdot)$, then $F_n \Rightarrow F$ if and only if $\phi_n(t) \to \phi(t)\ \forall t \in \Re$.

We finish this sub-section (as well as this chapter) with a couple of examples, the last one being of theoretical nature demonstrating weak convergence.

Example 3.9 (Continued): The c.f. of the r.v. here is given by

$$\begin{aligned}
&\phi(t) \\
&= E\left[e^{itX}\right] \\
&= \int_{-\infty}^{\infty} e^{itx} f(x)dx \qquad \text{(by the law of unconscious statistician)} \\
&= \lambda \int_0^{\infty} e^{-(\lambda - it)x} dx \\
&= \frac{\lambda}{\lambda - it} \int_0^{\infty} e^{-u} du \qquad \text{(by substituting } u = (\lambda - it)x) \\
&= \frac{\lambda}{\lambda - it}.
\end{aligned}$$

Note that

$$\phi(-is) = \frac{\lambda}{\lambda - i(-is)} = \frac{\lambda}{\lambda + i^2 s} = \frac{\lambda}{\lambda - s} = M(s).$$

Using the c.f.

$$E[X] = \frac{1}{i}\phi'(0) = \frac{1}{i}\,\frac{d}{dt}\frac{\lambda}{\lambda - it}\bigg|_{t=0} = \frac{1}{i}\,\frac{i\lambda}{(\lambda - it)^2}\bigg|_{t=0} = \frac{1}{\lambda}$$

$$E\left[X^2\right] = \frac{1}{i^2}\phi''(0) = \frac{1}{i^2}\,\frac{d}{dt}\frac{i\lambda}{(\lambda - it)^2}\bigg|_{t=0} = \frac{1}{i^2}\,\frac{2i^2\lambda}{(\lambda - it)^3}\bigg|_{t=0} = \frac{2}{\lambda^2}$$

so that $V[X] = E\left[X^2\right] - (E[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$ $\bigtriangledown$

Example 3.28: Consider the Binomial r.v. introduced in **Example 25**. For $n = 1, 2, \ldots$ let $X_n \sim \text{Binomial}(n, p_n)$ *i.e.* $X_n = \sum_{j=1}^n Y_j$ where each Y_j are i.i.d. with $Y_j = \begin{cases} 1 & \text{with probability } p_n \\ 0 & \text{with probability } q_n \end{cases}$, where $q_n = 1 - p_n$. Now assume the sequence p_n is such that $\lim_{n\to\infty} np_n = \lambda > 0$. Note that then $\lim_{n\to\infty} p_n = 0$. Now the question is do these X_n's converge in distribution to anywhere, and if so to which distribution? We shall provide a more direct proof of this in the next chapter, but now we shall present a different proof utilizing the Weak Convergence Theorem involving the characteristic function. We begin by computing $\phi_{X_n}(t)$,the c.f. of X_n. Note that since Y_j's are i.i.d. $\phi_{X_n}(t) = \left[\phi_{Y_j}(t)\right]^n$, where $\phi_{Y_j}(t)$ is the c.f. of Y_j given by

$$\phi_{Y_j}(t) = E\left[e^{itY_j}\right] = e^{it\times 1}p_n + e^{it\times 0}q_n = p_n e^{it} + (1 - p_n) = 1 + p_n\left(e^{it} - 1\right).$$

Therefore $\phi_{X_n}(t) = \left[1 + p_n\left(e^{it} - 1\right)\right]^n$. In order to evaluate $\lim_{n\to\infty}\phi_{X_n}(t)$ we begin by considering $\lim_{n\to\infty}\log\left(\phi_{X_n}(t)\right)$.

$$\begin{aligned}
&\lim_{n\to\infty}\log\left(\phi_{X_n}(t)\right) \\
&= \lim_{n\to\infty} n\log\left[1 + p_n\left(e^{it} - 1\right)\right] \\
&= \lim_{n\to\infty} n\left[p_n\left(e^{it} - 1\right) - \frac{1}{2}p_n^2\left(e^{it} - 1\right)^2 - \frac{1}{3}p_n^3\left(e^{it} - 1\right)^3 + \cdots\right] \quad \left(\text{as } \log(1 + x) = x - \frac{x^2}{2}\right.
\end{aligned}$$

$+\frac{x^3}{3} - \cdots$ for $|x| < 1$ and $\left|p_n\left(e^{it}-1\right)\right| < 1 \forall t \in \Re$ for sufficiently large n as $\left|e^{it}-1\right| \leq 2$ and $\lim_{n\to\infty} p_n = 0$)

$$= \lim_{n\to\infty}\left[(np_n)\left(e^{it}-1\right) - \frac{1}{2}p_n(np_n)\left(e^{it}-1\right)^2 - \frac{1}{3}p_n^2(np_n)\left(e^{it}-1\right)^3 + \cdots\right]$$

$$= \lambda\left(e^{it}-1\right) \quad (\text{as } \lim_{n\to\infty}(np_n)\left(e^{it}-1\right) = \left(e^{it}-1\right)\lim_{n\to\infty}(np_n) = \lambda\left(e^{it}-1\right) \text{ and}$$

$\lim_{n\to\infty}\frac{1}{k}p_n^{k-1}(np_n)\left(e^{it}-1\right)^k = \frac{1}{k}\left(e^{it}-1\right)^k\left\{\lim_{n\to\infty}(np_n)\right\}\left\{\lim_{n\to\infty}p_n^{k-1}\right\} = \frac{1}{k}\left(e^{it}-1\right)^k\lambda \times 0 = 0 \forall k \geq 2)$

Since $\lim_{n\to\infty}\log\left(\phi_{X_n}(t)\right) = \lambda\left(e^{it}-1\right)$, it follows that, $\lim_{n\to\infty}\phi_{X_n}(t) = \exp\left\{\lambda\left(e^{it}-1\right)\right\}$ and this is easily recognizable as the c.f. of a Poisson(λ) distribution, introduced in **Example 27** as the c.f. of Poisson(λ) r.v. is given by

$$\phi(t) = e^{-\lambda}\sum_{x=0}^{\infty} e^{itx}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^{it}\right)^x}{x!} = e^{-\lambda}e^{\lambda e^{it}} = \exp\left\{\lambda\left(e^{it}-1\right)\right\}.$$

Thus we have shown that the c.f. of Binomial(n, p_n) sequence with the property $\lim_{n\to\infty} np_n = \lambda$ converges to the c.f. of a Poisson(λ) distribution. Therefore by the Weak Convergence Theorem Binomial$(n, p_n) \xrightarrow{\mathcal{D}}$ Poisson(λ) provided $\lim_{n\to\infty} np_n = \lambda$. $\bigtriangledown$

Problems

3.1. A company has launched 4 products, and let the probability of any one of the products being successful is 0.3. Assume that the products behave independently of each other. Let the Number of Successful Products be denoted by X. Answer the following:

a. Plot the probability mass function of X.

b. What is the expected value of X?

c. What is the most likely value of X?

d. Plot the c.d.f. of X and find the median and IQR of X.

e. Find the standard deviation of X.

3.2. The shop floor of a factory has 4 machines. At any given point of time the probabilities of the 4 machines going out of order are 0.05, 0.06, 0.08 and 0.1 respectively. Since the machines operate independently of each other, it may be assumed that the break down of machine-i is statistically independent of break down of machine-j for $i \neq j$, $i, j = 1, \ldots, 4$. Let X denote the number of machines which are out of order at any given point of time. Answer the following:

a. Find the p.m.f. of X.

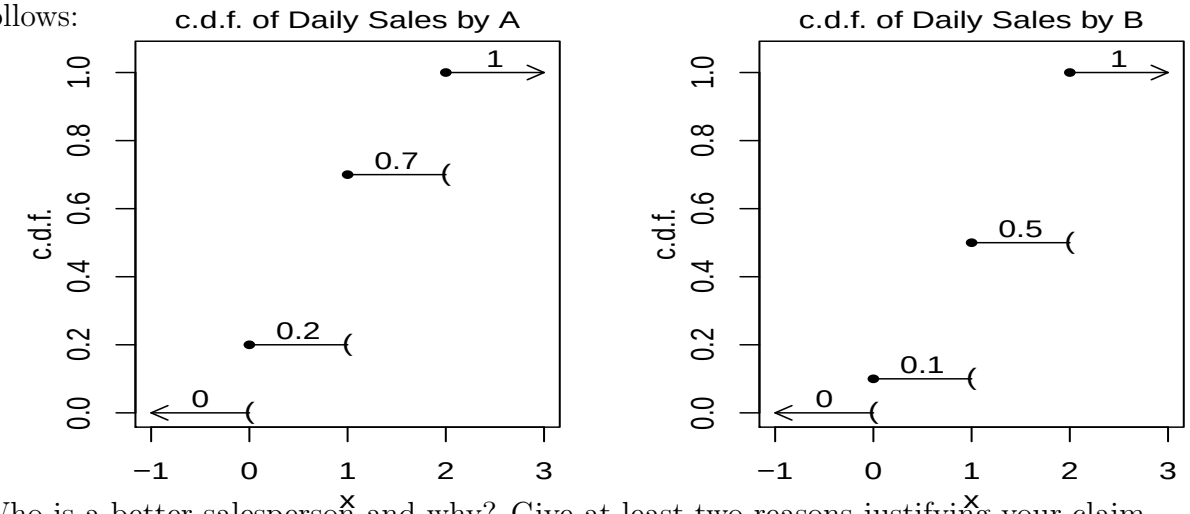
b. Find the c.d.f. of X.

c. What is the probability that at least one machine is out of order at any given point of time?

d. At any given point of time how many machines are most likely to be out of order? How many machines would you *expect* to be out of order at any given point of time?

e. Find the standard deviation of X and interpret its value.

3.3. The c.d.f.'s of the number of cars sold on a given day by sales-persons A and B are as follows:



Who is a better salesperson and why? Give at least two reasons justifying your claim.

3.4. The state of a stock at a given point of time t, say X_t, can be in one of the three states *viz.* below par (A), at par (B), or above par (C). The conditional probabilities of X_{t+1}, the state of the stock at time $t+1$, transiting to one of these three states, given X_t, its state at time t, are summarized in the following table:

$P(X_{t+1}|X_t)$

$X_{t+1} \rightarrow$ $X_t \downarrow$	A	B	C
A	0.7	0.2	0.1
B	0.3	0.5	0.2
C	0.3	0.3	0.4

Find the equilibrium (marginal) probability distribution of X_t *i.e.* that distribution of X_t which yields the same distribution for X_{t+1}.
(Motivation: It can be shown that, if the behavior of the stock is Markovian, then eventually the state of the stock with the above transition probabilities will have this *equilibrium* distribution.)

3.5. The number of days it takes to dear the bill of payments of an external supplier is unimodal and has a mean of 2 days, median of 3 days and an SD of 0.5 days. Answer the following:

a. Comment on the symmetry of the distribution.

b. What upper limit of the number of days should you quote to an external supplier who has just submitted his bill for payments, so that you are 95% certain that his bill would get cleared by then?

c. Sketch an approximate shape of the distribution, showing the scale of abscissa.

3.6. The cost of manufacturing an IC chip is Rs. 100 per unit up to 1000 units and it is Rs. 75 afterwards. Suppose the monthly demand for the chip, X (say), has the p.d.f. of X, $f(x) = \begin{cases} 0.001e^{-x/1000} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$. The chip is priced at Rs. 125 per unit and are manufactured in lots of the expected monthly demand of 1000. Answer the following:

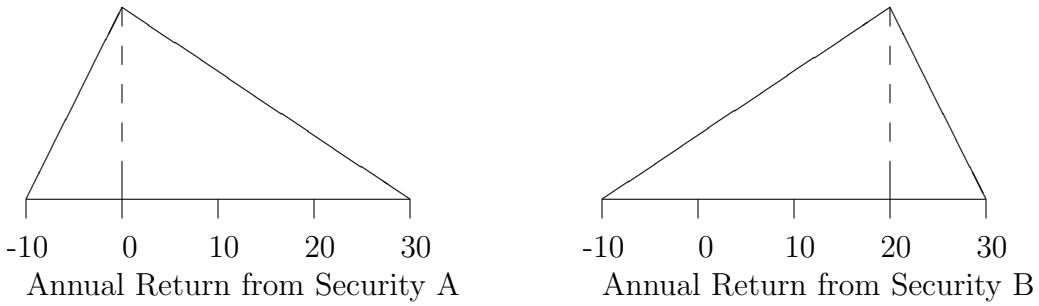
a. What is the probability of incurring a financial loss from the manufacture of the chip in a given month?

b. What can be done to minimize the loss, other than increasing the price?

c. Find the probability distribution of the monthly profit from the manufacture of the chip.

d. What is the expected monthly profit from the manufacture of the chip?

3.7. The p.d.f.'s of Annual Returns (the percentage of gain from an investment) from Securities A and B fare as follows:



Answer the following:

a. Show that Security B is a better investment than Security A.

b. If Rs. 1000 is invested in Security A for one year what is the probability that the investment will return more than Rs. 200 by the end of the year?

c. Find the distribution of bi-annual return (the percentage of gain) from Security B, assuming that an investment is compounded annually.

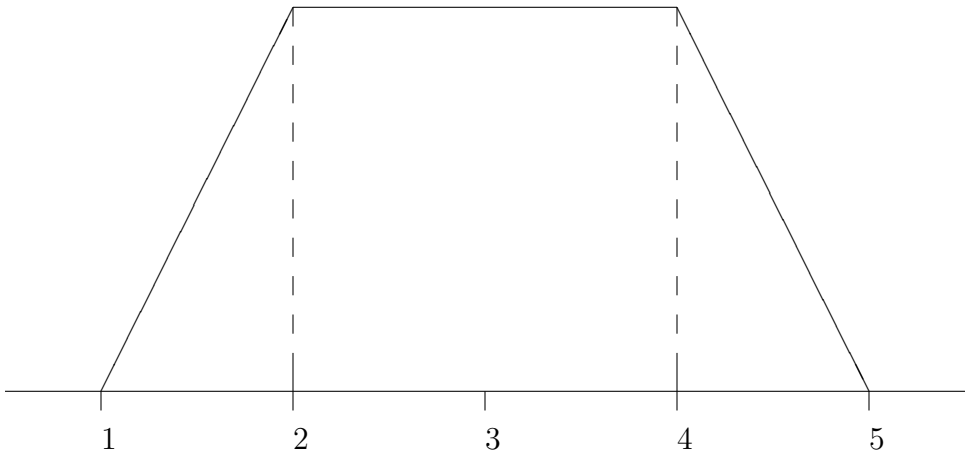
3.8. Minimum selling price acceptable to the owner of an apartment, which is up for sale, say X in lakhs of Rs., has the p.d.f. $f_x(x) = \begin{cases} c\, e^{-\frac{1}{2}(x-8)} & \text{if } x > 8 \\ 0 & \text{otherwise} \end{cases}$; while the maximum price, which a potential buyer is willing to pay, say Y, also in lakhs of Rs., is uniformly distributed over 6 to 10. Assume that X and Y are independent of each other. Answer the following:

a. Find the value of c.

b. What is *expected* minimum selling price of the owner of the apartment?

c. What is the probability of the apartment getting sold?

3.9. The probability density function of X, denoting the amount of time (in hours) it takes for a file to move from point A to point B in an organization is as follows:



Answer the following:

a. Find the height of the trapezium.

b. Find the c.d.f. of X.

c. What is the probability that it will take at least 4 hours for a file to move from point A to point B?

d. What is the probability that it will take between 1.5 to 3.5 hours for a file to move?

e. Find the mean and median amount of time it takes for a file to move.

f. Within how many hours 90% of the files have moved from point A to point B?

g. 5 files have been initiated from point A at 10 AM. Let Y denote the number of these reaching point B by 12 noon.

i. Find the p.m.f. of Y.

ii. What is the most likely number (out of 5) of files reaching point B by 12 noon?

iii. How many (of the 5) do you *expect* to reach point B by 12 noon? Interpret the *expected* value.

3.10. Let X be a positive continuous random variable having p.d.f. $f_X(x)$. Find a formula for the p.d.f. of $\frac{1}{X+1}$.

3.11. Agent 001 is trapped between two narrow abysmal walls. He swung his gun around in a vertical circle touching the walls and fired a wild (random) shot. Assume that the angle which his pistol makes with the horizontal is uniformly distributed between 0^0 and 90^0. Find the distribution of the height where the bullet landed and its mean.

3.12. It has been empirically found that the productivity of the top management (measured in terms of their contribution towards the growth of the company) Y, is related to their annual salary (in lakhs of Rs.) X, through the equation $Y = -X^2 + 16X - 45$. In a certain large corporate if the annual salary of the top management is uniformly distributed between Rs.5 and Rs.11 lakhs, find the distribution of their productivity.

3.13. The p.d.f. of the radius of buoys manufactured by a process is triangular, centered at a and spread over $[a-b, a+b]$ for some $a > 0$ and $b > 0$. Find the distribution of the volume of the manufactured buoys, assuming that they are perfect spheres.

3.14. Suppose the supply curve of a product is given by the equation $P = Q^2 e^{0.1*Q}$ where P denotes the Price and Q denotes the Quantity supplied. The p.d.f. of the Price of the product is modeled by $f_P(p) = \begin{cases} 0.1e^{-0.1p} & \text{if } p > 0 \\ 0 & \text{otherwise} \end{cases}$. Find the distribution of supply.

3.15. The joint p.m.f. of age, X, and the number of completed projects, Y, by a group of trainee engineers during the training period is as follows:

$Y \rightarrow$ $X \downarrow$	0	1	2	3
22	4	6	3	2
25	1	4	5	2
28	2	4	2	2

Answer the following:

a. What is the probability that a randomly chosen trainee is 25 year or older and has completed at least two projects?

b. What is the probability that a 25 year or older trainee has completed at least two projects?

c. If a randomly chosen trainee has completed at least two projects, what are the **(i)** most likely, and **(ii)** expected, ages of the trainee?

d. Find the regression of Y on X and plot it. Write your conclusion about the way age affects the number of projects completed by the trainees.

e. Find the correlation coefficient between X and Y and interpret its value.

3.16. The summary of complaints, received by a dealer of a particular automobile company, classified according to the model and the type of problem is as follows:

Problem $\rightarrow$ Model $\downarrow$	Engine	Transmission	Brake	Suspension	Body	Other
Small	8	14	6	2	4	1
Luxury	1	4	3	1	0	1
Heavy	4	8	2	3	1	2

Above table is based on 500 Small, 100 Luxury and 200 Heavy vehicles sold by the dealer so far. Answer the following:

a. Which model is most problem prone?

b. Which component, irrespective of the model, requires attention?

c. If a vehicle has a Suspension problem, what is the probability that it is Heavy?

d. What is the probability that a Luxury car has problem with its Brake?

e. What is the probability that the problem of a problematic Heavy vehicle is classified as Other?

f. What is the probability that a car manufactured by the company is Small and has Transmission problem?

g. What is the probability that a car has Body problem?

Serious problems are those with the Engine and the Transmission.

h. If a car is brought in with a Serious problem what is its most likely model?

i. What is the probability that a Small or a Luxury car has Serious problem?

3.17. The number of sales-persons in a company, which assembles personal computers, when classified according to their Sales Record (median number of machines sold per day) and Experience (rounded to the nearest number of years) is as follows:

Sales Record → Experience ↓	0	1	2	3
0	4	3	3	0
1	2	4	4	2
2	1	2	4	1

Answer the following:

a. What is the probability that a randomly chosen sales-person from the company has at least one year experience?

b. What is the probability that a randomly chosen sales-person from the company sells at least one machine on a given day?

c. What is the probability that a randomly chosen sales-person from the company has at least one year experience and sells at least one machine on a given day?

d. What is the probability that on a given day, a sales-person with one or more years of experience sells at least two machines?

e. What is the probability that a sales-person selling two or more machines on a given day has at least one year experience?

f. If a sales-person sells at least one machine on a given day, what is the most likely number of years of experience the sales-person has got?

g. Find and plot the regression of the Sales Record on Experience.

h. Find the correlation coefficient between the Sales Record and Experience and interpret its value.

i. Does the variability of the Sales Record remain unchanged for changing Experience?

3.18. The joint distribution of number of education years (X) and monthly salaries (Y in thousands of Rs.) of managers in the IT industry with 2 years of experience is as follows:

$X \rightarrow$ $Y \downarrow$	15	16	17	≥ 18
25-35	0.08	0.06	0.04	0.02
35-45	0.05	0.36	0.24	0.05
45-55	0.02	0.03	0.02	0.03

a. Draw the p.d.f. of monthly salaries of managers in the IT industry with 2 years of experience.

b. What proportion of managers in the IT industry with 2 years of experience and drawing a salary of at least Rs.35,000 per month have at least 17 years of education?

c. Show that as far as the salary is concerned it is immaterial whether a manager in the IT industry with 2 years of experience have 16 or 17 years of education.

d. Show that managers in the IT industry with 2 years of experience and at least 18 years of education earn more than those with 15 years of education.

e. Show that *on an average* managers in the IT industry with 2 years of experience tend to earn more with their number of years of education.

3.19. A company is registered in three different stock exchanges A, B and C, say. On any given working day, let X and Y denote the proportion of shares changing hands in A and B respectively. Assume that the joint distribution of (X, Y) is uniform on their natural domain. Answer the following.

a. What is the probability that on any given working day more than 50% of the shares change hands in exchange A?

b. On any given working day, what proportion of shares do you expect to change hands in exchange C?

c. Find the correlation coefficient between X and Y and interprete its value.

3.20. Consider the problem of allocating 5 trainees at random to 4 regional head offices. Let N denote the number of head offices not receiving any trainee and X_1 denote the number of trainees allocated to city 1. Answer the following:

a. Find the joint p.m.f. of N and X_1.

b. Find the marginal p.m.f.'s N and X_1.

c. Find the conditional p.m.f.'s N and X_1.

d. Find the correlation coefficient between N and X_1.

e. Find the two regression functions.

3.21. Consider a tied match-point between players A and B in a tennis match, a situation in which from that point onwards the player to win two successive points, wins the match. Let p denote the probability of A winning a point (at any given point) and the wins and losses at successive points be independent. Let the number of points (services) that are played to arrive at a winner be denoted by X. Answer the following:

a. Show that the p.m.f. of X is given by $p(x) = \begin{cases} (p^2+q^2)(pq)^{n-1} & \text{if } x = 2n \\ (pq)^n & \text{if } x = 2n+1 \end{cases}$, for $n = 1, 2, \ldots$, where $q = 1 - p$=P(B winning a point). Also check that $p(x)$ is indeed a legitimate p.m.f.

b. Show that the probability of A winning is $\frac{p^2(1+q)}{1-pq}$.

3.22. A travel portal among other things, sells vacations to its members. For this travel portal, for each of its members, let X denote the number of vacations the member had taken last year, and Y denote the average price (in Rs.) per vacation for that member. The p.m.f. of X is given by

x	0	1	2	3	4
$p(x)$	0.6	0.15	0.1	0.1	0.05

. For $x \neq 0$ given $X = x$, Y has the conditional p.d.f. $f(y|x) = \begin{cases} 0.001xe^{-0.001x(y-1000)} & \text{if } y \geq 1000 \\ 0 & \text{otherwise} \end{cases}$ (this is because the minimum price of a vacation is Rs.1000). Answer the following:

a. Find the expected average price per vacation paid by a member, who has taken at least one vacation.

b. What is the probability of a member spending a total of more than Rs.2500 on vacation last year from that portal?

c. Assuming that the same pattern continues this year, what would be the most likely number of vacations that a member would take this year, given that she has just booked a vacation of Rs.1500?

3.23 The joint p.m.f. of age, X, and the number of completed projects, Y, by a group of trainee engineers during the training period is as follows:

$Y \rightarrow$ $X \downarrow$	0	1	2	3
22	4	6	3	2
25	1	4	5	2
28	2	4	2	2

Answer the following:

a. Find the regression of Y on X and plot it. Write your conclusion about the way age affects the number of projects completed by the trainees. **[10]**

b. Find the correlation coefficient between X and Y and interpret its value. **[10]**

3.24. For a product like cell-phone the cost of production increases as the number of features increases. It is postulated that the number of features that a future cell-phone will contain, say X, is going to have a Poisson distribution with mean λ, that is X will have a p.m.f. $p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$. Now given $X = x$, that is a cell-phone having x number of features, the cost of production, say Y, which is a non-negative continuous random variable, is envisaged to have the p.d.f. $f(y|X=x) = \begin{cases} e^{-\{(y-\theta^{x+1}+\phi^{x/2})/\phi^{x/2}\}}/\phi^{x/2} & \text{if } y > \theta^{x+1} - \phi^{x/2} \\ 0 & \text{otherwise} \end{cases}$ for some $\theta > 1$ and $\phi > 0$. Find the correlation coefficient of X and Y.

[*Hint:* A random-variable Y having a p.d.f. $f(y) = \begin{cases} e^{-\{(y-\psi)/\tau\}}/\tau & \text{if } y > \psi \\ 0 & \text{otherwise} \end{cases}$ is said to have a two parameter exponential distribution with location parameter ψ and scale parameter τ or rate $1/\tau$. For such a random variable, $E[Y] = \psi + \tau$ and $V[Y] = \tau^2$.]

3.25. In a large R&D lab, let X and Y respectively denote the number of domestic and foreign patents filed by a scientist in any given year. Based on the past data (X, Y) is found to have the following joint p.m.f.:

$Y \rightarrow$ $X \downarrow$	0	1	2
0	0.02	0.20	0.05
1	0.15	0.25	0.13
2	0.10	0.15	0.05

Answer the following:

a. Show that the number of domestic patents filed by a scientist is larger than that of the number of foreign patents.

b. Find the median of the total number of patients filed by a scientist in a given year.

c. Find the correlation coefficient of the number of domestic and foreign patents filed by a scientist and interpret its value.

d. Find the regression of Y on X.

3.26 Let X denote the Bid Price of a Buy order in 100's of Rs., and Y denote the number of hours it takes for the order to get executed, for a particular (not very liquid) security, in an on-line order driven market (such as BSE or NSE). Based on empirical observations, the random vector (X, Y) appears to have a joint probability density function

$$f_{X,Y}(x,y) = \frac{1}{2}x^2 \exp\left\{-\left(\frac{y}{2}+1\right)x\right\} I_{[x>0,\, y>0]}(x,y),$$

where $I_A(x,y)$ is the indicator function of the set $A \subseteq \Re^2$ *i.e.* $I_A(x,y) = \begin{cases} 1 & \text{if } (x,y) \in A \\ 0 & \text{otherwise} \end{cases}$.
Answer the following:

a. Show that X and Y are not independent.

b. What is the probability that it takes at least an hour for a Buy order with a Bid Price of Rs.200 to get executed?

c. Find the regression of Y on X.

d. Qualitatively describe how X is affecting Y.

3.27. Minimum selling price acceptable to the owner of an apartment, which is up for sale, say X in lakhs of Rs., has the p.d.f. $f_x(x) = \begin{cases} c\, e^{-\frac{1}{2}(x-8)} & \text{if } x > 8 \\ 0 & \text{otherwise} \end{cases}$; while the maximum price, which a potential buyer is willing to pay, say Y, also in lakhs of Rs., is uniformly distributed over 6 to 10. Assume that X and Y are independent of each other. Answer the following:

a. Find the value of c.

b. What is *expected* minimum selling price of the owner of the apartment?

c. What is the probability of the apartment getting sold?

Appendix 3.A: Properties of the Distribution Function

Here we collect together a few useful properties of the c.d.f. $F(x) = P[X \leq x]$ of an arbitrary r.v. X, which could be discrete, continuous or a combination of both.

Property 1 (a): $\lim_{x\to-\infty} F(x) = 0$

Proof: Take and fix any sequence of real numbers $\{x_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} x_n = -\infty$. Let $A_n = \{\omega \in \Omega : X(\omega) \leq x_n\}$. Then $A_n \downarrow \phi$, the null set. Thus,

$$
\begin{aligned}
&\lim_{x\to-\infty} F(x) \\
&= \lim_{n\to\infty} F(x_n) && \text{(by definition of limit)} \\
&= \lim_{n\to\infty} P[X \le x_n] && \text{(by definition of a c.d.f.)} \\
&= \lim_{n\to\infty} P(A_n) && \text{(by definition of } A_n\text{)} \\
&= P(\lim_{n\to\infty} A_n) && \text{(by the continuity property of the Probability function)} \\
&= P(\phi) && \text{(as } \lim_{n\to\infty} A_n = \phi\text{)} \\
&= 0
\end{aligned}
$$

$\bigtriangledown$

Property 1 (b): $\lim_{x\to\infty} F(x) = 1$

Proof: Take and fix any sequence of real numbers $\{x_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} x_n = \infty$. Let $A_n = \{\omega \in \Omega : X(\omega) \le x_n\}$. Then $A_n \uparrow \Omega$. Thus,

$$
\begin{aligned}
&\lim_{x\to\infty} F(x) \\
&= \lim_{n\to\infty} F(x_n) && \text{(by definition of limit)} \\
&= \lim_{n\to\infty} P[X \le x_n] && \text{(by definition of a c.d.f.)} \\
&= \lim_{n\to\infty} P(A_n) && \text{(by definition of } A_n\text{)} \\
&= P(\lim_{n\to\infty} A_n) && \text{(by the continuity property of the Probability function)} \\
&= P(\Omega) && \text{(as } \lim_{n\to\infty} A_n = \phi\text{)} \\
&= 1
\end{aligned}
$$

$\bigtriangledown$

Property 2: $F(x)$ is a monotonically increasing function of x *i.e.* $x_1 < x_2 \Rightarrow F(x_1) \le F(x_2)$.

Proof: For $n = 1, 2$ let $A_n = \{\omega \in \Omega : X(\omega) \le x_n\}$. Then since $x_1 < x_2$, $A_1 \subseteq A_2$ and the result follows from the definition of c.d.f. and the monotonicity property of the Probability function. $\bigtriangledown$

Property 3: $F(x)$ is a right-continuous function *i.e.* $F(x+) = \lim_{y\to x+} F(y) = F(x)\ \forall x \in \Re$.

Proof: Fix any $x \in \Re$. Next take and fix any decreasing sequence of real numbers $\{x_n\}_{n=1}^{\infty}$ such that $x_n \downarrow x$ as $n \to \infty$. Let $A_n = \{\omega \in \Omega : X(\omega) \le x_n\}$ and $A = \{\omega \in \Omega : X(\omega) \le x\}$. Then A_n is a decreasing sequence of sets with $\lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n = A$. For checking the last equality, take any $\omega \in \cap_{n=1}^{\infty} A_n$. If $X(\omega) > x$, $\exists N \in \mathcal{P} \ni x_n < X(\omega)$ implying $\omega \notin A_n$ $\forall n \ge N$ and thus $\omega \notin \cap_{n=1}^{\infty} A_n$. Hence $\omega \in \cap_{n=1}^{\infty} A_n \Rightarrow X(\omega) \le x \Rightarrow \omega \in A$ proving that $\cap_{n=1}^{\infty} A_n \subseteq A$. Since $A \subseteq A_n\ \forall n \in \mathcal{P}$, $A \subseteq \cap_{n=1}^{\infty} A_n$ proving that $\cap_{n=1}^{\infty} A_n = A$. Now

$$
\begin{aligned}
&\lim_{y\to x+} F(y) \\
&= \lim_{n\to\infty} F(x_n) && \text{(by definition of limit)} \\
&= \lim_{n\to\infty} P[X \le x_n] && \text{(by definition of a c.d.f.)} \\
&= \lim_{n\to\infty} P(A_n) && \text{(by definition of } A_n\text{)} \\
&= P(\lim_{n\to\infty} A_n) && \text{(by the continuity property of the Probability function)} \\
&= P(A) && \text{(as } \lim_{n\to\infty} A_n = A\text{)} \\
&= P[X \le x] && \text{(by definition of } A\text{)} \\
&= F(x) && \text{(by definition of } F(x)\text{)}
\end{aligned}
$$

$\bigtriangledown$

Property 4: $P[X < x] = \lim_{y \to x-} F(y) = F(x-)$, the left-hand limit of $F(\cdot)$ at x.
Proof: Fix any $x \in \Re$. Next take and fix any increasing sequence of real numbers $\{x_n\}_{n=1}^{\infty}$ such that $x_n \uparrow x$ as $n \to \infty$. Let $A_n = \{\omega \in \Omega : X(\omega) \leq x_n\}$ and $A = \{\omega \in \Omega : X(\omega) < x\}$. Then A_n is an increasing sequence of sets with $\lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n = A$. For checking the last equality, take any $\omega \in \cup_{n=1}^{\infty} A_n$. Then $\exists n \in \mathcal{P} \ni X(\omega) \leq x_n < x \Rightarrow \omega \in A$ proving that $\cup_{n=1}^{\infty} A_n \subseteq A$. Now take any $\omega \in A$. Since $X(\omega) < x$ and $x_n \uparrow x$, $\exists N \in \mathcal{P} \ni x_n > X(\omega)$ $\forall n \geq N$ implying $\omega \in A_n$ $\forall n \geq N$ and thus $\omega \in \cup_{n=1}^{\infty} A_n$ proving that $A \subseteq \cup_{n=1}^{\infty} A_n$. Therefore $\cup_{n=1}^{\infty} A_n = A$. Now

$$
\begin{aligned}
&\lim_{y \to x-} F(y) \\
&= \lim_{n\to\infty} F(x_n) && \text{(by definition of limit)} \\
&= \lim_{n\to\infty} P[X \leq x_n] && \text{(by definition of a c.d.f.)} \\
&= \lim_{n\to\infty} P(A_n) && \text{(by definition of } A_n\text{)} \\
&= P(\lim_{n\to\infty} A_n) && \text{(by the continuity property of the Probability function)} \\
&= P(A) && \text{(as } \lim_{n\to\infty} A_n = A\text{)} \\
&= P[X < x] && \text{(by definition of } A\text{)}
\end{aligned}
$$

$\bigtriangledown$

From ***Property 4*** and the definition of c.d.f. it follows that $P[X = x] = P[X \leq x] - P[X < x] = F(x) - F(x-)$. At any point $x \in \Re$ thus $F(x) - F(x-)$ gives the quantum of jump the c.d.f. experiences which exactly equals the probability mass given at the point x. This observation is needed in understanding the proof of the next property.

Property 5: Let $D \subseteq \Re$ denote the set of points where $F(\cdot)$ is discontinuous *i.e.* $x \in D \Rightarrow F(x-) = \lim_{y\to x-} F(y) \neq F(x) = \lim_{y\to x+} F(y) = F(x+)$. The set D is countable.
Proof: For $n \in \mathcal{P}$ let $D_n = \left\{x \in \Re : F(x) - F(x-) \geq \frac{1}{n}\right\}$. D_n can have at most n elements, otherwise the sum of the probability masses of the distinct elements in D_n would exceed 1. Since the maximum amount of jump the c.d.f. can face is 1 in which case such a point is contained in D_1, and for any other amount of jump $\epsilon > 0$ at a point x $\exists n \ni \frac{1}{n} < \epsilon$ implying $x \in D_n$, $D = \cup_{n=1}^{\infty} D_n$. Thus D, the set of points of discontinuity of $F(\cdot)$, is a countable union of finite sets, and is thus countable. $\bigtriangledown$

Property 6: $F(x) = F_1(x) + F_2(x)$ where $F_1(x)$ is a step function and $F_2(x)$ is everywhere continuous.
Proof: Let D denote the set of points of discontinuity of $F(\cdot)$ as in ***Property 5***. Then since D is countable, let $D = \{x_1, x_2, \ldots\}$. Define $F_1(x) = \sum_{x_i \in D : x_i \leq x} [F(x_i) - F(x_i-)]$. Then obviously $F_1(x)$ is a step function with jumps at the points $\{x_1, x_2, \ldots\}$. Let $F_2(x) = F(x) - F_1(x)$. If $F(\cdot)$ is continuous at x then so is $F_1(\cdot)$ and hence $F_2(\cdot)$, and for $x_i \in D$ it is a trivial matter to see that $F_2(x_i-) = F_2(x_i)$. $\bigtriangledown$

Based on these properties now a clear picture of the nature of a general r.v. emerges, which may be summarized as follows.

1. At any point $x \in \Re$, $F(\cdot)$ is either continuous or experiences an upward jump.
2. If $F(\cdot)$ is continuous at a point x, then $F(x) = F(x-)$ and therefore $P[X = x] = 0$.

3. If $F(\cdot)$ gets a jump at point x, then there is a positive probability mass at x whose value equals the quantum of jump $F(x) - F(x-)$.
4. If $F(\cdot)$ is flat or constant on any interval $(a, b]$ then $P[a < X \leq b] = P[\{\omega \in \Omega : X(\omega) \leq b\} - \{\omega \in \Omega : X(\omega) \leq a\}] = P[X \leq b] - P[X \leq a]$ (since $A = \{\omega \in \Omega : X(\omega) \leq a\} \subseteq \{\omega \in \Omega : X(\omega) \leq b\} = B$, $P(B-A) = P(B) - P(A)) = F(b) - F(a) = 0$.
5. The number of points where X can have a positive probability mass is countable, justifying the definition of a discrete r.v.
6. It is enough to study just the discrete and continuous r.v. as any r.v. can be decomposed into having just these two components.

Finally as in point 4 above, let us list down the formulæ for computing probabilities of all types of events involving a r.v. using its c.d.f.. The proofs follow from the definition of $F(\cdot)$, ***Property 4*** and arguments similar to the one in point 4 above.

1. $P[X \leq a] = F(a)$.
2. $P[X > a] = 1 - F(a)$.
3. $P[a < X \leq b] = F(b) - F(a)$
4. $P[X < a] = F(a-)$
5. $P[X \geq a] = 1 - F(a-)$
6. $P[X = a] = F(a) - F(a-)$
7. $P[a \leq X \leq b] = F(b) - F(a-)$
8. $P[a < X < b) = F(b-) - F(a)$
9. $P[a \leq X < b] = F(b-) - F(a-)$

Appendix 3.B: Properties of Moments and Related Quantities

Here we collect together some useful properties of Expectation, Variance, Covariance and Correlation Coefficient. Since the formulæ for the moments are different for the discrete and continuous cases, we shall prove these results only for the continuous case. The proofs for the discrete case are similar, with the p.d.f. replaced by p.m.f. and the integrals replaced by summation. Thus let $f(x)$ denote the p.d.f. of a continuous r.v. X.

Properties of Expectation

Property E1: For constants a and b, $E[a + bX] = a + bE[X]$
Proof:

$$
\begin{aligned}
&E[a+bX] \\
&= \int_{-\infty}^{\infty}(a+bx)f(x)dx \\
&= a\int_{-\infty}^{\infty}f(x)dx + b\int_{-\infty}^{\infty}xf(x)dx \\
&= a+bE[X] \qquad \bigtriangledown
\end{aligned}
$$

Property E2: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$
Proof: Let $Y = g(X)$ be a many-to-one function. Then $f_Y(y)$, the p.d.f. of Y is as given in (6) and thus by definition

$$
\begin{aligned}
&E[g(X)] = E[Y] \\
&= \int_{-\infty}^{\infty} yf_y(y)dy \\
&= \sum_{i=1}^{k}\int_{-\infty}^{\infty} I[y\in\mathcal{Y}_i]yf\left(g_i^{-1}(y)\right)\left|\frac{d}{dy}g_i^{-1}(y)\right|dy \quad \text{(by (6))} \\
&= \sum_{i=1}^{k}\int_{\mathcal{Y}_i} yf\left(g_i^{-1}(y)\right)\left|\frac{d}{dy}g_i^{-1}(y)\right|dy \\
&= \sum_{i=1}^{k}\int_{\mathcal{X}_i} g(x)f(x)\left|\frac{d}{d(g(x))}x\right|\left|\frac{d}{dx}g(x)\right|dx \quad \text{(by substituting } y=g(x) \text{ we get that, for } y\in\mathcal{Y}_i, \\
&\quad g_i^{-1}(g(x)) = x \text{ and } x = g_i^{-1}(y) \text{ so that } y\in\mathcal{Y}_i \Rightarrow x\in\mathcal{X}_i \text{ and the rest follows from the} \\
&\quad \text{routine change of variable method for integration)} \\
&= \int_{-\infty}^{\infty} g(x)f(x)dx \quad (\text{as } \mathcal{X}_i\cap\mathcal{X}_j = \phi \text{ for } i\neq j \text{ and } \cup_{i=1}^{k}\mathcal{X}_i = \mathcal{X} = (-\infty,\infty)) \qquad \bigtriangledown
\end{aligned}
$$

For the next three properties, let $\boldsymbol{X} = (X_1, X_2)'$ be a random vector with joint p.d.f. $f(x_1, x_2)$. Let the marginal p.d.f. of X_i be denoted by $f_i(x_i)$ for $i = 1, 2$. Also let the conditional p.d.f. of $X_1|X_2 = x_2$ be denoted by $f_{1|2}(x_1|x_2)$.

Property E3: For constants c_1 and c_2, $E[c_1X_1 + c_2X_2] = c_1E[X_1] + c_2E[X_2]$
Proof:

$$
\begin{aligned}
&E[c_1X_1 + c_2X_2] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(c_1x_1 + c_2x_2)f(x_1,x_2)dx_1dx_2 \\
&= c_1\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1f(x_1,x_2)dx_1dx_2 + c_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_2f(x_1,x_2)dx_1dx_2 \\
&= c_1\int_{-\infty}^{\infty}\left\{x_1\int_{-\infty}^{\infty}f(x_1,x_2)dx_2\right\}dx_1 + c_2\int_{-\infty}^{\infty}\left\{x_2\int_{-\infty}^{\infty}f(x_1,x_2)dx_1\right\}dx_2 \\
&= c_1\int_{-\infty}^{\infty}x_1f_1(x_1)dx_1 + c_2\int_{-\infty}^{\infty}x_2f_2(x_2)dx_2 \\
&= c_1E[X_1] + c_2E[X_2] \qquad \bigtriangledown
\end{aligned}
$$

Property E4: If X_1 and X_2 are independent, $E[X_1X_2] = E[X_1]E[X_2]$
Proof: If X_1 and X_2 are independent, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Thus

$$
\begin{aligned}
& E[X_1X_2] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2f(x_1,x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2f_1(x_1)f_2(x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\left\{x_2f_2(x_2)\int_{-\infty}^{\infty} x_1f_1(x_1)dx_1\right\}dx_2 \\
&= E[X_1]\int_{-\infty}^{\infty} x_2f_2(x_2)dx_2 \\
&= E[X_1]E[X_2]
\end{aligned}
$$

▽

Property E5: $E\left[E[X_1|X_2]\right] = E[X_1]$
Proof:

$$
\begin{aligned}
& E\left[E[X_1|X_2]\right] \\
&= \int_{-\infty}^{\infty} E[X_1|X_2=x_2]f_2(x_2)dx_2 \\
&= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} x_1f_{1|2}(x_1|x_2)dx_1\right\}f_2(x_2)dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1\frac{f(x_1,x_2)}{f_2(x_2)}f_2(x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1f(x_1,x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\left\{x_1\int_{-\infty}^{\infty} f(x_1,x_2)dx_2\right\}dx_1 \\
&= \int_{-\infty}^{\infty} x_1f_1(x_1)dx_1 \\
&= E[X_1]
\end{aligned}
$$

▽

Properties of Covariance

We begin with a formula for $\mathrm{Cov}(X_1,X_2)$, in the same spirit of (1), that is simpler than ***Definition 14***.

$$
\begin{aligned}
& \mathrm{Cov}(X_1,X_2) \\
&= E\left[(X_1-E[X_1])(X_2-E[X_2])\right] \\
&= E\left[X_1X_2 - X_1E[X_2] - X_2E[X_1] + E[X_1]E[X_2])\right] \\
&= E\left[X_1X_2\right] - E\left[X_1E[X_2]\right] - E\left[X_2E[X_1]\right] + E\left[E[X_1]E[X_2]\right] \quad (\text{by } \textbf{\textit{E3}}) \\
&= E\left[X_1X_2\right] - E[X_2]E\left[X_1\right] - E[X_1]E\left[X_2\right] + E[X_1]E[X_2] \quad (\text{by } \textbf{\textit{E1}}) \\
&= E\left[X_1X_2\right] - E[X_1]E[X_2]
\end{aligned}
$$

Property C1: For a constant c, $\mathrm{Cov}(X,c)=0$
Proof:

$$
\begin{aligned}
&\mathrm{Cov}(X, c) \\
&\quad = E[cX] - E[c]E[X] \\
&\quad = cE[X] - cE[X] \\
&\quad = 0
\end{aligned}
$$
$\bigtriangledown$

Property C2: $\mathrm{Cov}(X, X) = V[X]$
Proof:

$$
\begin{aligned}
&\mathrm{Cov}(X, X) \\
&\quad = E[X \times X] - E[X]E[X] \\
&\quad = E[X^2] - (E[X])^2 \\
&\quad = V[X]
\end{aligned}
$$
$\bigtriangledown$

Property C3: For constants a, b, c and d, $\mathrm{Cov}(aX_1 + b, cX_2 + d) = ac\mathrm{Cov}(X_1, X_2)$.
Proof:

$$
\begin{aligned}
&\mathrm{Cov}(aX_1 + b, cX_2 + d) \\
&\quad = E\left[(aX_1 + b)(cX_2 + d)\right] - E[aX_1 + b]E[cX_2 + d] \\
&\quad = E[acX_1X_2 + adX_1 + bcX_2 + bd] - (aE[X_1] + b)(cE[X_2] + d) \quad (\text{by } \textbf{\textit{E1}}) \\
&\quad = acE[X_1X_2] + adE[X_1] + bcE[X_2] + bd - (acE[X_1]E[X_2] + adE[X_1]bcE[X_2] + bd) \\
&\quad = ac(E[X_1X_2] - E[X_1]E[X_2]) \\
&\quad = ac\mathrm{Cov}(X_1, X_2)
\end{aligned}
$$
$\bigtriangledown$

Property C4: $\mathrm{Cov}(X_1, X_2 + X_3) = \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, X_3)$
Proof:

$$
\begin{aligned}
&\mathrm{Cov}(X_1, X_2 + X_3) \\
&\quad = E\left[X_1(X_2 + X_3)\right] - E[X_1]E[X_2 + X_3] \\
&\quad = E[X_1X_2 + X_1X_3] - E[X_1](E[X_2] + E[X_3]) \\
&\quad = (E[X_1X_2] - E[X_1]E[X_2]) + (E[X_1X_3] - E[X_1]E[X_3]) \\
&\quad = \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, X_3)
\end{aligned}
$$
$\bigtriangledown$

Property C5: If X_1 and X_2 are independent then $\mathrm{Cov}(X_1, X_2) = 0$ but the converse is not true.
Proof:

$$
\begin{aligned}
&\mathrm{Cov}(X_1, X_2) \\
&\quad = E[X_1X_2] - E[X_1]E[X_2] \\
&\quad = E[X_1]E[X_2] - E[X_1]E[X_2] \quad (\text{by } \textbf{\textit{E4}}) \\
&\quad = 0
\end{aligned}
$$

Example 17 is an instance where $\mathrm{Cov}(X_1, X_2) = 0$ but X_1 and X_2 are not independent. $\bigtriangledown$

Properties of Variance

Property V1: For constants a and b, $V[a+bX] = b^2V[X]$
Proof:

$$\begin{aligned} & V[a+bX] \\ &= E[(a+bX)^2] - (E[a+bX])^2 \\ &= E[a^2+2abX+b^2X^2] - \left(a^2+2abE[X]+b^2E^2[X]\right) \quad (\text{by } \boldsymbol{E1}) \\ &= a^2+2abE[X]+b^2E[X^2]-a^2-2abE[X]-b^2E^2[X] \\ &= b^2\left(E[X^2]-E^2[X]\right) \\ &= b^2V[X] \end{aligned}$$

▽

Property V2: For constants a and b, $V[aX_1+bX_2] = a^2V[X_1]+2ab\text{Cov}(X_1,X_2)+b^2V[X_2]$
Proof:

$$\begin{aligned} & V[aX_1+bX_2] \\ &= E[(aX_1+bX_2)^2] - (E[aX_1+bX_2])^2 \\ &= E[a^2X_1^2+2abX_1X_2+b^2X_2^2] - \left(a^2E^2[X_1]+2abE[X_1]E[X_2]+b^2E^2[X_2]\right) \quad (\text{by } \boldsymbol{E1}) \\ &= a^2(E[X_1^2]-E^2[X_1])+2ab(E[X_1X_2]-E[X_1]E[X_2])+b^2(E[X_2^2]-E^2[X_2]) \\ &= a^2V[X_1]+2ab\text{Cov}(X_1,X_2)+b^2V[X_2] \end{aligned}$$

▽

Property V2 is extended for a linear combination of p r.v.'s as follows. Let $\boldsymbol{X}_{p\times 1} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ be a $p\times 1$ random vector and $\boldsymbol{\ell} = \begin{pmatrix} \ell_1 \\ \vdots \\ \ell_p \end{pmatrix}$ be a $p\times 1$ vector of constants. Then

$$E[\boldsymbol{\ell}'\boldsymbol{X}] = \boldsymbol{\ell}'E[\boldsymbol{X}] \quad \text{and} \quad V[\boldsymbol{\ell}'\boldsymbol{X}] = \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}$$

where $\boldsymbol{\Sigma}$ is a $p\times p$ matrix, called the variance-covariance or dispersion matrix of $\boldsymbol{X}$ which is also denoted by $D[\boldsymbol{X}]$. $\boldsymbol{\Sigma} = ((\sigma_{ij}))_{p\times p}$ and $\sigma_{ij} = \text{Cov}(X_i,X_j)$, such that by ***C2***, the diagonal elements of $\boldsymbol{\Sigma}$ are $V[X_i]$'s and the off-diagonal elements are $\text{Cov}(X_i,X_j)$'s. If $X_1,\ldots,X_p$ are independent of each other by ***C5***, $\sigma_{ij}=0$ for $i\neq j$ and thus $V[\boldsymbol{\ell}'\boldsymbol{X}] = \sum_{i=1}^p \ell_i^2V[X_i]$.

Property V3: $V[X_1] = E[V[X_1|X_2]] + V[E[X_1|X_2]]$
Proof:

$$\begin{aligned} & E[V[X_1|X_2]] + V[E[X_1|X_2]] \\ &= E\left[E[X_1^2|X_2]-E^2[X_1|X_2]\right] + E\left[E^2[X_1|X_2]\right] - (E[E[X_1|X_2]])^2 \\ &= E[X_1^2] - E\left[E^2[X_1|X_2]\right] + E\left[E^2[X_1|X_2]\right] - (E[X_1])^2 \quad (\text{by } \boldsymbol{E5}) \\ &= E[X_1^2]-E^2[X_1] \\ &= V[X_1] \end{aligned}$$

▽

Properties of Correlation Coefficient

Property R1: Correlation coefficient ρ between two random variables X_1 and X_2 is a number such that always $-1 \leq \rho \leq 1$ and $\rho = \pm 1$ if and only if $X_2 = a + bX_1$ with probability 1 for some constants a and b.
Proof: For any real number λ, $V[\lambda X_1 + X_2]$ is always non-negative. Thus

$$\begin{aligned}
&V[\lambda X_1 + X_2] \geq 0 \ \ \forall \lambda \in \Re \\
&\quad \Leftrightarrow \ \lambda^2 V[X_1] + 2\lambda \mathrm{Cov}(X_1, X_2) + V[X_2] \geq 0 \ \ \forall \lambda \in \Re \ \ (\text{by } \textbf{\textit{V2}}) \\
&\quad \Leftrightarrow \ 4\mathrm{Cov}^2(X_1, X_2) - 4V[X_1]V[X_2] \leq 0 \ \ (\text{if } ax^2 + bx + c, \text{ a quadratic in } x \text{ is non-negative} \\
&\qquad\qquad\qquad\qquad \forall x \in \Re \text{ then its discriminant } b^2 - 4ac \text{ must be less than or equal to } 0) \\
&\quad \Leftrightarrow \ \rho^2 \leq 1 \\
&\quad \Leftrightarrow \ -1 \leq \rho \leq 1
\end{aligned}$$

If $X_2 = a + bX_1$ for some constants a and b,

$$\begin{aligned}
&\rho \\
&\quad = \ \frac{\mathrm{Cov}(X_1, a + bX_1)}{\sqrt{V[X_1]V[a + bX_1]}} \\
&\quad = \ \frac{bV[X_1]}{|b|V[X_1]} \ \ (\text{by } \textbf{\textit{C2, C3}} \text{ and } \textbf{\textit{V1}}) \\
&\quad = \ \pm 1
\end{aligned}$$

On the other hand if $\rho = \pm 1$, then $\mathrm{Cov}^2(X_1, X_2) = V[X_1]V[X_2]$ and thus

$$\begin{aligned}
&V\left[X_2 - \frac{\mathrm{Cov}(X_1, X_2)}{V[X_1]} X_1\right] \\
&\quad = \ V[X_2] - 2\frac{\mathrm{Cov}^2(X_1, X_2)}{V[X_1]} + \frac{\mathrm{Cov}^2(X_1, X_2)}{V[X_1]^2} V[X_1] \ \ (\text{by } \textbf{\textit{V1}}) \\
&\quad = \ V[X_2] - 2V[X_2] + V[X_2] \\
&\quad = \ 0
\end{aligned}$$

Now if for some r.v. Y, $V[Y] = 0$, that implies $E[(Y - E[Y])^2] = 0$. Since $(Y - E[Y])^2$ is a non-negative quantity, its expectation can be 0 if and only if $(Y - E[Y])^2 = 0$ with probability 1, or if and only if $Y = E[Y]$, a constant, with probability 1. Thus $X_2 - \frac{\mathrm{Cov}(X_1, X_2)}{V[X_1]} X_1 = a$, a constant, with probability 1, and therefore $X_2 = a + bX_1$ with probability 1, for $b = \frac{\mathrm{Cov}(X_1, X_2)}{V[X_1]}$. ▽

Property R2: If X_1 and X_2 are independent then $\rho_{X_1, X_2} = 0$ but the converse is not true.
Proof: Follows immediately from ***C5***. ▽

Property R3: Correlation coefficient is a pure number whose absolute value does not depend on scale or origin shift *i.e.* $|\rho_{a+bX_1, c+dX_2}| = |\rho_{X_1, X_2}|$ for all constants a, b, c, $d \in \Re$.

Proof:

$$
\begin{aligned}
&|\rho_{a+bX_1,c+dX_2}| \\
&= \left| \frac{\mathrm{Cov}(a+bX_1, c+dX_2)}{\sqrt{V[a+bX_1]V[c+dX_2]}} \right| \\
&= \left| \frac{bd\mathrm{Cov}(X_1, X_2)}{|bd|\sqrt{V[X_1]V[X_2]}} \right| \quad (\text{by } \boldsymbol{C3} \text{ and } \boldsymbol{V1}) \\
&= |\rho_{X_1,X_2}|
\end{aligned}
$$