2.8 Bayes' Rule

We shall next introduce an important concept useful in conditional probability computations. Consider a finite or infinite sequence of "nonempty" events

$$\Pi_1 = \{E_n\}_n$$

in the probability space $(\Omega, \mathcal{B}, P)$, such that they give a partition of Ω, that is

$$E_i \cap E_j = \phi \text{ for } i \neq j \text{ and} \tag{2.8.1}$$

$$\Omega = \bigcup_n E_n \tag{2.8.2}$$

(Such a collection of events are also referred to as "**collectively exhaustive**").
For any set $A \in \mathcal{B}$ we have

$$\begin{aligned} A &= A \cap \Omega \\ &= A \cap \left(\bigcup_n E_n \right) \\ &= \bigcup_n (A \cap E_n) \\ \Longrightarrow & \\ P(A) &= P\left(\bigcup_n (A \cap E_n) \right) \\ &= \sum_n P\left(A \cap E_n\right) \text{ since } E_n \text{ are all disjoint} \\ &= \sum_n P(A|E_n)P(E_n) \end{aligned}$$

Thus we have

$$P(A) = \sum_n P(A|E_n)P(E_n) \tag{2.8.3}$$

This is called the "**Law of Total Probability**".
We further have

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

$$\begin{aligned} &= \frac{P(A|E_n)P(E_n)}{P(A)} \\ &= \frac{P(A|E_n)P(E_n)}{\sum_k P(A|E_k)P(E_k)} \text{ (using 2.8.3)} \end{aligned}$$

This is known as Bayes' Rule. Thus we have

Theorem 2.8.1 **Bayes' Rule**
Let $\{E_n\}_n$ be collectively exhaustive events in a probability space $(\Omega, \mathcal{B}, P)$. For any$A \in \mathcal{B}$ we have

$$P(E_n|A) = \frac{P(A|E_n)P(E_n)}{\sum_k P(A|E_k)P(E_k)} \tag{2.8.4}$$

Remark 2.8.1 Suppose $\{E_n\}_n$ and $\{F_n\}_n$ are two collectively exhaustive sets of events then we have from 2.8.4

$$P(E_n|F_m) = \frac{P(F_m|E_n)P(E_n)}{\sum_k P(F_m|E_k)P(E_k)} \text{ for every } m \text{ and } n \tag{2.8.5}$$

It is in this form that Bayes' Rule is used often in the computation of conditional probabilities.

2.9 Independence of Random Variables

The notion of independence of events induces a notion of independence of random variables. We shall first look at two discrete random variables X and Y on a probability space $(\Omega, \mathcal{B}, P)$ which take a finite number of values. Let the values taken by these two random variables be

$$\begin{aligned} \mathcal{R}_X &= \{x_1, x_2, x_3, \cdots, x_m\} \\ \mathcal{R}_Y &= \{y_1, y_2, y_3, \cdots, y_n\} \end{aligned}$$

Let us now consider the sets

$$\begin{aligned} E_i &= \{\omega \in \Omega : X(\omega) = x_i\} \\ F_j &= \{\omega \in \Omega : Y(\omega) = y_j\} \end{aligned}$$

Since X and Y are random variables the sets E_i and F_j are events. These events may or may not be independent. If these two events are independent for every i and j we say that the two random variables X and Y are independent. We can do the same thing even if the sets $\mathcal{R}_X$ or/and $\mathcal{R}_Y$ are infinite sequences. Thus we have

Definition 2.9.1 Two discrete random variables X, Y on a probability space $(\Omega, \mathcal{B}, P)$ are said to be independent if the events $\{\omega \in \Omega : X(\omega) = x_i\}$ and $\{\omega \in \Omega : Y(\omega) = y_j\}$ are independent for every $x_i \in \mathcal{R}_X$ and every $y_j \in \mathcal{R}_Y$. Using 2.7.2 we can also write this as follows:
Two discrete random variables X, Y on a probability space $(\Omega, \mathcal{B}, P)$ are said to be independent if

$$\begin{aligned} P(X = x_i \text{ and } Y = y_j) &= P(\{X = x_i\} \cap \{Y = y_j\}) \\ &= P(X = x_i) \times P(Y = y_j) \qquad (2.9.1) \end{aligned}$$

If p_X and p_Y are the pmfs of X and Y respectively then we can write the above definition as

Definition 2.9.2 Two discrete random variables X, Y on a probability space $(\Omega, \mathcal{B}, P)$ are said to be independent if

$$P(X = x_i \text{ and } Y = y_j) = p_X(x_i) \times p_Y(y_j) \qquad (2.9.2)$$

Analogously for any two general random variables we can define independence through the independence of the basic events $\{X \leq x\}$ and $\{Y \leq y\}$. We have the following

Definition 2.9.3 Two random variables X, Y on a probability space $(\Omega, \mathcal{B}, P)$ are said to be independent if

$$\begin{aligned} P(X \leq x \text{ and } Y \leq y) &= P(X \leq x) \times P(Y \leq y) \\ &= F_X(x) F_Y(y) \text{ for every } x, y \in \mathbb{R} \qquad (2.9.3) \end{aligned}$$

(where F_X and F_Y are the cdfs of X and Y respectively).

Remark 2.9.1 We say a collection $\{X_\alpha\}_{\alpha \in \mathcal{I}}$ of random variables on a probability space $(\Omega, \mathcal{B}, P)$ is independent if

$$P\left(\bigcap_{j=1}^{n} \{X_{\alpha_j} \leq x_j\} \right) = \prod_{j=1}^{n} F_{X_{\alpha_j}}(x_j)$$

for every positive integer $k \geq 2$, and for every $\alpha_j \in \mathcal{I}$ and every $x_j \in \mathbb{R}$.

2.10 Joint Distribution

In this section we shall study two or more random variables together. We shall first consider two discrete random variables. Consider two discrete random variables X and Y on a probability space $(\Omega, \mathcal{B}, \mathcal{P})$. Let

$$\begin{aligned} \mathcal{R}_X &= \{x_1, x_2, x_3 \cdots, x_m\} \\ \mathcal{R}_Y &= \{y_1, y_2, y_3, \cdots, y_n\} \end{aligned}$$

be the values taken by these random variables. Let p_X and p_Y be the pmfs of X and Y. Let

$$\begin{aligned} p_X(x_i) &= p_i \text{ for } 1 \leq i \leq m \text{ and} \\ p_Y(y_j) &= q_j \text{ for } 1 \leq j \leq n \end{aligned}$$

We now want to look at every point $\omega \in \Omega$ and analyse simulataneously X and Y at that point and we want to do this at every $\omega \in \Omega$. We do this as follows:
Let

$$\begin{aligned} E_i &= \{\omega \in \Omega : X(\omega) = x_i\} \\ F_j &= \{\omega \in \Omega : Y(\omega) = y_j\} \end{aligned}$$

Since X and Y are random variables, for every i and j, the sets E_i and F_j are events, (that is $E_i, F_j \in \mathcal{B}$), and hence the set $E_{ij} = E_i \cap F_j$ is also an event, (that is $E_{ij} \in \mathcal{B}$). We can therefore define its probability. We let

$$p_{ij} = P(E_{ij}) = P(E_i \cap F_j) = P(\omega \in \Omega : X(\omega) = x_i \text{ and } Y(\omega) = y_j)$$

Thus we get an $m \times n$ matrix

$$P_{XY} = (p_{ij})_{m \times n}$$

whose (i, j)th entry is the probability of the set of all those points where simultaneously X and Y take the values x_i and y_j respectively. Note that there may not be any point at which this simultaneous event takes place. In such a case $E_{ij} = \phi$ and $p_{ij} = P(E_{ij}) = 0$.
We shall now look at some simple examples:

Example 2.10.1 Let us consider the experiment of rolling a fair die., (with $\mathcal{B} = \mathcal{P}(\Omega)$). We have

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Conisder the random variables X and Y defined as follows:

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is even} \\ -1 & \text{if } \omega \text{ is odd} \end{cases}$$

$$Y(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is a prime number} \\ -1 & \text{if } \omega \text{ is not a prime number} \end{cases}$$

The corresponding pmfs are given below:

X	p_X
1	$\frac{1}{2}$
-1	$\frac{1}{2}$

Y	p_Y
1	$\frac{1}{2}$
-1	$\frac{1}{2}$

Note that X and Y take the same set of values with the same probabilities. (Such random variables are said to be identically distributed random variables). Let us now find the joint pmf. We have

$Y \rightarrow$ $X \downarrow$	$y_1 = -1$	$y_2 = 1$
$x_1 = -1$	$\frac{1}{6}$	$\frac{2}{6}$
$x_2 = 1$	$\frac{2}{6}$	$\frac{1}{6}$

Thus the joint pmf matrix is given by

$$P_{XY} = \begin{pmatrix} \frac{1}{6} & \frac{2}{6} \\ \frac{2}{6} & \frac{1}{6} \end{pmatrix}$$

Let us look at the row sums and column sums of this matrix. We have

$Y \rightarrow$ $X \downarrow$	$y_1 = -1$	$y_2 = 1$	*Row Sum*
$x_1 = -1$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{2}$
$x_2 = 1$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
Column Sum	$\frac{1}{2}$	$\frac{1}{2}$	1

These are called the "**marginal distributions**" of X and Y. Note that the Row Sums give the pmf P_X of X and the column sums give the pmf p_Y of Y

Example 2.10.2 Let us again consider the random experiment of rolling a fair die. Consider the random variables X and Y defined as follows:

$$
\begin{aligned}
X(\omega) &= \begin{cases} -1 & \text{if} \quad \omega \text{ is odd} \\ 1 & \text{if} \quad \omega \text{ is even} \end{cases} \\
Y(\omega) &= \begin{cases} -2 & \text{if} \quad \omega \leq 2 \\ 2 & \text{if} \quad \omega > 2 \end{cases}
\end{aligned}
$$

The pmfs of X and Y are as given below:

$X(\omega)$	$x_1 = -1$	$x_2 = 1$
$p(x_i)$	$\frac{1}{2}$	$\frac{1}{2}$

$Y(\omega)$	$y_1 = -2$	$x_2 = 2$
$p(x_i)$	$\frac{1}{3}$	$\frac{2}{3}$

The joint pmt is as given below:

$Y \rightarrow$ $X \downarrow$	$y_1 = -2$	$y_2 = 2$	*Row Sum*
$x_1 = -1$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
$x_2 = 1$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
Column Sum	$\frac{1}{3}$	$\frac{2}{3}$	1

The joint pmf matrix P_{XY} is given by

$$P_{XY} = \begin{pmatrix} \frac{1}{6} & \frac{1}{3} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}$$

Example 2.10.3 Consider two rvs X and Y on a probability space $(\Omega, \mathcal{B}, P)$ such that

$$\mathcal{R}_X = \mathcal{R}_y = \{-1, 0, 1\}$$

with the joint pmf given by

$Y \rightarrow$ $X \downarrow$	$y_1 = -1$	$y_2 = 0$	$y_3 = 1$
$x_1 = -1$	0	$\frac{1}{4}$	0
$x_2 = 0$	$\frac{1}{4}$	0	$\frac{1}{4}$
$x_3 = 1$	0	$\frac{1}{4}$	0

From the joint pmf we can find the pmfs p_X of X and p_y of Y using the marginal distributions. For the marginal distributions we find the row and column sums. We get

$Y \rightarrow$ $X \downarrow$	$y_1 = -1$	$y_2 = 0$	$y_3 = 1$	*Row Sums*
$x_1 = -1$	0	$\frac{1}{4}$	0	$\frac{1}{4}$
$x_2 = 0$	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
$x_3 = 1$	0	$\frac{1}{4}$	0	$\frac{1}{4}$
Column Sums	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Hence we get

X	p_X
$x_1 = -1$	$\frac{1}{4}$
$x_2 = 0$	$\frac{1}{2}$
$x_3 = 1$	$\frac{1}{4}$

and

Y	p_Y
$y_1 = -1$	$\frac{1}{4}$
$y_2 = 0$	$\frac{1}{2}$
$y_3 = 1$	$\frac{1}{4}$

Thus we see that we can easily find the individual pmfs from the joint pmf using the marginal distributions. However, it is not easy to get the joint pmf knowing only the individual pmfs.

We shall now see that the joint pmf is much easier to compute from the individual pmfs if the random variables are independent. Suppose X and Y are independent random variables on a probability space $(\Omega.\mathcal{B}, P)$, and

$$\begin{aligned} \mathcal{R}_X &= \{x_1, x_2, \cdots, x_m\} \\ \mathcal{R}_Y &= \{y_1, y_2, \cdots, y_n\} \end{aligned}$$

Let $p_{_X}$ and $p_{_Y}$ be the pmfs of X and Y respectively.

$$\begin{aligned} p_{_X}(x_i) &= p_i \text{ for } 1 \leq i \leq m \\ p_{_Y}(y_j) &= q_j \text{ for } 1 \leq j \leq n \end{aligned}$$

Then we have the joint pmf $P_{_{XY}} = (p_{ij})_{m\times n}$ where

$$\begin{aligned} p_{ij} &= P(\omega \in \Omega : X(\omega) = x_i \text{ and } Y(\omega) = y_j) \\ &= P(\{\omega \in \Omega : X(\omega) = x_i\} \cap \{\omega \in \Omega : Y(\omega_= y_j\}) \\ &= P(\{\omega \in \Omega : X(\omega) = x_i\}) \times P(\{\omega \in \Omega : Y(\omega_= y_j\}) \\ &\quad \text{(by using independence)} \\ &= p_{_X}(x_i) \times p_{_Y}(y_j) \\ &= p_i q_j \end{aligned}$$

By marginal distribution, we have

$$\begin{aligned} p_i &= i - th\ Row\ Sum \\ q_j &= j - th\ Column\ Sum \end{aligned}$$

Thus we see that if X and Y are independent then (for every i and j), the (i, j)-th entry in the joint pmf matrix is the product of the i-th Row Sum with the j-th Column Sum. Conversely if the (i, j)-th entry in the joint pmf matrix is the product of the i-th Row Sum with the j-th Column Sum (for every i and j) then the two random variables are independent.
The random variables in Example 2.10.2 the two random variables are independent since every entry in the joint pmf matrix is the product of the corresponding Row Sum and Column Sum. The random variables of Examples 2.10.1 and 2.10.3 are not independent.
While we used the pmfs to describe the joint pmf in the case of discrete random variables, we use the cdfs to describe the joint distribution of continuous

random variables. If X and Y are continuous random variables whose cdfs are respectively

$$\begin{aligned} F_X(x) &= P(X \leq x) \text{ and} \\ F_Y(y) &= P(Y \leq y) \end{aligned}$$

we define their joint distribution as

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

We define the two marginal distributions as $\lim_{y \to \infty} F_{XY}(x, y)$ and $\lim_{x \to \infty} F_{XY}(x, y)$. These are the individual cdfs. Thus

$$\begin{aligned} F_X(x) &= \lim_{y \to \infty} F_{XY}(x, y) \\ F_Y(y) &= \lim_{x \to \infty} F_{XY}(x, y) \end{aligned}$$

The corresponding joint PDF, $f_{XY}(x, y)$ is defined as

$$f_{XY} = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) \tag{2.10.1}$$

We then have

$$\begin{aligned} F_{XY}(x, y) &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x, y) dx dy && (2.10.2) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y) dy dx && (2.10.3) \end{aligned}$$

Let $a < b$ and $c < d$. For any rectangle $I \times J$ where I is an interval with left and right end points as a and b respectively, and J is an interval with left and right end points as c and d respectively, we have,

$$\begin{aligned} \mathcal{P}\left(\{\omega : X(\omega) \in I \text{ and } Y(\omega) \in J\}\right) &= \int_{I \times J} f_{XY}(x, y) dx dy \\ &= \int_a^b \left(\int_c^d f_{XY}(x, y) dy \right) dx \\ &= \int_c^d \left(\int_a^b f_{XY}(x, y) dx \right) dy \end{aligned}$$

The marginal pdfs are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \tag{2.10.4}$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx \tag{2.10.5}$$

2.11 Expectation and Variance

We next study two important parameters associated with a random variable. These measure the mean value of the random variable and its deviation from the mean value.
Consider a discrete random variable X on a Probability Space $(\Omega, \mathcal{B}, P)$, with

$$\mathcal{R}_X = \{x_1, x_2, \cdots, x_N\}$$

and with pmf p_X given by $p_X(x_i) = p_i$ for $1 \leq i \leq n$. The "**Expectation of *X***" is denoted by $E(X)$ (or μ_X) and is defined as the weighted average,

$$\begin{aligned} E(X) &= \sum_{i=1}^{N} x_i P(X = x_i) \\ &= \sum_{i=1}^{N} x_i p_i \end{aligned}$$

(The Expectation of a random variable X is also referred to as the "**mean**" or "**Expected value of *X***").
Analogously for a continuous random variable X the Expectation is defined through the pdf $f_X(x)$ as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

Let $\mathcal{R}_X$ be the Range of X, that is, $\mathcal{R}_X$ is the set of values taken by X. If $g : \mathcal{R}_X \longrightarrow \mathbb{R}$ is a "reasonably smooth" real valued function defined on the Range $\mathcal{R}_X$ of X and $Y = g(X)$ then we define

$g : \mathbb{R}_x \to \mathbb{R}$
$y = g(X)$

$$E(Y) = E(g(X)) = \sum_{j=1}^{N} g(x_j)p_j \tag{2.11.1}$$

$E(y) = E(g(X)) = \sum_{i=1}^{N} g(x_i)p_i$

in the case of a discrete random variable X (where $\mathcal{R}_X = \{x_1, x_2, \cdots, x_N\}$), and

$$E(Y) = E(g(X)) \quad = \quad \int_{-\infty}^{\infty} g(x) f_X(x) dx \qquad (2.11.2)$$

in the case of a continuous random variable X with pdf $f_X(x)$.
It is to easy that the Expectation satisfies the following properties:

1. $E(\alpha X) = \alpha E(X)$ for any $\alpha \in \mathbb{R}$
2. If X is a constant random variable $X = C$ then $E(X) = E(C) = C$. In particular, $E(1) = 1$
3. If X and Y are any two random variables then
$$E(X + Y) = E(X) + E(Y)$$
4. Combining the first and third properties above we get that E is a linear function on the collection of all random variables on $(\Omega, \mathcal{B}, P)$, that is, if $X_1, X_2, \cdots, X_n$ are a finite number of real valued random variables on $(\Omega, \mathcal{B}, P)$ and $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{R}$ then
$$E(\alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n) = \alpha_1 E(X_1) + \alpha_2 E(X_2) + \cdots + \alpha_n E(X_n)$$

Example 2.11.1 Consider the sample space $\Omega = \{H, T\}$ of the experiment of tossing a fair coin. Let X and Y be the random variables defined as follows:

ω	H	T
Random Variable X	1	−1
Random Variable Y	−1	1

Then we have

$$p_X(1) = 0.5 = p_Y(1) \text{ and } p_X(-1) = 0.5 = p_Y(-1)$$

Hence we have

$$E(X) = (1)(0.5) + (-1)(0.5) = 0$$
$$E(Y) = (1)(0.5) + (-1)(0.5) = 0$$

We also have $X+Y$ is the random variable which takes the value 0 for both outcomes and hence $X+Y$ is the Zero random variable and we have $E(0)=0$. Thus we have

$$E(X+Y) \quad = \quad E(X)+E(Y)$$

Consider the random variable $Z = XY$. Then $Z(\omega) = -1$ for both $\omega = H$ and $\omega = T$. Hence Z is the constant random variable $Z = -1$ and we have $E(Z) = -1$. Note that in this case $E(XY) \neq E(X)E(Y)$

Example 2.11.2 Consider the random variables X and Y of Example 2.10.2. We had the joint pmf given by

$Y \rightarrow$ $X \downarrow$	$y_1 = -2$	$y_2 = 2$	*Row Sum*
$x_1 = -1$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
$x_2 = 1$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
Column Sum	$\frac{1}{3}$	$\frac{2}{3}$	1

We have

$$\begin{aligned} E(X) &= (-1) \times (\frac{1}{2}) + 1 \times (\frac{1}{2}) \\ &= 0 \\ E(Y) &= (-2) \times (\frac{1}{3}) + 2 \times (\frac{2}{3}) \\ &= \frac{2}{3} \end{aligned}$$

Now consider the random variable $U = X+Y$. We have $\mathcal{R}_U = \{-3, -1, 1, 3\}$ and the pmf of U given as

U	-3	-1	1	3
p_U	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$

Hence we get

$$\begin{aligned} E(X+Y) = E(U) &= (-3) \times (\frac{1}{6}) + (-1) \times (\frac{1}{6}) + 1 \times (\frac{1}{3}) + 3 \times (\frac{1}{3}) \\ &= \frac{2}{3} \\ &= E(X) + E(Y) \end{aligned}$$

Next consider the random variable $Z = XY$. We have

$$\mathcal{R}_Z = \{-2, 2\}$$

The pmf of Z is given by

Z	-2	2
p_Z	$\frac{1}{2}$	$\frac{1}{2}$

Hence we get

$$\begin{aligned} E(XY) = E(Z) &= (-2) \times (\frac{1}{2}) + 2 \times (\frac{1}{2}) \\ &= 0 \end{aligned}$$

We see that in this case we get $E(XY) = E(X)E(Y)$ whereas in the above Example 2.11.1 we had $E(XY) \neq E(X)E(Y)$

Example 2.11.3 Consider the random variables X, Y of Example 2.10.3.

We have the joint pmf

$Y \rightarrow$ $X \downarrow$	$y_1 = -1$	$y_2 = 0$	$y_3 = 1$	*Row Sums*
$x_1 = -1$	0	$\frac{1}{4}$	0	$\frac{1}{4}$
$x_2 = 0$	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
$x_3 = 1$	0	$\frac{1}{4}$	0	$\frac{1}{4}$
Column Sums	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Hence we get

$$\begin{aligned} E(X) = E(Y) &= \frac{1}{4} \times (-1) + \frac{1}{2} \times 0 + \frac{1}{4} \times 1 \\ &= 0 \end{aligned}$$

Now consider the random variable $U = X + Y$. The possible values of Z is given by

$$\mathcal{R}_U = \{-2, -1, 0, 1, 2\}$$

We have the pmf of U given as below:

U	-2	-1	0	1	2
p_U	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0

We have

$$\begin{aligned} E(X + Y) = E(U) &= (-2) \times 0 + (-1) \times (\frac{1}{2}) + 0 \times 0 + 1 \times \frac{1}{2}) + 2 \times 0 \\ &= 0 \\ &= E(X) + E(Y) \end{aligned}$$

Now consider the random variable $Z = XY$. We have

$$\mathcal{R}_Z = \{-1, 0, 1\}$$

The pmf of Z is given by

Z	-1	0	-1
p_Z	0	1	0

Hence we get

$$\begin{aligned} E(Z) &= (-1) \times 0 + o \times 1 + 1 \times 0 \\ &= 0 \\ &= E(X)E(Y) \end{aligned}$$

In this example we have

$$E(XY) = E(X)E(Y)$$

We can also compute the Expectation $E(X+Y)$ and $E(XY)$ using the joint pmf as follows:
If X and Y are discrete random variables with $\mathcal{R}_X = \{x_1, x_2, \cdots, x_m\}$ and $\mathcal{R}_Y = \{y_1, y_2, \cdots, y_n\}$ with joint pmf p_{XY} as $p_{XY}(x_i, y_j) = p_{ij}$ then

$$\begin{aligned} E(X+Y) &= \sum_{i=m}^{N} \sum_{j=1}^{n} (x_i + y_j) p_{XY}(x_i, y_j) \\ &= \sum_{i=m}^{N} \sum_{j=1}^{n} (x_i + y_j) p_{ij} \end{aligned}$$

$$\begin{aligned} E(XY) &= \sum_{i=m}^{N} \sum_{j=1}^{n} (x_i y_j) p_{XY}(x_i, y_j) \\ &= \sum_{i=m}^{N} \sum_{j=1}^{n} (x_i y_j) p_{ij} \end{aligned}$$

Analogously, if X and Y are continuous random variables with joint pdf $f_{XY}(x,y)$ then

Limits $-\infty$ to ∞ ??

$$
\begin{aligned}
E(X+Y) &= \int_0^\infty \int_0^\infty (x+y) f_{XY}(x,y)dxdy \\
E(XY) &= \int_0^\infty \int_0^\infty (xy) f_{XY}(x,y)dxdy
\end{aligned}
$$

Using the linearity property of the Expectation we see that for any random variable

$$
\begin{aligned}
E(X-\mu_X) &= E(X)-\mu_X \\
&= \mu_X-\mu_X \\
&= 0
\end{aligned}
$$

Thus the random variable Y has $E(Y)=0$. Hence give any random variable we can always standardize it to have mean zero by introducing the random variable $Y = X - \mu_X$.
We next introduce the notion of Variance of a random variable. The variance measures the average of the square of the deviation of X from its mean. We have

Definition 2.11.1 Let X be a random variable with expectation $E(X)=\mu_X$. We then define the variation of X as the expectation of the random variable $(X-\mu_X)^2$

$$
Var(X) = E((X-\mu_X)^2) \tag{2.11.3}
$$

We have

$$
\begin{aligned}
Var(X) &= E((X-\mu_X)^2) \\
&= E(X^2 - 2\mu X + \mu_X^2) \\
&= E(X^2) - 2\mu E(X) + \mu_X^2 \\
&\quad \text{(using the properties of expectation)} \\
&= E(X^2) - \mu_X^2 \\
&= E(X^2) - (E(X))^2
\end{aligned}
$$

Thus we have

$$
Var(X) = E(X^2) - (E(X)^2) \tag{2.11.4}
$$

We define "**standard deviation**" (denoted by σ) of the random variable as

$$\sigma_X = \sqrt{Var(X)} \tag{2.11.5}$$

Example 2.11.4 For the random variable X and Y of Example 2.11.1 we have both X^2 and Y^2 are the constant random variables $X^2 = 1$ and $Y^2 = 1$ and both have expectation 0. Hence we have

$$\begin{aligned} Var(X) &= E(X^2) - E(X) \\ &= 1 - 0 = 1 \\ Var(Y) &= E(Y^2) - E(Y) \\ &= 1 - 0 = 1 \end{aligned}$$

Hence $\sigma_X = 1 = \sigma_Y$

Example 2.11.5 For the uniform random variable $X \sim Uni[a, b]$ we have the pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for} \quad a \leq x \leq b \\ 0 & \text{other} \quad \text{values of } x \end{cases}$$

Hence we get

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_a^b x(\frac{1}{b-a}) dx \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{b+a}{2} \end{aligned}$$

Hence we get

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= \int_a^b x^2 \frac{1}{b-a} dx - \frac{(b+a)^2}{4} \\ &= \frac{b^3 - a^3}{3(b-a)} - \frac{(b+a)^2}{4} \end{aligned}$$

$$\begin{aligned} &= \frac{b^2 + a^2 + 2ab}{12} \\ &= \frac{(b+a)^2}{12} \end{aligned}$$

Hence we get

$$\sigma_X = \frac{b+a}{2\sqrt{3}}$$

We observe the following properties of the variance:

1. We have, for any real number α,

$$\begin{aligned} Var(\alpha x) &= E(\alpha^2 X^2 - (E(\alpha X))^2 \\ &= \alpha^E(X^2) - (\alpha E(X))^2 \\ &= \alpha^2(E(X^2) - (E(X))^2) \\ &= \alpha^2 Var(X) \end{aligned}$$

Thus we have

$$Var(\alpha X) = \alpha^2 Var(X) \text{ for every } \alpha \in \mathbb{R} \tag{2.11.6}$$

2. Let X and Y be two random variables. We have

$$\begin{aligned} Var(X+Y) &= E((X+Y)^2) - (E(X+Y))^2 \\ &= \left\{E(X^2 + Y^2 + 2XY)\right\} \\ &\quad - \{E(X) + E(Y)\}^2 \\ &= \left\{E(X^2) + E(Y^2) + 2E(XY)\right\} \\ &\quad - \left\{(E(X))^2 + (E(Y))^2 + 2E(X)E(Y)\right\} \\ &= \left\{E(X^2) - (E(X))^2\right\} \\ &\quad + \left\{E(Y^2) - (E(Y))^2\right\} \\ &\quad +2\{E(XY) - E(X)E(Y)\} \\ &= Var(X) + Var(Y) + 2Cov(X,Y) \end{aligned}$$

where

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{2.11.7}$$

is called the "**Covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$**. Thus we have

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \tag{2.11.8}$$

Thus we see that in general $Var(X+Y)$ need not be equal to $Var(X)+Var(Y)$. This is because $Cov(X, Y)$ need not be zero in general. However, a "**sufficient condition**" under which this happens is that of independence of X and Y. We have for two discrete random variables X and Y, using joint pmf, $p_{XY}(x_i, y_j) = p_{ij}$,

$$\begin{aligned}
E(XY) &= \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j p_{ij} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j p_i p_j \text{ (if } X \text{ and } Y \text{ are independent)} \\
&= \sum_{i=1}^{m} x_i p_i \left\{\sum_{j=1}^{n} p_j\right\} + \sum_{j=1}^{n} y_j p_j \left\{\sum_{i=1}^{m} p_i\right\} \\
&= \sum_{i=1}^{m} x_i p_i + \sum_{j=1}^{n} y_j p_j \\
&\quad \text{(since } \sum_{j=1}^{n} p_j = 1 = \sum_{i=1}^{m} p_i) \\
&= E(X) + E(Y)
\end{aligned}$$

(We can give a similar proof for the continuous case also), Thus we have

Theorem 2.11.1

$$\begin{aligned}
&X, Y \text{ are independent random variables} \\
&\Longrightarrow \\
&E(X+Y) = E(X) + E(Y)
\end{aligned}$$

Remark 2.11.1 In Example 2.11.2 we obtained $E(X + Y) = E(X) + E(Y)$ as the two random variables are independent and hence $cov(X, Y) = E(XY) - E(X)E(Y) = 0$.

Remark 2.11.2 It must be stressed that the requirement of independence is only a sufficient condition for $cov(X, Y)$ to be zero. We may get

$$E(XY) - E(X)E(Y) = 0$$

even without the two random variables being independent. For instance in Example 2.11.3 we had $E(XY) = E(X)E(Y)$ but the two random variables are not independent.

2.12 Tail Distribution

Let X be a nonnegative random variable on a Probability Space $(\Omega, \mathcal{B}, P)$. The probability that X takes values beyond a certain threshold value is what is known as the Tail Distribution. More precisely we define $F_X^T(x)$, the tail distribution of X as

$$F_X^T(x) \quad = \quad P(X \geq x) \tag{2.12.1}$$

If X is real valued random variable with expectation μ_X, then we are interested in the probability that the X does not deviate from its mean beyond a certain threshold value and hence we are interesting in the tail distribution of $|X-\mu|$, that is, we are interested in $P(|X-\mu| \geq x)$. We now look at some inequalities that give certain estimates for this tail distribution.

I Markov's Inequality

Suppose we have a class of 90 students whose average score in a test is 20. Let us say we are interested in the probability that a randomly chosen student"s score is 60 or more. We have,

$$\begin{aligned}
\text{the number of students who score 60 or more} &= k \\
\Longrightarrow & \\
\text{The total score of these } k \text{ students} &\geq 60k \\
\Longrightarrow & \\
\text{The total score of the class} &\geq 60k \\
\Longrightarrow & \\
\text{The average score must be} &\geq \frac{60k}{90} \\
\Longrightarrow & \\
20 &\geq \frac{60k}{90} \\
\Longrightarrow & \\
\frac{k}{90} &\leq \frac{20}{60} \\
P(Score \ \geq 60) &\leq \frac{20}{60}
\end{aligned}$$

(since $\frac{k}{90}$ is the proportion of students getting a score of at least 60)

If we now replace the scores by a general nonnegative random variable X, the average by its expectation $E(X)$, and the threshold score 60 by a general positive real number k, we should get

$$P(X \geq k) \quad \leq \quad \frac{E(X)}{k}$$

That this is true is what Markov's inequality establishes. We have

Theorem 2.12.1 **Markov's Inequality**

If $\boldsymbol{X}$ is any nonnegative random variable, on a Probability Space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}, \boldsymbol{P})$, with expectation $\boldsymbol{E(X)}$ then

$$\boldsymbol{P(X \geq k) \quad \leq \quad \frac{E(X)}{k}} \textbf{ for any } \boldsymbol{k > 0} \tag{2.12.2}$$

Proof:
Case 1: Discrete random variable
Let X be a nonnegative random variable taking values $x_1, x_2, \cdots, x_n$ with respective probabilities $p_1, p_2, \cdots, p_n$. Then we have for any $k > 0$,

$$\begin{aligned} E(X) &= \sum_{j=1}^{n} p_j x_j \\ &= \sum_{\{j: x_j < k\}} p_j x_j + \sum_{\{j: x_j \geq k\}} p_j x_j \\ &\geq \sum_{\{j: x_j \geq k\}} p_j x_j \\ &\geq k \sum_{\{j: x_j \geq k\}} p_j \\ \Longrightarrow & \\ E(X) &\geq kP(X \geq k) \\ \Longrightarrow & \\ P(X \geq k) &\leq \frac{E(X)}{k} \end{aligned}$$

thus proving the inequality. Analogously we prove in the case of continuous random variables as follows:
Case 2: Continuous random variable

Let $f(x)$ be the pdf of the random variable X. Then we have

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x)dx \\
&= \int_{0}^{\infty} x f_X(x)dx \text{ (since } X \text{ is nonnegative)} \\
&= \int_{0}^{k} x f_X(x)dx + \int_{k}^{\infty} x f_X(x)dx \\
&\geq \int_{k}^{\infty} x f_X(x)dx \\
&\geq k\int_{k}^{\infty} f_X(x)dx \text{ (since } x \geq k \text{ in the domain of integration)} \\
&\geq kP(X \geq k) \\
\Longrightarrow & \\
P(X \geq k) &\leq \frac{E(X)}{k}
\end{aligned}
$$

Remark 2.12.1 The inequality will not be of any use at all if $E(X) = \infty$. Hence without loss of generality we can assume in the above statement that $E(X) < \infty$

Remark 2.12.2 The inequality may give some times some bizarre results even if $E(X) < \infty$, as shown in the following example.

Example 2.12.1 Consider the random experiment of throwing a fair die. We have, in this case,

$$\Omega = \{1, 2, 3, 4, 5, 6\} \text{ and } p(j) = \frac{1}{6} \text{ for } 1 \leq j \leq 6$$

Let X be the random variable defined as $X(j) = j$.
We have

$$
\begin{aligned}
E(X) &= \frac{1+2+3+4+5+6}{6} \\
&= 3.5
\end{aligned}
$$

Hence by Markov's inequality we get

$$P(X \geq 3) \leq \frac{E(X)}{3} = \frac{3.5}{3}$$

Since the rhs above is > 1, in this case, we get no information from Markov's inequality since we already know that the probability is anyway ≤ 1.
We also get from Markov's inequality

$$P(X \geq 5) \quad \leq \quad \frac{E(X)}{5} = \frac{3.5}{5} = 0.7$$

The actual value is

$$P(X \geq 5) \quad = \quad p(5) + p(6) = \frac{2}{6} = \frac{1}{3} = 0.33$$

Thus we see that the estimate we get from Markov's inequality is a highly exaggerated overestimate. Thus we see from this example that the Markov's inequality may give sometimes highly exaggerated overestimates.

Remark 2.12.3 Despite the above Remark, it must be noted that the inequality is of a very general nature in the sense that it holds for all nonnegative random variables and for all positive k. If we keep this in mind, this inequality is "tight", that is there will be at least one nonnegative random variable X and one positive real number k for which the inequality becomes an equality as shown in the following example. Hence we cannot make the lhs any smaller if the inequality has to hold fro all nonnegative random variables and all $k > 0$

Example 2.12.2 Let $a > 1$ be any positive real number. Consider a random variable X which takes only two values a and 0 with respective probabilities $\frac{1}{a}$ and $\frac{a-1}{a}$. For this nonnegative random variable we get

$$E(x) \quad = \quad a \times \frac{1}{a} = 1$$

$$P(X \geq a) \quad = \quad P(X = a) = \frac{1}{a}$$

On the other hand we get from Markov's inequality,

$$P(X \geq a) \quad \leq \quad \frac{E(X)}{a} = \frac{1}{a}$$

Comparing with the exact value of this probability found above we see that equality holds in the Markov's inequality for this X and for this a.

II Chebychev's Inequality:
Let X be a real valued random variable with $E(X) = \mu < \infty$. Then $Y = |X - \mu|$ is a nonnegative random variable and we have for any positive real number k and any positive increasing function $f(t) : [0, \infty) \longrightarrow \mathbb{R}$,

$$\begin{array}{lcl} Y(\omega) \geq k & \Longleftrightarrow & f(Y(\omega)) \geq f(k) \\ \quad\quad \Longrightarrow & & \\ P(Y \geq k) & = & P(f(Y) \geq f(k)) \\ & \leq & \dfrac{E(f(Y))}{f(k)} \text{ by Markov's inequality 2.12.2} \\ & & \text{ applied to the random variable } f(X) \text{ and with } k \text{ as } f(k)) \end{array}$$

As a special case let us take $f(t) = t^2$. Then we have from above,

$$P(Y \geq k) \leq \frac{E(Y^2)}{k^2}$$

Recalling that $Y = |X - \mu|$ we get

$$P(|X - \mu| \geq k) \leq \frac{E(|X - \mu|^2)}{k^2}$$

But

$$E(|X - \mu|^2) = E((X - \mu)^2) = Var(X) \text{ (since } X \text{ and } \mu \text{ are real)}$$

Substituting above we get

$$P(|X - \mu|) \leq \frac{Var(X)}{k^2}$$

Thus we have

Theorem 2.12.2 **Chebychev's Inequality**

If X is any real valued random variable with expectation $E(X) = \mu < \infty$ and finite Variance $Var(X)$, then

$$\boldsymbol{P(|X - \mu|) \leq \frac{Var(X)}{k^2}} \textbf{ for any real number } \boldsymbol{k > 0} \quad (2.12.3)$$

III Chernoff Bound

Let X be any real valued random variable. Now consider the function

$$f : \mathbb{R} \longrightarrow \mathbb{R}$$

defined as $f(t) = e^{\alpha t}$, where α is a positive constant. Then $f(t)$ is a nonnegative increasing function. The random variable defined as $Y = f(X) = e^{\alpha X}$ is a nonnegative random variable. Further since $f(t)$ is increasing we see that for any real number k we have

$$\begin{aligned} X \geq k &\iff f(X) \geq f(k) \\ &\iff e^{\alpha X} \geq e^{\alpha k} \\ &\iff Y \geq f(k) \end{aligned}$$

Hence we get

$$P(X \geq a) \;=\; P(Y \geq e^{\alpha k}) \tag{2.12.4}$$

Since Y is a nonnegative random variable we can apply Markov's inequality to Y to get

$$P(Y \geq e^{\alpha k}) \;\leq\; \frac{E(Y)}{e^{\alpha k}} \tag{2.12.5}$$

From the above two equations we get

$$P(X \geq k) \;\leq\; \frac{E(Y)}{e^{\alpha k}} = e^{-\alpha k}E(e^{\alpha X}) \tag{2.12.6}$$

Similarly consider the decreasing function $g(t) = e^{-\alpha t}$ and let Y be the nonnegative random variable defined as $Y = g(X) = e^{-\alpha X}$. Using the fact that $X \leq k$ if and only if $g(X) \geq g(k)$, we get

$$\begin{aligned} P(X \leq k) &= P(g(X) \geq g(k)) \\ &= P(Y \geq g(k)) \\ &\leq \frac{E(Y)}{g(k)} \end{aligned}$$

Substituting $Y = g(X) = e^{-\alpha X}$ and $g(k) = e^{-\alpha k}$ we get

$$P(X \leq k) \;\leq\; \frac{E(e^{-\alpha X})}{e^{-\alpha k}} = e^{\alpha k}E(e^{-\alpha X}) \tag{2.12.7}$$

Since 2.12.6 and 2.12.7 hold whatever $\alpha > 0$ we choose we get

$$P(X \geq k) \leq \min_{\alpha > 0} \left\{ e^{-\alpha k} E(e^{\alpha X}) \right\} \tag{2.12.8}$$

$$P(X \leq k) \leq \min_{\alpha > 0} \left\{ e^{\alpha k} E(e^{-\alpha X}) \right\} \tag{2.12.9}$$

Equations 2.12.8 and 2.12.9 are known as Chernoff bounds.