STOCHASTIC PROCESSES IN ENGINEERING SCIENCE[1]

Ravi R. Mazumdar
Department of Electrical and Computer Engineering
University of Waterloo
Canada

2010

Chapter 1

Introduction and Overview of Probability

The basis for understanding and analyzing stochastic processes is probability theory and real analysis.This chapter presents an overview of probability and the associated concepts in order to develop the basis for the theory of stochastic or random processes in subsequent chapters. The results given in this chapter are the key ideas and results with direct bearing on the tools necessary to develop the results in the sequel. Although it is not necessary to have had a prior course in probability, graduate students in engineering usually have had an undergraduate course in probability and statistics. Prior background will enable readers to go through the early parts faster however the emphasis in this book is understanding the basic concepts and their use and this is different from an undergraduate course in which the emphasis is on computation and calculation.

1.1 Definition of a probability space

Let us begin at the beginning. In order to introduce the notion of probabilities and operations on them we first need to set up the mathematical basis or hypotheses on which we can construct our edifice. This mathematical basis is the notion of a *probability space*. The first object we need to define is the space Ω which is known as the space of all possible outcomes (or observations). This is the mathematical artifact which is the setting for studying the likelihood of *occurrence* or outcome of an *experiment* based on some assumptions on how we expect the quantities of interest to behave. By experiment it is meant the setting in which quantities are observed or measured. For example the experiment may be the measurement of rainfall in a particular area and the method of determining the amount. The outcome i.e. $\omega \in \Omega$ is the actual value measured (or observed) at a given time. Another, classical example is the roll of a die. The space Ω is just the possible set of values which can appear i.e. $\{1, 2, 3, 4, 5, 6\}$ in one roll. In the case of measurements of rainfall it is the numerical value typically any real non-negative number between $[0, \infty)$. Once Ω is specified the next notion we need is the notion of an *event*. Once an experiment is performed the outcome ω is observed and it is possible to tell whether an event of interest has occurred or not. For example, in a roll of the die we can say whether the number is even or odd. In the case of rainfall whether it is greater than 5mm but less than 25 mm or not. The set of all events is usually denoted by $\mathcal{F}$. $\mathcal{F}$ is just the collection of subsets of Ω which satisfy the following axioms:

i) $\Omega \ \in \mathcal{F}$

ii) $A \ \in \mathcal{F}$ then $\bar{A} \ \in \ \mathcal{F}$ where $\bar{A}$ denotes the complement of A.

iii) If the sequence $\{A_n\}$ with $A_n \ \in \ \mathcal{F}$ then $\bigcup_{n=1}^{\infty} A_n \ \in \mathcal{F}$.

The above axioms imply that $\mathcal{F}$ is a σ-algebra or field. Property i) says that the entire space Ω is an event. Property ii) states that if A is an event then *not* A is also an event. Property iii) implies that the union of a countable number of events is an event and by property ii) it also implies that the intersection of a countable set of events is also an event.

To complete the mathematical basis we need the notion of a *probability* measure which is denoted by $\mathbb{P}$. This is a function which attaches a numerical value between [0,1] to events $A \ \in \mathcal{F}$ with the following properties:

i) $\mathbb{P} : \mathcal{F} \rightarrow \ [0,1]$ i.e. for any $A \ \in \mathcal{F}$ $\mathbb{P}(A) \in \ [0,1]$.

ii) $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(\phi) = 0$

iii) If $\{A_n\}_{n=1}^{\infty}$ is a countable collection of disjoint sets with $A_n \in \ \mathcal{F}$ then $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

The second property states that the probability of the entire space is a certain event. Property iii) is called the property of σ-additivity (or countably additive) and plays a crucial role in the definition of probability measures and the notion of random variables (which we will see soon) but it is beyond the scope of this text. Note σ-additivity does not follow if we define (iii) in terms of a finite collection of disjoint sets and state that it holds for every n.

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the probability space. In the sequel it must be remembered that except when specifically given, all the quantities are defined with respect to a given probability space i.e. the 'triple' always is lurking in the background and all claims are made w.r.t. to that space even though it is not specified every time—this will be commented upon later.

Definition 1.1.1 *The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be a complete probability space if $\mathcal{F}$ is completed with all $P-$null sets.*

1.1.1 Some properties of probability measures

¿From the definition of the probability measure as a mapping from $\mathcal{F} \rightarrow \ [0,1]$ and axioms ii) and iii) above we have :

$$\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$$

and

$$A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$$

or the probability measure is monotone non-decreasing. We append a proof of the second property below:

Indeed, $B = (B - A) + A$ where $B - A$ denotes the events in B which are not in A and therefore $(B - A)\bigcup A = B$ and we use the notation of $+$ for the union of disjoint sets. Now, using the countably additive axiom we have

$$\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$$

and by definition $\mathbb{P}(B - A) \geq 0$.

Hence for any collection $\{A_n\}$ of sets in $\mathcal{F}$ it follows that:

$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

In general $\mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \bigcap B)$.

An immediate consequence of the countable additivity of the probability measure is the notion of *sequential continuity* of the measure by which it is meant that the limit of the probabilities of events can be replaced by the probability of the limit of the events. This is clarified below.

Proposition 1.1.1 *a) Let $\{A_n\}$ be an increasing sequence of events i.e. $A_n \subset A_{n+1}$ then*

$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \mathbb{P}(A_n)$$

b) Let $\{B_n\}$ be a decreasing sequence of events i.e. $B_{n+1} \subset B_n$ then

$$\mathbb{P}(\bigcap_{n=1}^{\infty} B_n) = \lim_{n \to \infty} \mathbb{P}(B_n)$$

Proof:

First note that for any $n \geq 2$,

$$A_n = A_1 + \sum_{m=1}^{n-1}(A_{m+1} - A_m)$$

and

$$\bigcup_{n=1}^{\infty} A_n = A_1 + \sum_{m=1}^{\infty}(A_{m+1} - A_m)$$

Now using the countable additivity assumption and noting that $\{A_{m+1} - A_m\}$ are disjoint we have

$$\begin{aligned} \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) &= \mathbb{P}(A_1) + \textstyle\sum_{m=1}^{\infty} \mathbb{P}(A_{m+1} - A_m) \\ &= \mathbb{P}(A_1) + \lim_{n \to \infty} \textstyle\sum_{m=1}^{n-1} \mathbb{P}(A_{m+1} - A_m) \\ &= \lim_{n \to \infty} \mathbb{P}(A_n) \end{aligned}$$

The proof of part b) follows by using DeMorgan's law i.e.

$$\overline{\bigcap B_n} = \bigcup \overline{B_n}$$

and the result of part a) noting that $\mathbb{P}(B_n) = 1 - \mathbb{P}(\bar{B_n})$.

Remark: Note that in the case of a) $\bigcup_{n=1}^{\infty} A_n$ is the limit of an increasing sequence of events and thus a) is equivalent to $P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n)$ which is what is meant by the interchange of limits. Part b) corresponds to the limit of a decreasing sequence of events. Note the limits exist since the sequences of probabilities are increasing (or decreasing).

An important notion in probability theory which sets it apart from real analysis is the notion of independence.

Definition 1.1.2 *The events $\{\mathcal{A}_n\}$ are said to be mutually independent if for any subset $\{k_1, k_2, \cdots, k_r\}$ of $\{1, 2, \cdots, n\}$*

$$\mathbb{P}\left(\bigcap_{j=1}^{r} \mathcal{A}_{k_j}\right) = \prod_{k=1}^{r} \mathbb{P}(\mathcal{A}_{k_j}) \tag{1.1. 1}$$

Remarks:

i. Note pairwise independence between events does not imply mutual independence of all the events. Here is a simple example.

   Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ i.e. Ω consists of 4 elements such that: $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2) = \mathbb{P}(\omega_3) = \mathbb{P}(\omega_4) = \frac{1}{4}$. Now define the following events: $A = \{\omega_1, \omega_2\}, B = \{\omega_2, \omega_3\}, C = \{\omega_1, \omega_3\}$. Then it is easy to see that: $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$ and $\mathbb{P}(A \bigcap B) = \mathbb{P}(\omega_2) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B)$ and similarly $\mathbb{P}(B \bigcap C) = \mathbb{P}(B)\mathbb{P}(C)$ and $\mathbb{P}(A \bigcap C) = \mathbb{P}(A)\mathbb{P}(C)$ showing that the events A, B and C are pairwise independent. However $\mathbb{P}(A \bigcap B \bigcap C) = \mathbb{P}\{\phi\} = 0 \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{8}$ showing that A, B and C are not independent.

   Once can also construct an example where A, B, C are independent but not pairwise independent.

ii. The second point which is often a point of confusion is that disjoint events are not independent since if two events A and B are dis joint and independent then since $0 = \mathbb{P}(A \bigcap B) = \mathbb{P}(A)\mathbb{P}(B)$ then it would imply that at least one of the events has 0 probability. In fact disjointness of events is a strong form of dependence since the occurrence of one precludes the occurrence of the other.

To conclude our discussion of probabilitity measures we present an important formula which is very useful to compute probabilities of events with complicated dependence.

Proposition 1.1.2 *(Inclusion-Exclusion Principle or Poincaré's Formula) Let $\{A_i\}_{i=1}^n \in \mathcal{F}$ be any collection of events. Then:*

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{\substack{S \subset \{1,2,..,n\} \\ S \neq \emptyset}} (-1)^{|S|-1} \mathbb{P}(\bigcap_{j \in S} A_j) \tag{1.1. 2}$$

Proof: The result clearly holds for $n = 2$ i.e.:

$$\mathbb{P}(A_1 \bigcup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \bigcap A_2)$$

Now define $B = \bigcup_{i=1}^{n-1} A_i$. Then from above:

$$\begin{aligned} \mathbb{P}(\bigcup_{i=1}^{n} A_i) &= \mathbb{P}(B \bigcup A_n) \\ &= \mathbb{P}(B) + \mathbb{P}(A_n) - \mathbb{P}(B \bigcap A_n) \\ &= \mathbb{P}(\bigcup_{i=1}^{n-1} A_i) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i-1}^{n-1} (A_i \bigcap A_n)\right) \end{aligned}$$

where we used the distributivity property of unions and intersection in the last step.

To prove the result we first note that:

$$\sum_{S \subset \{1,..,n\}} (-1)^{|S|-1} \mathbb{P}(\bigcap_{j \in S} A_j) = \sum_{S \subset \{1,...,n-1\}} (-1)^{|S|-1} \mathbb{P}(\bigcap_{j \in S} A_j) + \mathbb{P}(A_n) - \sum_{S \text{ } containing \text{ } n, |S| \geq 2} (-1)^{|S|-1} \mathbb{P}(\bigcap_{j \in S} A_j)$$

Identifying the terms on the r.h.s. with the events defined upto $(n-1)$ of the inclusion-exclusion identity above the rest of the induction argument can be completed.

A particular form of the above result is sometimes useful in applications.

Corollary 1.1.1 *(Bonferroni inequalities)*
Let $r < n$ be even. Then the following holds:

$$\sum_{S \neq \emptyset, S \in \{1,2,..,n\}, |S| \leq r} (-1)^{|S|-1} \mathbb{P}\left(\bigcap_{j \in S} A_j\right) \leq \mathbb{P}(\bigcup_{j=1}^{n} A_j) \qquad (1.1.\ 3)$$

If $r < n$ is odd then

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) \leq \sum_{S \neq \emptyset, S \in \{1,2,..,n\}, |S| \leq r} (-1)^{|S|-1} \mathbb{P}\left(\bigcap_{j \in S} A_j\right) \qquad (1.1.\ 4)$$

The above corollary states that if we truncate the Poincaré inequality to include terms upto r then, if r is odd the resulting probability is smaller than the required probability while if r is odd then the resulting probability will be larger. A very simple consequence is:

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) \geq \sum \mathbb{P}(A_i) - \sum_{i,j,i \neq j} \mathbb{P}(A_i \bigcap A_j)$$

1.2 Random variables and probability distributions

One of the most important concepts for a useful operational theory is the notion of a *random variable* or r.v. for short. In this course we will usually concern ourselves with so-called real valued random variables or countably valued (or discrete-valued) random variables. In order to define random variables we first need the notion of what are termed *Borel sets.* In the case of real spaces then the Borel sets are just sets formed by intersections and unions of open sets (and their complements) and hence typical Borel sets are any open intervals of the type (a, b) where $a, b \in \Re$ and by property ii) it can also include half-open, closed sets of the form $(a, b]$ or $[a, b]$ and by the intersection property a point can be taken to be a Borel set etc. Borel sets play an important role in the definition of random variables. Note for discrete valued r.v's a single value can play the role of a Borel set. In the sequel we always will restrict ourselves to $\Re$ or $\Re^n$ unless explicitly specified to the contrary.

Definition 1.2.1 *A mapping $X(.) : \Omega \to \Re$ is said to be a random variable if the inverse image of any Borel set C in $\Re$ belongs to $\mathcal{F}$ i.e.*

$$\{\omega : X(\omega) \ \in C\} \in \ \mathcal{F}$$

or what is equivalent

$$\{X^{-1}(C) \ \in \ \mathcal{F}\}$$

Remark: In words what the above definition says that if we consider all the elementary events ω which are mapped by $X(.)$ into C then the collection $\{\omega\}$ will define a valid event so that a probability can be assigned to it. At the level of these notes there will never be any difficulty as to whether the given mappings will define random variables, in fact we treat r.v's as "primitives". This will be clarified a little later on.

From the viewpoint of computations it is most convenient to work with an "induced measure" rather than on the original space. This amounts to defining the probability measure induced by or associated with a r.v. on its range so that rather than treat the points ω and the probability measure $\mathbb{P}$ we can work with a probability distribution on $\Re$ (or X) with $x \ \in \Re$ as the sample values. This is defined below.

Definition 1.2.2 *Let $X(\omega)$ be a real valued r.v. defined on $(\Omega, \mathcal{F}, \mathbb{P})$ then the function*

$$F(x) = \mathbb{P}(\{\omega : X(\omega) \leq x\})$$

is called the (cumulative) probability distribution function of X.

Remark: Note $F(x)$ is a probability distribution defined on $\Re$ i.e. it corresponds to the probability measure corresponding to $\mathbb{P}$ induced by $X(.)$ on $\Re$. Usually the short form $\mathbb{P}(X(\omega) \leq x)$ is used instead of $\mathbb{P}(\{\omega : X(\omega) \leq x\})$. The tail distribution given by $\mathbb{P}\{X(\omega) > x\} = 1 - F(x)$ is called the complementary distribution function.

By definition $F(x)$ is a monotone non-decreasing function since $\{\omega : X(\omega) \ \leq x_1\} \ \subset \{\omega : X(\omega) \leq x_2\}$ whenever $x_1 \ \leq x_2$. Also because of the definition in terms of the "$\leq$" sign rather than

$<$ sign F(x) is right continuous i.e., F(x) = F(x+), the value at x is the one obtained by shrinking to x from the right. This follows from the sequential continuity of the probability measure and we append a proof below. What right continuity means is that the distribution function can have discontinuities and the value at the point of discontinuity is taken to be the value of F(x) to the right of the discontinuity. Some authors define the distribution function in terms of the "$<$" sign in which case the function is left continuous.

Finally from now on, unless necessary for clarity, we will use X for $X(\omega)$. In general the uppercase variables will be used to represent random quantities.

Proof of right continuity: Let ϵ_n be a monotone decreasing sequence going to zero e.g. $\{\frac{1}{n}\}$. Define $B_n = \{X(\omega) \leq x + \epsilon_n\}$. Then $\bigcap_n B_n = \{X \leq x\}$. By sequential continuity

$$F(x) = lim_{n\to\infty} F(x + \epsilon_n) = F(x+)$$

Other properties of the distribution function:

i) $\lim_{x\to-\infty} F(x) = 0$

ii) $\lim_{x\to\infty} F(x) = 1$ since $\lim_{x\to\infty} F(x) = \mathbb{P}(X \leq \infty) = \mathbb{P}(\Omega) = 1$ since $X(.)$ being real valued takes values in $(-\infty, \infty)$.

Given a r.v. and the definition of the distribution function one can consider $X(\omega)$ as a 'coordinate' or sample point on the probability space defined by $\{X, \mathcal{B}, F(dx)\}$ where $\mathcal{B}$ is the Borel σ-field on X. Since typically in applications we have a given r.v. and other r.v's constructed by operations on the r.v. this is a concrete way of constructing a probability space.

If the distribution function $F(x)$ is differentiable with respect to x then :

$$p(x) = \frac{dF(x)}{dx} \tag{1.2. 5}$$

is called the probability density function (abbreviated as p.d.f) of X and

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

1.2.1 Expectations, moments and characteristic functions

Given a r.v. X we define the (mathematical) expectation or mean as

$$\mathbf{E}[X] = \int_\Omega X(\omega)dP(\omega) = \int_\Re xdF(x) \tag{1.2. 6}$$

provided the integrals exist. If a density exists then:

$$\int_\Re xdF(x) = \int_\Re xp(x)dx$$

Similarly, the $kth.$ moment, if it exists, is defined as :

$$\mathbf{E}[X^k(\omega)] = \int_\Omega X^k(\omega)d\mathbb{P}(\omega) = \int_\Re x^k dF(x) \tag{1.2. 7}$$

Of particular interest in applications is the second centralized moment called the variance denoted by $\sigma^2 = Var(X)$ which is given by

$$Var(X) = \mathbf{E}\Big[(X(\omega) - \mathbf{E}[X])^2\Big] \tag{1.2. 8}$$

The quantity $\sigma = \sqrt{Var(X)}$ is referred to as the standard deviation.

¿From the definition of expectations it can be seen that it is a linear operation i.e.

$$\mathbf{E}\Big[\sum_1^n a_i X_i(\omega)\Big] = \sum_1^n a_i \mathbf{E}[X_i(\omega)]$$

for arbitrary scalars $\{a_i\}$ provided the individual expectations exist.

Also the expectation obeys the order preservation property i.e. if $X(\omega) \geq Y(\omega)$ then

$$\mathbf{E}[X(\omega)] \geq \mathbf{E}[Y(\omega)]$$

So far we have assumed that the r.v's are continuous. If on the other hand the r.v's take discrete values say $\{a_n\}$ then the distribution of the discrete r.v. is specified by the probabilities :

$$p_n = \mathbb{P}(X(\omega) = a_n); \quad \sum_n p_n = 1 \tag{1.2. 9}$$

and

$$F(x) = \sum_{n:a_n \leq x} p_n$$

and by definition $F(x)$ is right continuous. Indeed in the discrete context the σ-field $\mathcal{F}$ is be generated by disjoint events $\{a_n\}$.

In light of the definition of the distribution function it can be readily seen that for non-negative r.v's the following holds:

$$\mathbf{E}[X(\omega)] = \int_0^\infty (1 - F(x))dx \tag{1.2. 10}$$

Finally, a third quantity of interest associated with a r.v. is the *characteristic function* defined by

$$C(h) = \mathbf{E}[e^{\imath hX(\omega)}] = \int_{\Re} e^{\imath hx} dF(x) \tag{1.2. 11}$$

where $\imath = \sqrt{-1}$. In engineering literature rather than working with the Fourier transform of the p.d.f. (which is what the characteristic function represents when a density exists) we often work with the Laplace transform which is called the *moment generating function* but it need not be always defined.

$$M(h) = \mathbf{E}[e^{hX(\omega)}] = \int_{\Re} e^{hx} dF(x) \tag{1.2. 12}$$

The term moment generating function is used since knowledge of $M(h)$ or for that matter $C(h)$ completely specifies the moments since:

$$\mathbf{E}[X(\omega)^k] = \frac{d^k}{dh^k} M(h)|_{h=0}$$

i.e. the kth derivative w.r.t. h evaluated at $h = 0$. In the case of the characteristic function we must add a multiplication factor of $\imath^k$.

In the case of discrete r.v's the above definition can be replaced by taking $z-transforms$. In the following we usually will not differentiate (unless explicitly stated) between discrete or continuous r.v's and we will use the generic integral $\int g(x)dF(x)$ to denote expectations with respect to the probability measure $F(x)$. In particular such an integral is just the so-called Stieltjes integral and in the case that X is discrete

$$\int g(x)dF(x) = \sum_n g(a_n)p_n$$

As moment generating functions are not always defined, we will work with characteristic functions since $|e^{ihx}| = 1$ and therefore $|C(h)| \leq \int dF(x) = 1$ showing that they are always defined. Below we list some important properties associated with characteristic functions:

i) C(0) = 1

ii) C(-h)= $C^*(h)$ where C^* denotes the complex conjugate of C.

iii) The characteristic function of a probability distribution is a non-negative definite function (of h).

By non-negative definite function it is meant that given any complex numbers $\{\lambda_k\}_{k=1}^N$ then denoting the complex conjugate of a complex number λ by λ^*,

$$\sum_{k=1}^N \sum_{l=1}^N \lambda_k \lambda_l^* C(h_k - h_l) \geq 0$$

for all N. Let us prove this result. Define $Y = \sum_{k=1}^N \lambda_k e^{ih_k X}$. Then

$$|Y|^2 = YY^* = \sum_{k=1}^N \sum_{l=1}^N \lambda_k \lambda_l^* e^{\imath(h_k - h_l)X}$$

Hence taking expectations and noting the the expectation of a non-negative r.v. is non-negative we obtain:

$$\mathbf{E}[|Y|^2] = \sum_{k=1}^N \sum_{l=1}^N \lambda_k \lambda_l^* C(h_k - h_l) \geq 0$$

It is an important fact (which is beyond the scope of the course) that characteristic functions completely determine the underlying probability distribution. What this means is that if two r.v's have the same characteristic function then they have the same distribution or are probabilistically equivalent. Note this does not imply that they are identical. Also an important converse to the result holds which is due to Paul Lévy, any function which satisfies the above properties corresponds to the characteristic function of a probability distribution.

1.2.2 Functions of random variables

It is evident that given a r.v., any (measurable) mapping of the r.v. will define a random variable. By measurability it is meant that the inverse images of Borel sets in the range of the function belong to the σ-field of X; more precisely, given any mapping $f() : X \to Y$, then $f(.)$ is said to be measurable if given any Borel set $C \in \mathcal{B}_Y$ the Borel σ-field in Y then

$$f^{-1}(C) \in \mathcal{B}_X$$

where $\mathcal{B}_X$ denotes the Borel σ-field in X. This property assures us that the mapping $f(X(\omega))$ will define a r.v. since we can associate probabilities associated with events.

Hence, the expectation can be defined (if it exists) by the following:

$$\mathbf{E}[f(X(\omega))] = \int_\Omega f(X(\omega))d\mathbb{P}(\omega) = \int_X f(x)dF_X(x) = \int_Y ydF_Y(y)$$

where the subscripts denote the induced distributions on X and Y respectively.

It is of interest when we can compute the distribution induced by $f(.)$ on Y in terms of the distribution on X. This is not obvious as will be illustrated by the following example.

Example : Let $X(\omega)$ be defined on $\Re$ with distribution $F_X(x)$. Let $Y(\omega) = X^2(\omega)$. Then $Y(\omega)$ takes values in $[0, \infty)$. Hence the distribution of Y denoted by $F_Y(.)$ has the following property : $F_Y(x) = 0$ for $x < 0$. Let us now obtain $F_Y(x)$ for non-negative values of x. By definition:

$$F_Y(x) = \mathbb{P}(Y(\omega) \leq x) = \mathbb{P}(X^2(\omega) \leq x)$$

This implies that the event $\{Y(\omega) \leq x) = \{X(\omega) \leq \sqrt{x}\} \bigcap \{X(\omega) \geq -\sqrt{x}\}$ where $\sqrt{x}$ is the positive square root. Now by definition of $F_X(x)$ and noting that it is right-continuous we obtain

$$F_Y(x) = F_X(\sqrt{x}) - F_X(-\sqrt{x}) + P(X = -\sqrt{x})$$

and if the distribution is continuous then the third term is 0.

The above example shows the difficulty in general of obtaining the distribution of a r.v. corresponding to a transformation of another r.v. in terms of the distribution of the transformed r.v. This is related to the fact that inverses of the mappings need not exist. However, when the mappings are monotone (increasing or decreasing) then there is a very simple relation between the distributions.

First suppose that $f(.)$ is monotone increasing then its derivative is positive. Then the event:

$$\{Y(\omega) \leq y\} \equiv \{X(\omega) \leq f^{-1}(y)\}$$

and hence

$$F_Y(y) = F_X(f^{-1}(y))$$

since the inverse exists (since monotonicity implies that the function is 1:1).

Similarly if $f(.)$ is monotone decreasing then its derivative is negative and the event:

$$\{Y(\omega) \leq y\} \equiv \{X(\omega) \geq f^{-1}(y)\}$$

Hence, assuming that the distribution is continuous then:

$$F_Y(y) = 1 - F_X(f^{-1}(y))$$

Hence, in both cases if a $F_X(.)$ possesses a density then:

$$p_Y(y) = \frac{p_X(f^{-1}(y))}{|f'(f^{-1}(y))|} \tag{1.2. 13}$$

by the usual chain rule for differentiation where $f'(.)$ denotes the derivative of f.

When the mapping is non-monotone there is no general expression for obtaining $F_Y(.)$. But we can still obtain the distribution by the following way. Let $\phi(y)$ denote the set of all points x such that $f(x) \leq y$ i.e.

$$\phi(y) = \{x : f(x) \leq y\}$$

Then:

$$F_Y(y) = \int_{\phi(y)} dF_X(x)$$

In the next section we will obtain a generalization of these results for the case of vector valued r.v's.. We will then present several examples.

In the sequel we will generally omit the argument ω for r.v's and capital letters will usually denote random variables while lowercase letters will denote values.

1.3 Joint distributions and Conditional Probabilities

So far we have worked with one r.v. on a probability space. In order to develop a useful theory we must develop tools to analyze collection of r.v's and in particular interactions between them. Of course they must be defined on a common probability space. Let $\{X_i(\omega)\}_{i=1}^n$ be a collection of r.v's defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Specifying their individual distributions $F_i(x)$ does not account for their possible interactions. In fact, to say that they are defined on a common probability space is equivalent to specifying a probability measure $\mathbb{P}$ which can assign probabilities to events such as $\{\omega : X_1(\omega) \in \mathcal{A}_1, X_2(\omega) \in \mathcal{A}_2, ..., X_n(\omega) \in \mathcal{A}_n\}$. In the particular case where the r.v's are real valued this amounts to specifying a joint probability distribution function defined as follows:

$$F(x_1, x_2, .., x_n) = \mathbb{P}\{X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, ..., X_n(\omega) \leq x_n\} \tag{1.3. 14}$$

for arbitrary real numbers $\{x_i\}$. Since n is finite the collection of r.v's can be seen as a vector valued r.v. taking values in $\Re^n$.

The joint distribution being a probability must sum to 1 i.e.

$$\int_{\Re}\int_{\Re} ... \int_{\Re} dF(x_1, x_2, ..., x_n) = 1$$

and in addition if we 'integrate out' any m out of the n variables over $\Re^m$ then what remains is a joint distribution function of the remaining (n-m) r.v's i.e.

$$F(x_1, x_2, ..., x_{n-m}) = F(x_1, x_2, ..., x_{n-m}, \infty, \infty, .., \infty)$$

Another way of stating this is that if the joint density exists i.e. $\exists$ $p(x_1, , x_n)$ such that

$$F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1} .. \int_{-\infty}^{x_n} p(y_1, y_2, .., y_n) dy_1 dy_2 ... dy_n$$

Then,

$$F(x_1, x_2, , x_{n-m}) = \int_{-\infty}^{x_1} .. \int_{-\infty}^{x_{n-m}} \int_{\Re} .... \int_{\Re} p(y_1, y_2, .., y_{n-m}.y_{n-m+1}, .., y_n) dy_1 dy_2 ... dy_n$$

The above property is known as the *consistency* of the distribution. In particular if we integrate out any (n-1) r.v's then what remains is the probability distribution of the remaining r.v. which has not been integrated out.

We now come to an important notion of *independence* of r.v's in light of the definition of independent events.

Let $\{X_i\}_{i=1}^n$ be a finite collection of r.v's. This amounts to also considering the vector valued r.v. whose components are X_i. Then we have the following definition:

Definition 1.3.1 *The r.v's $\{X_i\}_{i=1}^n$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be mutually independent if for any Borel sets $\mathcal{A}_i$ in the range of X the following holds:*

$$\mathbb{P}\Big(X_{i_1}(\omega) \in \mathcal{A}_{i_1}, X_{i_2}(\omega) \in \mathcal{A}_{i_2}, ..., X_{i_r}(\omega) \in \mathcal{A}_{i_r}\Big) = \prod_{j=1}^{r} \mathbb{P}(X_{i_j}(\omega) \in \mathcal{A}_{i_j}) \qquad (1.3.\ 15)$$

for all $\{i_1, i_2, \cdots, i_r\}$ of r-tuples of $\{1, 2, \cdots, n\}$ and $1 \leq r \leq n$.

Remark: The above just follows from the definition of independence of the corresponding events $X_i^{-1}(A_i)$ in $\mathcal{F}$ by virtue of the definition of random variables.

In light of the above definition, if the X_i's take real values then the following statements are equivalent.

Proposition 1.3.1 *The real valued r.v's X_i are mutually independent if any of the following equivalent properties holds:*

a)$F(x_1, x_2, ..., x_n) = \prod_{i=1}^n F_{X_i}(x_i)$

b)$\mathbf{E}[g_1(X_1)g_2(X_2)...g_n(X_n)] = \prod_{i=1}^n \mathbf{E}[g_i(X_i)]$
for all bounded measurable functions $g_i(.)$.

c) $\mathbf{E}[e^{\imath \sum_{j=1}^n h_j X_j}] = \prod_{j=1}^n C_j(h_j)$ *where $C_k(h_k)$ corresponds to the characteristic function of X_k.*

It is immediate from a) that if the joint probability density exists (then marginal densities exist) then:

$$p(x_1, x_2, ..., x_n) = \prod_{i=1}^n p_{X_i}(x_i)$$

Note it is not enough to show that $C_{X+Y}(h) = C_X(h)C_Y(h)$ for the random variables to be independent. You need to show that $C_{X+Y}(h_1, h_2) = \mathbf{E}[e^{i(h_1 X + h_2 Y)}] = C_X(h_1)C_Y(h_2)$ for all $H - 1, h_2$.

Remark 1.3.1 *An immediate consequence of independence is that it is a property preserved via transformations i.e. if $\{X_i\}_{i=1}^N$ are independent then $\{f_i(X_i)\}_{i=1}^N$, where the $f_i(.)'s$ are measurable mappings (i.e. they define r.v's), form an independent collection of random variables.*

Related to the concept of independence but actually much weaker is the notion of uncorrelatedness.

Definition 1.3.2 *Let X_1 and X_2 be two r.v's with means $m_1 = \mathbf{E}[X_1]$ and $m_2 = \mathbf{E}[X_2]$ respectively. Then the covariance between X_1 and X_2 denoted by $cov(X_1, X_2)$ is defined as follows:*

$$cov(X_1, X_2) = \mathbf{E}[(X_1 - m_1)(X_2 - m_2)] \qquad (1.3.\ 16)$$

Definition 1.3.3 *A r.v. X_1 is said to be uncorrelated with X_2 if $cov(X_1, X_2) = 0$.*

Remark: If two r.v's are independent then they are uncorrelated but not vice versa. The reverse implication holds only if they are jointly Gaussian or normal (which we will see later on).

In statistical literature the normalized covariance (or correlation) between two r.v's is referred to as the correlation coefficient or coefficient of variation.

Definition 1.3.4 *Given two r.v's X and Y with variances σ_X^2 and σ_Y^2 respectively. Then the coefficient of variation denoted by $\rho(X, Y)$ is defined as*

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \qquad (1.3.\ 17)$$

We conclude this section with the result on how the distribution of a vector valued (or a finite collection of r.v's) which is a transformation of another vector valued r.v. can be obtained.

Let $\mathbf{X}$ be a $\Re^n$ valued r.v. and let $\mathbf{Y} = f(\mathbf{X})$ be a $\Re^n$ valued r.v. and $f(.)$ be a 1:1 mapping. Then just as in the case of scalar valued r.v's the joint distribution of Y_i can be obtained from the joint distribution of the $X_i's$ by the extension of the techniques for scalar valued r.v's. These transformations are called the Jacobian transformations.

First note that since $f(.)$ is 1:1, we can write $\mathbf{X} = f^{-1}(\mathbf{Y})$. Let $x_k = [f^{-1}(\mathbf{y})]_k$ i.e. the kth. component of f^{-1}.

Define the so called Jacobian matrix:

$$J_Y(y) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_1}{\partial y_n} & \frac{\partial x_2}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

Then the fact that f is 1:1 implies that $|J_Y| \neq 0$ where $|J| = detJ$. Then as in the scalar case it can be shown that :

$$P_Y(y) = P_X(f^{-1}(y))|J_Y(y)| \qquad (1.3.\ 18)$$

or equivalently

$$\int f(x)p_X(x)dx = \int_D y p_X(f^{-1}(y))|J_Y(y)|dy$$

where D is the range of f.

When the mapping $f(.) : \Re^n \rightarrow \Re^m$ with $m < n$, then the above formula cannot be directly applied. However in this case the way to overcome this is to define a 1:1 transformation $\tilde{f} = col\ (f, f_1)$ with $f_1(.)$ mapping $\Re^n$ to $\Re^{n-m}$. Then use the formula above and integrate out the (n-m) variables to obtain the distribution on $\Re^m$.

Let us now see some typical examples of the application of the above result.

Example 1: Let X and Y be two jointly distributed real valued r.v's with probability density functions $p_{X,Y}(x,y)$. Define $Z = X + Y$ and $R = \frac{X}{Y}$.

Find the probability density functions of Z and R respectively. Let $p_Z()$ denote the density function of Z. Since both Z and R are one dinensional mappings in the range we need to use the method given at the end.

To this end, define the function $f(.,.) : (x,y) \rightarrow (v,w)$ with $V = X$ and $W = X + Y$. Then $f(.)$ defines a 1:1 mapping from $R^2 \rightarrow R^2$. In this case the inverse mapping $f^{-1}(v,w) = (v, w - v$. Therefore the Jacobian is given by:

$$Jacobian \quad = \quad \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

hence

$$p_Z(z) = \int_v p_{X,Y}(v, z-v) dv$$

In particular if X and Y are independent then:

$$p_Z(z) = \int_v p_Y(z-v) P_X(v) dv = \int_v p_X(z-v) p_Y(v) dv = (p_X * p_Y)(z)$$

where $f * g$ denotes the convolution operator. This result can also be easily derived as a consequence of conditioning which is discussed in the next section.

In particular if $Y = \sum_{i=1}^N X_i$ where $\{X_i\}$ are i.i.d. then :

$$p_Y(y) = p_X^{*N}(y)$$

where p^{*N} denotes the N-fold convolution of $p_X(.)$ evaluated at y.

In a similar way to find the density of R, define $f(.,.) = (x,y) \rightarrow (v,w)$ with $v = x$ and $w = \frac{x}{y}$. Then noting that $f^{-1}(v,w) = (v, \frac{v}{w})$ giving $detJ = |\frac{v}{w^2}|$ we obtain:

$$p_R(r) = \int_v p_{X,Y}(v, \frac{v}{r}) |\frac{v}{r^2}| dv$$

Once again if X and Y are independent, then:

$$p_R(r) = \int_v p_X(v) p_Y(\frac{v}{r}) |\frac{v}{r^2}| dv$$

Example 2: Suppose $Y = AX$ where $X = col(X_1, \ldots, X_N)$ is a column vector of N jointly distributed r.v.'s whose joint density is $p_X(x_1, \ldots, x_N) = p_X(x)$. Suppose A^{-1} exists.

Then:

$$p_Y(y) = \frac{1}{|det(A)|} p_X(A^{-1}y)$$

1.3.1 Conditional probabilities and distributions

Definition 1.3.5 *Let A and B be two events with $\mathbb{P}(B) > 0$. Then the conditional probability of the event A occurring given that the event B has occurred denoted by $\mathbb{P}(A/B)$ is defined as:*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \bigcap B)}{\mathbb{P}(B)} \tag{1.3. 19}$$

Remark 1.3.2 *A conditional probability defines a* bona fide *probability measure on $(\Omega, \mathcal{F}$, i.e. if we denote $\mathbb{P}_A(.) = \mathbb{P}(.|A)$ then it defines a probability measure on Ω.*

To show this we need to show:

1. $$0 \leq \mathbb{P}_A(.) \leq 1$$

   with $\mathbb{P}_A(\phi) = 0$ and $\mathbb{P}_A(\Omega) = 1$.

2. *The σ-additivity property $\mathbb{P}_A(\bigcup B_i) = \sum_i \mathbb{P}_A(B_i)$ where $\{B_i\}$ is any countable collection of mutually disjoint events of $\mathcal{F}$.*

The proof of (i) is direct because $\mathbb{P}(\bigcup_i B_i \bigcap A) \leq \mathbb{P}(A)$ as $(\bigcup_i B_i) \bigcap A \subset A$. The second property is also is direct by noting that $(\bigcup_i B_i) \bigcup A = \bigcup_i (B_i \bigcup A)$ and the events $B_i \bigcup A$ are mutually disjoint.

Associated with the notion of conditional probabilities is the notion of conditional independence of two events given a third event. This property plays an important role in the context of Markov processes.

Definition 1.3.6 *Two events A and B are said to be conditionally independent given the event C if:*

$$\mathbb{P}(A \bigcap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C) \tag{1.3. 20}$$

Denote the conditional probability of A given B by $\mathbb{P}_B(A)$. Then from the definition it can be easily seen that $\mathbb{P}_B(A|C) = \mathbb{P}(A|B \bigcap C)$ provided $\mathbb{P}(B \bigcap C) > 0$.

¿From the definition of conditional probability the joint probability of a finite collection of events $\{\mathcal{A}_i\}_{i=1}^n$ such that $\mathbb{P}(\bigcap_{k=1}^n \mathcal{A}_k) > 0$ can be sequentially computed as follows:

$$\mathbb{P}(\bigcap_{k=1}^n \mathcal{A}_k) = \mathbb{P}(\mathcal{A}_1)\mathbb{P}(\mathcal{A}_2|\mathcal{A}_1)\mathbb{P}(\mathcal{A}_3|\mathcal{A}_1, \mathcal{A}_2)....\mathbb{P}(\mathcal{A}_n|\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_{n-1})$$

Remark: Of particular importance is the case when the events are *Markovian* i.e. when $\mathbb{P}(\mathcal{A}_k|\mathcal{A}_1, \mathcal{A}_2, ...\mathcal{A}_{k-1}) = \mathbb{P}(\mathcal{A}_k|\mathcal{A}_{k-1})$ for all k. Then the above sequential computation reduces to:

$$\mathbb{P}(\bigcap_{k=1}^n \mathcal{A}_k) = \mathbb{P}(\mathcal{A}_1) \prod_{k=2}^n \mathbb{P}(\mathcal{A}_k|\mathcal{A}_{k-1})$$

or in other words the joint probability can be deduced from the probability of the first event and the conditional probabilities of subsequent pairs of events. It is easy to see that such a property implies that the event $\mathcal{A}_k$ is conditionally independent of the events $\mathcal{A}_l;\ l \leq k-2$ given the event $\mathcal{A}_{k-1}$.

The definition of conditional probabilities and the sequential computation of the joint probabilities above leads to a very useful formula of importance in estimation theory called the *Bayes' Rule*

Proposition 1.3.2 *Let $\{B_k\}_{k=1}^{\infty}$ be a countable collection of disjoint events i.e. $B_i \bigcap B_j = \phi; i \neq j$; such that $\bigcup_{n=1}^{\infty} B_n = \Omega$. Then for any event A :*

$$\mathbb{P}(A) = \sum_{k=1}^{\infty} \mathbb{P}(A|B_k)\mathbb{P}(B_k) \tag{1.3. 21}$$

and

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i)} \tag{1.3. 22}$$

The first relation is known as the law of total probability which allows us to compute the probability of an event by computing the conditional probability on 'simpler' events which are disjoint and exhaust the space Ω. The second relation is usually referred to as Bayes' rule. The importance arises in applications where we cannot observe a particular event but instead observe another event whose probability can be inferred by the knowledge of the event conditioned on simpler events. Then Bayes' rule just states that the conditional probability can be computed from the knowledge of how the observed event interacts with other 'simpler' events. The result also holds for a finite decomposition with the convention that if $P(B_k) = 0$ then the corresponding conditional probability is set to 0.

Once we have the definition of conditional probabilities we can readily obtain conditional distributions. Let us begin by the problem of computing the conditional distribution of a real valued r.v. X given that $X \in \mathcal{A}$ where $\mathbb{P}(X \in \mathcal{A}) > 0$. In particular let us take $\mathcal{A} = (a, b]$. We denote the conditional distribution by $F_{(a,b]}(x)$.

Now by definition of conditional probabilities:

$$F_{(a,b]}(x) = \frac{\mathbb{P}(X \in (-\infty, x] \bigcap (a, b])}{\mathbb{P}(X \in (a, b])}$$

Now, for $x \leq a$ the intersection of the events is null. Hence, it follows that :

$$\begin{aligned} F_{(a,b]}(x) &= 0 \ \ x \leq a \\ &= \frac{F(x) - F(a)}{F(b) - F(a)} \ \ x \in (a, b] \\ &= 1 \ \ x \geq b \end{aligned}$$

More generally we would like to compute the conditional distribution of a random variable X given the distribution of another r.v. Y. This is straight forward in light of the definition of conditional probabilities.

$$F_{X|Y}(x; y) = \frac{F(x, y)}{F_Y(y)}$$

where $F_{X|Y}(x; y) = P(X \leq x|Y \leq y)$.

¿From the point of view of applications a more interesting conditional distribution is the conditional distribution of X given that $Y = y$. This corresponds to the situation when we 'observe' a particular occurrence of Y given by y. In the case when Y is continuous this presents a problem since the event that $\{Y = y\}$ has zero probability. However, when a joint density exists and hence individual densities will exist, then the conditional density $p(x|Y = y)$ can be defined by:

$$p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)} \tag{1.3. 23}$$

Note this is not a consequence of the definition since densities are not probabilities (except in the discrete case) and so we append a proof:

The way to derive this is by noting that by the definition of the conditional probability

$$\mathbb{P}(X \in (x, x + \Delta x]; Y \in (y, y + \Delta y]) = \frac{\mathbb{P}(X \in (x, x + \Delta x], Y \in (y, y + \Delta y])}{\mathbb{P}(Y \in (y, y + \Delta y])}$$

Now, for small $\Delta x, \Delta y$ we obtain by the definition of densities:

$$\mathbb{P}(X \in (x, x + \Delta x]; Y \in (y, y + \Delta y]) = \frac{p(x,y)\Delta x \Delta y}{p_Y(y)\Delta y}$$

hence taking limits as $\Delta x \to 0$ and denoting $\lim_{\Delta x \to 0} \frac{\mathbb{P}(X \in (x, x+\Delta x; Y \in (y, y+\Delta y])}{\Delta x} = \mathbb{P}_{X|Y}(x|y)$ (it exists) we obtain the required result.

Note from the definition of the conditional density we have:

$$\int p_{X|Y}(x|y)dx = 1$$

and

$$\int p_{X|Y}(x|y)p_Y(y)dy = p_X(x)$$

Remark: Although we assumed that X and Y are real valued r.v's the above definitions can be directly used in the case when X and Y are vector valued r.v's where the densities are replaced by the joint densities of the component r.v's of the vectors. We assumed the existence of densities to define the conditional distribution (through the definition of the conditional density) given a particular value of the r.v. Y. This is strictly not necessary but to give a proof will take us beyond the scope of the course.

Once we have the conditional distribution we can calculate conditional moments. From the point of view of applications the conditional mean is of great significance which we discuss in the next subsection.

1.3.2 Conditional Expectation

¿From the definition of the conditional distribution one can define the conditional expectation of a r.v. given the other. This can be done as follows:

$$\mathbf{E}[X|Y = y] = \int_X x p_{X|Y}(x|y)dx$$

note what remains is a function of y. Let us call it $g(y)$. Now if we substitute the random variable Y instead of y, then $g(Y)$ will be a random variable. We identify this as the result of conditioning

Xgiven the r.v. Y. In general it can be shown using measure theory that indeed one can define the conditional expectation :

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

where g(.) is a measurable function (as defined above).

The conditional expectation has the following properties:

Proposition 1.3.3 *Let X and Y be jointly distributed r.v's. Let $f(.)$ be a measurable function with $\mathbf{E}[f(X)] < \infty$ i.e. f(X) is* integrable. *Then:*

a) If X and Y are independent, $\mathbf{E}[f(X)|Y] = \mathbf{E}[f(X)]$.

b) If X is a function of Y say $X = h(Y)$, then $\mathbf{E}[f(X)|Y] = f(X) = f(h(Y))$.

c) $\mathbf{E}[f(X)] = \mathbf{E}[\mathbf{E}[f(X)|Y]]$.

d) $\mathbf{E}[h(Y)f(X)|Y] = h(Y)\mathbf{E}[f(X)|Y]$ for all functions $h(.)$ such that $\mathbf{E}[h(Y)f(X)]$ is defined.

Proof: The proof of a) follows from the fact that the conditional distribution of X given Y is just the marginal distribution of X since X and Y are independent. The proof of b) can be shown as follows: let $a < h(y)$ then $\mathbb{P}(X \leq a, Y \in (y, y+\delta]) = \mathbb{P}(h(Y) \leq a, Y \in (y, y+\delta])$ which is 0 for sufficiently small δ (assuming h(.) is continuous). If $a \geq h(y)$ then the above distribution is 1. This amounts to the fact that $F(x/y) = 0$ if $x < h(y)$ and $F(x/y) = 1$ for all $x \geq h(y)$ and thus the distribution function has mass 1 at $h(y)$. Hence $\mathbf{E}[f(X)|Y = y] = f(h(y))$ from which the result follows. The proof of c) follows (in the case of densities) by noting that : $\mathbf{E}[f(X)] = \int_Y \int_X x p_{X|Y}(x/y) p_Y(y) dx dy$ and $\int_Y p_{X|Y}(x|y) P_Y(y) dy = p_X(dx)$. Part d) follows from b) and the definition.

Properties c) and d) give rise to a useful characterization of the conditional expectations. This is called the *orthogonality principle.* This states that the difference between a r.v. and its conditional expectation is uncorrelated with any function of the r.v. on which it is conditioned i.e,

$$\mathbf{E}[(X - \mathbf{E}[X|Y])h(Y)] = 0 \tag{1.3. 24}$$

In the context of mean squared estimation theory the above is just a statement of the orthogonality principle with $\mathbf{E}[XY]$ playing the role of the inner-product on the Hilbert space of square integrable r.v's i.e. r.v's such that $\mathbf{E}[X]^2 < \infty$. Thus the conditional expectation can be seen as a projection onto the subspace spanned by Y with the inner-product as defined. This will be discussed in Chapters 3 and 4.

The next property of conditional expectations is a very important one. This is at the heart of estimation theory. Suppose X and Y are jointly distributed r.v's. Suppose we want to approximate X by a function of Y i.e. $\hat{X} = g(Y)$. Then, the following result shows that amongst all measurable functions of Y, g(.) that can be chosen, choosing $g(Y) = \mathbf{E}[X|Y]$ minimizes the mean squared error.

Proposition 1.3.4 *Let X and Y be jointly distributed. Let $g(Y)$ be a measurable function of Y such that both X and $g(Y)$ have finite second moments. Then the mean squared error $\mathbf{E}[X - g(Y)]^2$ is minimized over all choices of the functions $g(.)$ by choosing $g(Y) = \mathbf{E}[X|Y]$.*

Proof: This just follows by the completion of squares argument. Indeed,

$$\begin{aligned}
\mathbf{E}[X-g(Y)]^2 &= \mathbf{E}[X-\mathbf{E}[X|Y]+\mathbf{E}[X|Y]-g(Y)]^2 \\
&= \mathbf{E}[X-\mathbf{E}[X|Y]]^2+2\mathbf{E}[(X-\mathbf{E}[X|Y])(\mathbf{E}[X|Y]-g(Y))] \\
&\quad +\mathbf{E}[\mathbf{E}[X|Y]-g(Y)]^2 \\
&= \mathbf{E}[X-\mathbf{E}[X|Y]]^2+\mathbf{E}[\mathbf{E}[X|Y]-g(Y)]^2
\end{aligned}$$

where we have used properties b and c to note that $\mathbf{E}[(\mathbf{E}[X|Y]-g(Y))(X-\mathbf{E}[X|Y])]=0$. Hence since the right hand side is the sum of two squares and we are free to choose g(Y), the right hand side is minimized by choosing $g(Y)=\mathbf{E}[X|Y]$.

Remark 1.3.3 *Following the same proof it is easy to show that the constant C which minimizes $\mathbf{E}[(X-C)^2]$ is $C=\mathbf{E}[X]$ provided $E[X^2]<\infty$.*

Finally we conclude our discussion of conditional expectations by showing another important property associated with conditional expectations. This is related to the fact that if we have a 1:1 transformation of a r.v. then conditioning w.r.t to the r.v. or its transformed version gives the same result.

Proposition 1.3.5 *Let X and Y be two jointly distributed r.v's. Let $\phi(.)$ be a 1:1 mapping. Then:*

$$\mathbf{E}[X|Y]=\mathbf{E}[X|\phi(Y)]$$

Proof: Note by property c) of conditional expectations for any bounded measurable function $h(Y)$ we have:

$$\mathbf{E}[h(Y)X]=\mathbf{E}[h(Y)\mathbf{E}[X|Y]] \tag{A}$$

Denote $g(Y)=\mathbf{E}[X|Y]$ and $\tilde{Y}=\phi(Y)$.

Once again by property c) we have for all bounded measurable functions $j(.)$:

$$\mathbf{E}[j(\tilde{Y})X]=\mathbf{E}[j(\tilde{Y})\mathbf{E}[X/\tilde{Y}]]$$

Now since $\phi(.)$ is 1:1 it implies that ϕ^{-1} exists and moreover the above just states that

$$\mathbf{E}[j\circ\phi(Y)X]=\mathbf{E}[j\circ\phi(Y)\tilde{g}\circ\phi(Y)]$$

where we denote $\mathbf{E}[X/\tilde{Y}]=\tilde{g}(\tilde{Y})=\tilde{g}\circ\phi(Y)$ by definition of ϕ. Now since ϕ^{-1} exists then we can take $j(.)=h\circ\phi^{-1}$ for all bounded measurable functions h(.). Therefore we have

$$\mathbf{E}[h(Y)X]=\mathbf{E}[h(Y)\tilde{g}\circ\phi(Y)] \tag{B}$$

Comparing (A) and (B) we see that $\mathbf{E}[X|Y]=\mathbf{E}[X/\phi(Y)]$ with $\tilde{g}(.)=g\circ\phi^{-1}(.)$.

Let us see a non-trivial example of conditional expectations where we use the abstract property to obtain the result.

Example : Let X be a real valued r.v. with probability density $p_X(.)$. Let $h(.)$ be a measurable function such that $\mathbf{E}|h(X))|<\infty$. Show that the conditional expectation

$$\mathbf{E}[h(X)|X^2]=h(\sqrt{X^2})\frac{p_X(\sqrt{X^2})}{p_X(\sqrt{X^2})+p_X(-\sqrt{X^2})}+h(-\sqrt{X^2})\frac{p_X(-\sqrt{X^2})}{p_X(\sqrt{X^2})+p_X(-\sqrt{X^2})} \tag{1.3. 25}$$

Note this is not trivial since $X^2 : \Re \to \Re$ is not 1:1 (since knowing X^2 does not tell us whether X is positive or negative. Moreover calculating the conditional densities involves generalized functions (delta functions)(see below) . The way to show this is by using the property that for any measurable and integrable function $g(.)$ we have:

$$\mathbf{E}[h(X)g(X^2)] = \mathbf{E}[\mathbf{E}[h(X)|X^2]g(X^2)]$$

Indeed, we can actually show that the first term in the conditional expectation just corresponds to $\mathbf{E}[h(X)\mathbb{I}_{[X>0]}|X^2]$. Let is show it by using the property of conditional expectations.

Let us consider the first term on the r.h.s of (1.3. 25).:

$$\begin{aligned}
\mathbf{E}[h(\sqrt{X^2})\frac{p_X(\sqrt{X^2})}{p_X(\sqrt{X^2})+p_X(-\sqrt{X^2})}g(X^2)] &= \int_{-\infty}^{\infty} h(\sqrt{x^2})\frac{p_X(\sqrt{x^2})}{p_X(\sqrt{x^2})+p_X(-\sqrt{x^2})}g(x^2)p_X(x)dx \\
&= \int_0^{\infty} h(\sqrt{x^2})\frac{p_X(x)}{p_X(x)+p_X(-x)}g(x^2)p_X(x)dx + \\
&\quad \int_{-\infty}^{0} h(\sqrt{x^2})\frac{p_X(x)}{p_X(x)+p_X(-x)}g(x^2)p_X(x)dx \\
&= \int_0^{\infty} h(\sqrt{x^2})\frac{p_X(x)}{p_X(x)+p_X(-x)}g(x^2)p_X(x)dx + \\
&\quad \int_0^{\infty} h(\sqrt{x^2})\frac{p_X(x)}{p_X(x)+p_X(-x)}g(x^2)p_X(-x)dx \\
&= \int_0^{\infty} h(x)g(x^2)p_X(x)dx = \mathbf{E}[h(X)g(X^2)\mathbb{I}_{[X>0]}]
\end{aligned}$$

where in the 3rd. step we substitute $-x$ for x in the second integral.

Similarly for the second term. Indeed one way of looking at the result is

$$\mathbf{E}[h(X)|X^2] = h(\sqrt{X^2})\mathbb{P}(X>0|X^2) + h(-\sqrt{X^2})\mathbb{P}(X<0|X^2)$$

It is instructive to check that going via the route of explicitly calculating the conditional density gives the conditional density $p_{X|X^2}(x|y)$ as:

$$p_{X|X^2}(x|y) = \frac{1}{p_X(\sqrt{y})+p_X(-\sqrt{y})}\left(p_X(\sqrt{y})\delta(x-\sqrt{y}) + p_X(-\sqrt{y})\delta(x+\sqrt{y})\right)$$

where $\delta(.)$ denotes the Dirac delta function. Hence $\mathbf{E}[X|X^2 = y] = \int_{-\infty}^{\infty} h(x)p_{X|X^2}(x|y)dx$ which can be easily seen to be the relation above and also

$$\mathbb{P}(X>0|X^2=y) = \int_0^{\infty} p_{X|X^2}(x|y)dx = \frac{p_X(\sqrt{y})}{p_X(\sqrt{y})+p_X(-\sqrt{y})}$$

1.4 Gaussian or Normal random variables

We now discuss properties of Gaussian or Normal r.v's. since they play an important role in the modeling of signals and the fact that they also possess special properties which are crucial in the development of estimation theory. Throughout we will work with vector valued r.v's. Note in probability we usually refer to Gaussian distributions while statisticians refer to them as Normal distributions. In these notes we prefer the terminology Gaussian.

We first introduce some notation: For any two vectors $x, y \in \Re^n$, the inner-product is denoted by $[x, y] = \sum_{i=1}^n x_i y_i$. x^* denotes the transpose (row vector) of the (column) vector x and thus $[x, y] = x^* y$. For any $n \times n$ matrix A, A^* denotes the transpose or adjoint matrix. $[Ax, x]$ is the quadratic form given by $[Ax, x] = x^* A^* x$. If the vectors or matrices are complex then the $*$ operation denotes the complex conjugate transpose.

We first develop some general results regarding random vectors in $\Re^n$.

Given a vector valued r.v $X \in \Re^n$ (i.e. n r.v's which are jointly distributed) the mean is just the column vector with elements $m_i = \mathbf{E}[X_i]$ and the covariance is a matrix of elements $R_{i,j} = \mathbf{E}[(X_i - m_i)(X_j - m_j)]$ and hence $R_{i,j} = R_{j,i}$ or the matrix R is self-adjoint (symmetric). In vectorial terms this can be written as $R = \mathbf{E}[(X - m)(X - m)^*]$.

Definition 1.4.1 *Let X be a $\Re^n$ valued r.v. then the characteristic functional of X denoted by $C_X(h)$ is*

$$\mathbf{E}[e^{\imath[h,X]}] = \int_{\Re^n} e^{\imath[h,x]} dF(x) \tag{1.4. 26}$$

where $h \in \Re^n$ and $F(x)$ denotes the joint distribution of $\{X_i\}_{i=1}^n$.

It can be easily seen that the characteristic functional possesses the properties given in 1.2.1.

Let $X \in \Re^n$ and define the r.v.

$$Y = AX + b \tag{A}$$

where A is a $m \times n$ matrix and $b \in \Re^m$ is a fixed vector.

The following proposition establishes the relations between the mean, covariance, characteristic functions and distributions of Y in terms of the corresponding quantities of X.

Proposition 1.4.1 *a) For Y given above; $\mathbf{E}[Y] = A\mathbf{E}[X] + b$ and $cov(Y) = Acov(X)A^*$*

b) If A is a $n \times n$ (square matrix) then denoting $m = \mathbf{E}[X]$

$$\mathbf{E}[AX, X] = [Am, m] + Trace[Acov(X)]$$

c) Let $Y \in \Re^n$ be any r.v. with finite variance. Then there exists a $\Re^n$ valued r.v. X with mean zero and variance $= I$ (the identity $n \times n$ matrix) such that (A) holds.

d) $C_Y(h) = e^{\imath[h,b]} C_X(A^*h)$ with $h \in \Re^m$.

e) If X possesses a density $p_X(x)$ then if A is a non-singular $n \times n$ matrix then

$$p_Y(y) = \frac{1}{|det(A)|} p_X(A^{-1}(y - b))$$

Proof: a) The mean is direct.

$$\begin{aligned} cov(Y) &= \mathbf{E}[(Y - \mathbf{E}[Y])(Y - \mathbf{E}[Y])^*] \\ &= \mathbf{E}[A(X - \mathbf{E}[X])(X - \mathbf{E}[X])^* A^* \\ &= Acov(X)A^* \end{aligned}$$

b) Define $\tilde{X} = X - \mathbf{E}[X]$ then $\tilde{X}$ is a zero mean r.v. with $cov(\tilde{X}) = cov(X)$. Hence,

$$\begin{aligned}
\mathbf{E}[AX, X] &= \mathbf{E}[A(\tilde{X} + m), (\tilde{X} + m)] \\
&= \mathbf{E}[A\tilde{X}, \tilde{X}] + \mathbf{E}[A\tilde{X}, m] + \mathbf{E}[Am, \tilde{X}] + [Am, m] \\
&= \mathbf{E}[A\tilde{X}, \tilde{X}] + [Am, m] \\
&= \sum_{i,j=1}^{n} A_{i,j} cov(X)_{i,j} + [Am, m] \\
&= Trace(Acov(X)) + [Am, m]
\end{aligned}$$

c) The proof of part c follows from the fact that since the covariance of Y is $\mathbf{E}[(Y - \mathbf{E}[Y])(Y - \mathbf{E}[Y])^*]$ it is non-negative definite and (symmetric). Hence, from the factorization of symmetric non-negative definite matrices $Q = \Lambda\Lambda^*$ we can take $A = \Lambda$ and $b = \mathbf{E}[Y]$

d) This follows from the fact that :

$$C_Y(h) = \mathbf{E}[e^{\imath[h, AX+b]}] = e^{\imath[h,b]}\mathbf{E}[e^{\imath[A^*h, X]}] = e^{\imath[h,b]}C_X(A^*h)$$

e) This follows from the Jacobian formula by noting that $f^{-1}(x) = A^{-1}(x - b)$ where $f(x) = Ax + b$ and $detA^{-1} = \frac{1}{detA}$.

Let us now consider Gaussian r.v's.

A r.v. $X \in \Re$ is Gaussian with mean m and var σ^2, denoted $N(m, \sigma^2)$, if its probability density function is given by:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad (1.4.\ 27)$$

The corresponding characteristic function of a Gaussian r.v. X $\sim$ N(m,σ^2) is:

$$C(h) = e^{\imath mh - \frac{1}{2}\sigma^2 h^2} \qquad (1.4.\ 28)$$

The extension of these to vector $\Re^n$ valued r.v's (or a finite collection of r.v's) is done through the definition of the characteristic functional.

Definition 1.4.2 *A $\Re^n$ valued r.v. X (or alternatively a collection of $\Re$ valued r.v's $\{X_i\}_{i=1}^n$) is (are) said to be Gaussian if and only if the characteristic functional can be written as :*

$$C_X(h) = e^{\imath[h,m] - \frac{1}{2}[Rh,h]} \qquad (1.4.\ 29)$$

where $m = \mathbf{E}[X]$ and R denotes the covariance matrix with entries $R_{i,j} = cov(X_i X_j)$.

Remark: The reason that jointly Gaussian r.v's are defined through the characteristic functional rather than the density is that R^{-1} need not exist. Note it is not enough to have each r.v. to have a marginal Gaussian distribution for the collection to be Gaussian. An example of this is given in the exercises.

We now study the properties of jointly Gaussian r.v's.

Proposition 1.4.2 *a) Linear combinations of jointly Gaussian r.v's are Gaussian.*

b) Let $\{X_i\}_{i=1}^n$ be jointly Gaussian. If the r.v's are uncorrelated then they form a collection of independent r.v's

Proof: a) Let $\{X_i\}$ be jointly Gaussian. Define:

$$Y = \sum_{i=1}^{n} a_i X_i = AX$$

where A is a $1 \times n$ matrix.

Then applying part d) of the proposition 1.4.1 we obtain:

$$C_Y(h) = C_X(A^*h) = e^{\imath[A^*h,m]-\frac{1}{2}[ARA^*h,h]}$$

implying that the r.v. Y is Gaussian with mean Am and variance (ARA^*).

Note we could also have taken Y to be a $\Re^m$ valued r.v. formed by linear combinations of the vector $X = col(X_1, X_2, ..., X_n)$.

b) Since the r.v's are uncorrelated, $cov(X_i X_j) = 0$ for $i \neq j$. Hence, R is a diagonal matrix with diagonal elements $\sigma_i^2 = var(X_i)$. From the definition of jointly Gaussian r.v's in this case the characteristic functional can be written as:

$$\begin{aligned} C(h) &= e^{\imath \sum_{j=1}^{n} h_j m_j - \frac{1}{2}\sum_{j=1}^{n} \sigma_j^2 h_j^2} \\ &= \prod_{j=1}^{n} e^{\imath h_j m_j - \frac{1}{2}\sigma_j^2 h_j^2} \\ &= \prod_{j=1}^{n} C_{N_j}(h_j) \end{aligned}$$

where $C_{N_j}(h_j)$ corresponds to the characteristic function of a $N(m_j, \sigma_j^2)$ r.v. Hence since the characteristic functional is the product of individual characteristic functions the r.v's are independent.

With the aid of the above two properties one can directly calculate the joint density of a finite collection of Gaussian r.v's provided that the covariance matrix is non-singular. This is often used to define jointly Gaussian r.v's but is less general than the definition via characteristic functionals since the inverse of the covariance matrix must exist.

Proposition 1.4.3 *Let $\{X_i\}_{i=1}^n$ be jointly Gaussian with mean $m = col(m_1, m_2, ..., m_n)$ and covariance matrix R with $r_{i,j} = \mathbf{E}[(X_i - m_i)(X_j - m_j)] = cov(X_i X_j)$. If R is non-singular then the joint probability density is given by*

$$p(x_1, x_2, x_3, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}(detR)^{\frac{1}{2}}} e^{-\frac{1}{2}[R^{-1}(x-m),(x-m)]} \qquad (1.4.\ 30)$$

Proof: The proof of this follows by using parts c) and e) of proposition 1.4.1 and part a above. Suppose that first $\{Z_i\}$ are independent $N(0,1)$ r.v's then by independence the joint density can be written as :

$$p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{||x||^2}{2}}$$

Now use the fact that if R is non-singular then $R = LL^*$ with L non-singular and $X = LZ + m$ where m denotes the column vector of means of X i.e. $m_i = \mathbf{E}[X_i]$. Now use part e) of proposition 1.4.1 with $A = L$ and then noting that $det(L) = \sqrt{detR}$ and substituting above we obtain:

$$p(x_1, x_2, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}(detR)^{\frac{1}{2}}} e^{-\frac{1}{2}||L^{-1}(x-m)||^2}$$

and then $||L^{-1}(x-m)||^2 = [R^{-1}(x-m),(x-m)]$.

We now give a very important property associated with the conditional distributions of Gaussian r.v's. Let X and Y be jointly distributed Gaussian r.v's taking values in $\Re^n$ and $\Re^m$ respectively with means m_X and m_Y and covariances Σ_X and Σ_Y. Assume that the covariance of Y is non-singular. Denote $cov(X,Y) = \Sigma_{XY}$ and $cov(Y,X) = \Sigma_{YX}$. Note by definition $\Sigma_{XY} = \Sigma_{YX}^*$.

Then we have the following proposition.

Proposition 1.4.4 *If X and Y are two jointly distributed Gaussian vector valued r.v's. Then the conditional distribution of X given Y is Gaussian with mean:*

$$\mathbf{E}[X|Y] = m_{X/Y} = m_X + \Sigma_{XY}\Sigma_Y^{-1}(Y - m_Y) \qquad (1.4.\ 31)$$

and covariance

$$cov(X/Y) = \Sigma_{X/Y} = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^* \qquad (1.4.\ 32)$$

Proof: The idea of the proof is to transform the joint vector (X,Y) into an equivalent vector r.v. whose first component is independent of the second and then identifying the mean and covariance of the resultant first component.

Define the $\Re^{n+m}$ vector valued r.v. $Z = col(X,Y)$. Let M be the matrix defined by

$$M = \begin{bmatrix} I_n & -A \\ 0 & I_m \end{bmatrix}$$

Define $Z' = col(X',Y)$ by

$$Z' = MZ$$

Then Z' is a Gaussian $\Re^{n+m}$ valued r.v. with mean $E[Z'] = M\mathbf{E}[Z]$ and covariance $cov(Z') = Mcov(Z)M^*$. In particular, due to the structure of M we have:

$$Z' \sim N\left(\begin{pmatrix} m_X - Am_Y \\ m_Y \end{pmatrix} ; \begin{bmatrix} \Sigma_X - A\Sigma_{XY}^* - \Sigma_{XY}A^* + A\Sigma_Y A^* & \Sigma_{XY} - A\Sigma_Y \\ \Sigma_{XY}^* - \Sigma_Y A^* & \Sigma_Y \end{bmatrix} \right)$$

Now suppose we choose A such that the cross terms in the covariance of Z' are zero, this will imply that X' and Y are independent. This implies that

$$\Sigma_{XY} - A\Sigma_Y = 0$$

or

$$A = \Sigma_{XY}\Sigma_Y^{-1}$$

Substituting for A we obtain:

$$\mathbf{E}[X'] = m_X - \Sigma_{XY}\Sigma_Y^{-1}m_Y$$

and

$$cov(X') = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^*$$

It remains to compute the conditional distribution of X given Y and show that it is Gaussian with the mean and covariance as stated in the statement.

Note by the construction of $Z' = MZ$, it is easy to see that M is invertible with :

$$M^{-1} = \begin{bmatrix} I_n & A \\ 0 & I_m \end{bmatrix}$$

Now the conditional density of X given Y is just:

$$p_{X/Y}(x/y) = \frac{p(x,y)}{p_Y(y)} = \frac{p_Z(x,y)}{p_Y(y)}$$

Now since $Z = M^{-1}Z'$ we can compute p_Z from $p_{Z'}$ since Z and Z' are related by a 1:1 transformation and note that the two components of Z' are independent. Hence, noting that $Y = y$ and $X = x$ implies that $X' = x - Ay$ we obtain:

$$p_{X/Y}(x/y) = \frac{1}{|detM|}\frac{p_{Z'}(x - Ay, y)}{p_Y(y)}$$

Noting that $p_{Z'}(x,y) = p_{X'}(x)p_Y(y)$ and $detM = 1$, by independence of X' and Y we obtain that

$$p_{X/Y}(x/y) = p_{X'}(x - Ay)$$

But by construction X' is Gaussian and hence the above result states that the conditional distribution of X given $Y = y$ is just a Gaussian distribution with mean:

$$m_{X/Y} = m_{X'} + \Sigma_{XY}\Sigma_Y^{-1}y$$

and

$$cov(X/Y) = Cov(X') = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^*$$

and the proof is done.

Remark: The above proposition shows that not only is the conditional distribution of jointly Gaussian r.v's Gaussian but also the conditional mean is an affine function of the second r.v. and the conditional covariance does not depend on the second r.v. i.e. the conditional covariance is a constant and not a function of the r.v. on which the conditioning takes place. This result only holds for Gaussian distributions and thus the problem of finding the best mean squared estimate can be restricted to the class of linear maps.

1.5 Probabilistic Inequalities and Bounds

In this section we give some important inequalities for probabilities and bounds associated with random variables. These results will play an important role in the study of the convergence issues.

Proposition 1.5.1 *(***Markov Inequality)**

Let X be a r.v. and $f(.) : \Re \to \Re_+$ such that $\mathbf{E}[f(X)] < \infty$. Then

$$\mathbb{P}\left(f(X) \geq a\right) \leq \frac{\mathbf{E}[f(X)]}{a} \quad ; \quad a > 0 \qquad (1.5.\ 33)$$

Proof: This just follows from the fact that :

$$a\mathbf{1}_{(f(X)\geq a)} \leq f(X)$$

Hence taking expectations gives:

$$a\mathbb{P}\left(f(X) \geq a\right) \leq \mathbf{E}[f(X)]$$

A special case of the Markov inequality is the so-called Chebychev inequality which states that for a r.v. X with mean m and variance σ^2

$$\mathbb{P}\left(|X-m| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}$$

This amounts to choosing the function $f(X) = (X-m)^2$ in the Markov inequality.

An immediate consequence of Chebychev's inequality is the fact that if a r.v. X has variance equal to 0 then such a random variable is *almost surely* a constant.

Definition 1.5.1 *An event is said to be P-almost sure(*usually abbreviated as $\mathbb{P}$-a.s.*) if the probability of that event is 1 i.e.*

$$\mathbb{P}\{\omega : \omega \in A\} = 1$$

Now if X is any r.v. with variance 0 note that :

$$\{\omega : |X(\omega)-m| > 0\} = \bigcup_{n=1}^{\infty}\{\omega : |X(\omega)-m| \geq \frac{1}{n}\}$$

Hence,

$$\mathbb{P}\left(|X(\omega)-m| > 0\right) \leq \sum_{n=1}^{\infty}\mathbb{P}\left(|X(\omega)-m| \geq \frac{1}{n}\right)$$

Using Chebychev's inequality and noting $\sigma^2 = 0$ we obtain that the probability of the event that $|X-m| > 0$ is 0 or $\mathbb{P}\left(X(\omega) = m\right) = 1$.

The next set of inequalities concerns moments or expectations of r.v's. Perhaps one of the most often used ones is the Cauchy-Schwarz inequality. We state it in terms of vectors.

Proposition 1.5.2 (Cauchy-Schwarz Inequality)

Let X ad Y be two r.v's such that $\mathbf{E}[||X||^2] < \infty$ and $\mathbf{E}[||Y||^2] < \infty$. Then $\mathbf{E}(|[X,Y]|) < \infty$ and

$$(\mathbf{E}([X,Y]))^2 \leq \mathbf{E}[||X||^2]\mathbf{E}[||Y||^2] \tag{1.5. 34}$$

Proof: We assume that $\mathbf{E}[||Y||^2] > 0$ since if $\mathbf{E}[||Y||^2 = 0$ then in light of the result above we have $Y = 0$ a.s. and hence the inequality is trivially satisfied (it is an equality).

Hence for $\mathbf{E}[||Y||^2] > 0$ we have:

$$\begin{aligned}
0 &\leq \mathbf{E}[||Y||^2]\mathbf{E}\left[(X - \frac{\mathbf{E}([X,Y])}{\mathbf{E}[||Y||^2]}Y)(X - \frac{\mathbf{E}([X,Y])}{\mathbf{E}[||Y||^2]}Y))^*\right] \\
&\leq \mathbf{E}[||Y||^2]\left(\mathbf{E}[||X||^2 - \frac{(\mathbf{E}([X,Y]))^2}{\mathbf{E}[||Y||^2]}\right) \\
&\leq \mathbf{E}[||X||^2]\mathbf{E}[||Y||^2] - (\mathbf{E}([X,Y]))^2
\end{aligned}$$

which is the required result.

Using the Cauchy-Schwarz inequality we can prove a one-sided version of Chebychev's inequality that is often useful. This is sometimes referred to as Cantelli's inequality.

Proposition 1.5.3 *(Cantelli's inequality)*
Let X be a random variable with mean $E[X] = m$ and finite variance. Then:

$$\mathbb{P}\left(X(\omega) - m \geq \varepsilon\right) \leq \frac{var(X)}{var(X) + \varepsilon^2}$$

Moreover, we can obtain the following version of Chebychev's inequality:

$$\mathbb{P}\left(|X - m| \geq \varepsilon\right) \leq \frac{2var(X)}{var(X) + \varepsilon^2}$$

Proof: Without loss of generality we assume $m = 0$. Then,

$$\varepsilon - X \leq (\varepsilon - X)\mathbb{1}_{[\varepsilon > X]}$$

Taking expectations on both sides we obtain:

$$\varepsilon \leq \mathbf{E}[(\varepsilon - X)\mathbb{1}_{[\varepsilon > X]}$$

Hence squaring both sides and using the Cauchy=Schward inequality we obtain:

$$\begin{aligned} \varepsilon^2 &\leq (\mathbf{E}(\varepsilon - X)\mathbb{1}_{[\varepsilon > X]})^2 \\ &\leq \mathbf{E}(\varepsilon - X)^2 \mathbb{P}(\varepsilon > X) \end{aligned}$$

Now expanding the r.h.s. we obtain:

$$\mathbb{P}(X < \varepsilon) \geq \frac{\varepsilon^2}{\varepsilon^2 + var(X)}$$

and noting that $\mathbb{P}(X \geq \varepsilon) = 1 - \mathbb{P}(\varepsilon < X)$ the result follows.

In a similar way by considering $Y = -X$ and noting $var(Y) = var(X)$, using the previous result above we obtain:

$$\mathbb{P}(Y \geq \varepsilon) = \mathbb{P}(-\varepsilon \leq X) \leq \frac{var(X)}{var(X) + \varepsilon^2}$$

and therefore combining the two results we have:

$$\mathbb{P}(|Y| \geq \varepsilon) \leq \frac{2var(X)}{var(X) + \varepsilon^2}$$

Another important inequality known as Jensen's inequality allows us to relate the mean of a function of a r.v. and the function of the mean of the r.v.. Of course we need some assumptions on the functions. These are called convex functions. Recall, a function $f(.) : \Re \to \Re$ is said to be convex downward (or convex) if for any $a, b \geq 0$ such that $a + b = 1$:

$$f(ax + by) \leq af(x) + bf(y)$$

if the inequality is reverse then the function is said to be convex upward (or concave).

Proposition 1.5.4 (Jensen's Inequality)
Let $f(.)$ be convex downward (or convex) and $\mathbf{E}|X| < \infty$. Then:

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)] \tag{1.5. 35}$$

Conversely if f(.) is concave (or convex upward) then

$$\mathbf{E}[f(X)] \leq f(\mathbf{E}[X]) \tag{1.5. 36}$$

Proof: We prove the first part. The second result follows similarly by noting that if f(.) is convex then -f is concave.

Now if f(.) is convex downward then it can be shown that there exists a real number $c(y)$ such that

$$f(x) \geq f(y) + (x-y)c(y)$$

for any $x, y \in \Re$. (This can be readily seen if f is twice differentiable since the convex property implies that the second derivative is non-negative)

Hence choose $x = X$ and $y = \mathbf{E}[X]$. Then we have:

$$f(X) \geq f(\mathbf{E}[X]) + (X - \mathbf{E}[X])c(\mathbf{E}[X])$$

Now taking expectations on both sides we obtain the required result by noting that the expectation of the second term is 0.

Remark 1.5.1 *Noting that conditional distributions are bona-fide distributions, it follows that Jensen's inequality also holds for conditional expectations. i.e. if $f(.)$ is convex then:*

$$f(\mathbf{E}[X|Y]) \leq \mathbf{E}[f(X)|Y]$$

One immediate consequence of Jensen's inequality is that we can obtain relations between different moments of r.v's.. These inequalities are called Lyapunov's inequalities.

Proposition 1.5.5 (Lyapunov's Inequality) *Let $0 < m < n$ then:*

$$(\mathbf{E}[|X|^m])^{\frac{1}{m}} \leq (\mathbf{E}[|X|^n]^{\frac{1}{n}} \tag{1.5. 37}$$

Proof: This follows by applying Jensen's inequality to the convex downward function : $g(x) = |x|^r$ where $r = \frac{n}{m} > 1$.

A very simple consequence is of Lyapunov's inequality is the following relationship between the moments of r.v's (provided they are defined).

$$\mathbf{E}[|X|] \leq (\mathbf{E}[|X|^2])^{\frac{1}{2}} \leq .... \leq (\mathbf{E}[|X|^n])^{\frac{1}{n}}$$

Another consequence of Jensen's inequality is a generalization of the Cauchy-Schwarz inequality called Hölder's inequality which we state without proof.

Proposition 1.5.6 (Hölder's Inequality)
Let $1 < p < \infty$ *and* $1 < q < \infty$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$. *If* $\mathbf{E}[|X|^p] < \infty$ *and* $\mathbf{E}[|Y|^q] < \infty$. *Then* $\mathbf{E}|XY| < \infty$ *and*

$$\mathbf{E}[|XY|] \leq (\mathbf{E}[|X|^p])^{\frac{1}{p}} (\mathbf{E}[|Y|^q])^{\frac{1}{q}} \tag{1.5. 38}$$

Note a special case of Hölder's inequality with $p = q = 2$ corresponds to the Cauchy-Schwarz inequality. Another important inequality which can be shown using Hölder's inequality is Minkowski's inequality which relates the moments of sums of r.v's to sums of the moments of r.v's. Once again we state the result without proof.

Proposition 1.5.7 (Minkowski's inequality)
Let $1 \leq p < \infty$. *If* $\mathbf{E}[|X|^p]$ *and* $\mathbf{E}[|Y|^p]$ *are finite then* $\mathbf{E}[|X+Y|^p] < \infty$ *and*

$$(\mathbf{E}[|X+Y|^p])^{\frac{1}{p}} \leq (\mathbf{E}[|X|^p])^{\frac{1}{p}} + (\mathbf{E}[|Y|^p])^{\frac{1}{p}} \tag{1.5. 39}$$

Finally we conclude with a very useful inequality called the Association Inequality whose generalization to $\Re^n$ valued r.v.'s is called the FKG inequality.

Proposition 1.5.8 *(***Association Inequality)**
Let $f(.)$ *and* $g(.)$ *be non-decreasing (or on-increasing) functions from* $\Re \to \Re$. *The for any real valued r.v.* X

$$\mathbf{E}[f(X)g(X)] \geq \mathbf{E}[f(X)]\mathbf{E}[g(X)] \tag{1.5. 40}$$

and if $f(.)$ *is non-increasing and* $g(.)$ *is non-decreasing then:*

$$\mathbf{E}[f(X)g(X)] \leq \mathbf{E}[f(X)]\mathbf{E}[g(X)] \tag{1.5. 41}$$

Proof: Since the proof involves a trick, here are the details. Let X and Y be independent and identically distributed: then if $f(.)$ and $g(.)$ are non-decreasing:

$$(f(x) - f(y))\,(g(x) - g(y)) \geq 0 \ \forall\ x, y \in \Re$$

Hence substitute X and Y and take expectations and the result follows.

The second result follows in a similar way.

Remark: The FKG inequality states that if $f(.)$ and $g(.)$ are non-decreasing functions from $\Re^n \to \Re$ then the above result holds for $\Re^n$ valued r.v's X. Note here a function is non-decreasing if it is true for each coordinate.

We conclude this section with a brief discussion of the so-called Cramer's theorem (also called the Chernoff bound) which allows us to obtain finer bounds on probabilities than those provided by Chebychev's inequality particularly when the tails of the distributions are rapidly decreasing. This can be done if the moment generating function of a r.v. is defined and thus imposes stronger assumptions on the existence of moments. This result is the basis of the so-called *large deviations theory* which plays an important role in calculating tail distributions when the probabilities are small and important in information theory and in simulation.

Let X be a real valued r.v. such that the moment generating function $M(h)$ is defined for $h < \infty$. i.e.

$$M(h) = \mathbf{E}[e^{hX}] = \int e^{hx} dF(x) < \infty$$

Define $\phi_F(h) = log(M(h))$. Then by the definition of $\phi_F(h)$ we observe $\phi_F(0) = 0$ and $\phi'_F(0) = m = \mathbf{E}[X]$. Define the following function:

$$I(a) = \sup_h (ha - \phi_F(h)) \tag{1.5. 42}$$

Then $I(m) = 0$ and $I(a) > 0$ for all $a \neq m$. Let us show this: by definition of I(a) we see that the max of $ha - \phi_F(h)$ is achieved at $a - \phi'_F(h) = 0$ and thus if $a = m$ then $h = 0$ from above and hence $I(m) = -\phi_F(0) = 0$. On the other hand : note $\phi'_F(h) = \frac{M'(h)}{M(h)}$. Let h_a the point where $max_h(ah - \phi_F(h)$ is achieved.

Then h_a satisfies $\frac{M'(h_a)}{M(h_a)} = a$ and $I(a) = ah_a - \phi_F(h_a)$. By the definition of I(a) as the envelope of affine transformations of a it implies that I(a) is convex downward and note that $I'(a) = ah'_a + h_a - \frac{M'(h_a)}{M(h_a)}h'a$ and by definition of h_a we have that $I'(a) = h_a$. Hence $I'(m) = 0$ implying that the function $I(a)$ achieves its minimum at $a = m$ which we have shown to be 0. Hence, $I(a) > 0$ for $a \neq m$. Also note that if $a > m = E[X]$ then $h_a > 0$ or

$$I(a) = \sup_{h \geq 0}\{ha - \phi_F(h)\}$$

Proposition 1.5.9 *Let X be a r.v. whose moment generating function , M(h), is defined for h $< \infty$. Then :*

i) If $a > m$, $\mathbb{P}(X \geq a) \leq e^{-I(a)}$

ii If $a < m$, $\mathbb{P}(X \leq a) \leq e^{-I(a)}$

The function $I(a)$ is called the rate function associated with the distribution $F(.)$.

Proof: We show only part i) the proof of the second result follows analogously.

First note that :

$$\mathbf{1}_{\{X \geq a\}} \leq e^{u(X-a)} \ ; \ \ \forall\ u \geq 0$$

Hence,

$$\mathbb{P}(X \geq a) = \mathbf{E}[\mathbf{1}_{[X \geq a]} \leq \mathbf{E}[e^{u(X-a)} = e^{-(ua - \phi_F(u))}$$

Since the inequality on the rhs is true for all $u \geq 0$ we choose $u \geq 0$ to minimize the rhs which amounts to choosing $u \geq 0$ to maximize the exponent $(ua - \phi_F(u))$ which by definition is $I(a)$ since the maximum is achieved in the region $u \geq 0$.

Now if $\{X_i\}_{i=1}^n$ is a collection of independent, identically distributed (**i.i.d**) r.v's with common mean m then the above result reads as :

Corollary 1.5.1 *If $\{X_i\}_{i=1}^n$ are i.i.d. with common mean m and let $M(h)$ be the moment generating function of X_1 which is defined for all $h < \infty$. Then:*

i) If $a > m$, then $\mathbb{P}(\frac{1}{n}\sum_{k=1}^n X_k \geq a) \leq e^{-nI(a)}$

ii) If $a < m$, then $\mathbb{P}(\frac{1}{n}\sum_{k=1}^n X_k \leq a) \leq e^{-nI(a)}$

where $I(a)$ corresponds to the rate function of $F(.)$ the distribution of X_1.

Proof: The proof follows from above by noting that by the i.i.d. assumption

$$M_n(h) = \mathbf{E}[e^{h\sum_{k=1}^n X_k}] = (M(h))^n$$

Remark: The result also has important implications in terms of the so-called *quick simulation* techniques or *importance sampling.* Below we discuss the idea. Suppose we define a new probability distribution

$$dF_a(x) = \frac{e^{h_a x}}{M(h_a)} dF(x)$$

i.e. we 'modulate' the distribution of X with the term $\frac{e^{h_a x}}{M(h_a)}$ then $F_a(x)$ defines a probability distribution (it is easy to check that $\int dF_a(x) = 1$) such that the mean of the r.v. X under F_a i.e. $\int x dF_a(x) = a$ and $var_a(X) = \frac{M''(h_a)}{M(h_a)} - a^2$. Now the implication in terms of 'quick simulations' is the following. Suppose a is large (or very small) such that $\mathbb{P}(X \geq a)$ is very small (or $\mathbb{P}(X \leq a)$ is very small). Then the occurrence of such events will be 'rare' and inspite of many trials we might not 'observe' them. However if we perform the experiment under the change of coordinates as the modulation suggests then since a is the mean such events will likely occur frequently and thus we are more likely to observe them.

We can also use F_a to give a proof that $I(a)$ is convex downward. To see this note that $I''(a) = h'_a$. Now since h_a satisfies $\frac{M'(h_a)}{M(h_a)} = a$ differentiating gives $h'_a(\frac{M''(h_a)}{M(h_a)} - \left(\frac{M'(h_a)}{M(h_a)}\right)^2) = 1$ and noting that the term in the brackets is just $var_a(X)$ of the r.v. X with distribution $F_a(x)$ which must be positive it implies that $h'_a > 0$ or $I(a)$ is convex downward and hence attains its minimum at $a = m$ which has the value 0.

1.6 Borel-Cantelli Lemmas

We conclude our overview of probability by discussing the so-called *Borel-Cantelli Lemmas.* These results are crucial in establishing the *almost sure* properties of events and in particular in studying the almost sure behavior of limits of events. The utility of these results will be seen in the context of convergence of sequences of r.v's and in establishing limiting behavior of Markov chains.

Before we state and prove the results we need to study the notions of limits of sequences of events. We have already seen this before in the context of sequential continuity where the limit of an increasing sequence of events is just the countable union while the limit of a decreasing sequence of sets is just the countable intersection of the events. That the events form a σ-algebra implies that the limit is a valid event to which a probability can be defined. Now if $\{\mathcal{A}_n\}_{n=1}^\infty$ is an arbitrary collection of events then clearly a limit need not be defined. However, just as in the case of real sequences the $\limsup$ or limit superior and the $\liminf$ or limit inferior can always be defined.

Definition 1.6.1 *Let $\{\mathcal{A}_n\}_{n=1}^\infty$ be a countable collection of events. Then $\limsup_n \mathcal{A}_n$ is the event : $\{\omega : \omega \in \text{infinitely many } \mathcal{A}_n\}$ and is given by:*

$$\limsup_n \mathcal{A}_n = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty \mathcal{A}_k \tag{1.6. 43}$$

The event $\limsup_n \mathcal{A}_n$ is often referred to as the event $\{\mathcal{A}_n \ i.o.\}$ where i.o. denotes *infinitely often.*

Definition 1.6.2 *Let* $\{\mathcal{A}_n\}_{n=1}^{\infty}$ *be a countable collection of events. Then* $\liminf_n \mathcal{A}_n$ *is the event :* $\{\omega : \omega \in \text{ all but a finite number of } \mathcal{A}_n\}$ *and given by:*

$$\liminf_n \mathcal{A}_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \mathcal{A}_n \tag{1.6. 44}$$

Proposition 1.6.1 (Borel-Cantelli Lemma) *Let* $\{\mathcal{A}_n\}$ *be a countable collection of events. Then :*

a) If $\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{A}_n) < \infty$ *then* $\mathbb{P}(A_n \ i.o.) = 0$.

b) If $\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{A}_n) = \infty$ *and* $\mathcal{A}_1, \mathcal{A}_2, ...$ *are independent then* $\mathbb{P}(\mathcal{A}_n \ i.o.) = 1$.

Proof:

a) First note by definition of $\{\mathcal{A}_n \ i.o.\} = \lim_{n\to\infty} B_n$ where $B_n = \bigcup_{k=n}^{\infty} \mathcal{A}_k$ and B_n is a monotone decreasing sequence. Hence by the sequential continuity property:

$$\begin{aligned} \mathbb{P}(\mathcal{A}_n \ i.o.) &= \lim_{n\to\infty} \mathbb{P}(B_n) \\ &= \lim_{n\to\infty} \mathbb{P}(\bigcup_{k=n}^{\infty} \mathcal{A}_k) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(\mathcal{A}_k) = 0 \end{aligned}$$

b) Note the independence of $\{\mathcal{A}_n\}$ it implies that $\{\bar{\mathcal{A}}_n\}$ are also independent. Hence we have:

$$\mathbb{P}(\bigcap_{k=n}^{\infty} \bar{\mathcal{A}}_k) = \prod_{k=n}^{\infty} \mathbb{P}(\bar{\mathcal{A}}_k)$$

Now using the property that $log(1-x) \leq -x$ for $x \in [0,1)$ we have:

$$log(\prod_{k=n}^{\infty} \mathbb{P}(\bar{\mathcal{A}}_k) = -\sum_{k=n}^{\infty} log(1 - \mathbb{P}(\mathcal{A}_k)) = -\sum_{k=n}^{\infty} \mathbb{P}(\mathcal{A}_k) = -\infty$$

Hence it implies that $\mathbb{P}(\bigcap_{k=n}^{\infty} \bar{\mathcal{A}}_k) = 0$ or $\mathbb{P}(\mathcal{A}_n \ i.o) = 1$.

Concluding remarks

In this chapter we have had a very quick overview of probability and random variables. These results will form the basis on which we will build to advance our study of stochastic processes in the subsequent chapters. The results presented in this chapter are just vignettes of the deep and important theory of probability. A list of reference texts is provided in the bibliography which the reader may consult for a more comprehensive presentation of the results in this chapter.

Exercises

1. Let X be a continuous r.v. Use the sequential continuity of probability to show that $\mathbb{P}(X = x) = 0$. Show the reverse i.e. if the probability that $X = x$ is zero then the r.v. is continuous.

2. Let $A\Delta B$ be the symmetric difference between two sets. i.e. $A\Delta B = A\bigcap \bar{B}\bigcup(B\bigcap \bar{A}$. Define $d_1(A,B) = \mathbb{P}(A\Delta B)$ and $d_2(A,B) = \frac{\mathbb{P}(A\Delta B)}{\mathbb{P}(A\bigcup B)}$ if $\mathbb{P}(A\bigcup B) \neq 0$ and 0 otherwise. Show that $d_1(.,.)$ and $d_2(.,.)$ define 'metrics' i.e. they are non-negative, equal to 0 if the events are equal (modulo null events) and satisfy the triangle inequality.

3. a) If $X(\omega) \geq 0$, show that:
$$\mathbf{E}[X(\omega)] = \int_0^\infty (1 - F(x))dx$$
b) If $-\infty < X(\omega) < \infty$ then:
$$\mathbf{E}[X(\omega)] = \int_0^\infty (1 - F(x))dx - \int_{-\infty}^0 F(x)dx$$
c) More generally, if X is a non-negative random variable show that
$$\mathbf{E}[X^r] = \sum_0^\infty rx^{r-1}P(X > x)dx$$

4. Let X be a continuous r.v. with density function $p_X(x) = C(x - x^2) \;\; x \in [a,b]$.
   (a) What are the possible values of a and b ?
   (b) What is C?

5. Suppose the X is a non-negative, real valued r.v.. Show that:
$$\sum_{m=1}^\infty P(X \geq m) \leq E[X] \leq 1 + \sum_{m=1}^\infty P(X \geq m)$$

6. Let X and Y be two jointly distributed r.v's with joint density:
$$p_{XY}(x,y) = \frac{1}{x}, \quad 0 \leq y \leq x \leq 1$$
Find
   (a) $p_X(x)$ the marginal density of X
   (b) Show that , given $\{X = x\}$, the r.v. Y is uniformly distributed on $[0,x]$ i.e.the conditional density:
$$pY/X(y/x) = \frac{1}{x} \quad y \in [0,x]$$
   (c) Find $E[Y|X = x]$

(d) Calculate $\mathbf{P}(X^2 + Y^2 < 1|X = x)$ and hence show that $\mathbf{P}(X^2 + Y^2 \leq 1) = \ln(1 + \sqrt{2})$

7. Find the density function of $Z = X + Y$ ehen X and Y have the joint density:

$$p_{X,Y}(x, y) = \frac{(x + y)}{2} e^{-(x+y)} \quad , \quad x, y \ \geq 0$$

8. Let X and Y be independent, exponentially distriibuted r.v's with parameters λ and μ respectively. Let $U = \min\{X, Y\}$ and $V = \max\{X, Y\}$ and $W = V - U$.

   (a) Find $P(U = X) = P(X \leq Y)$

   (b) Show that U and W are independent.

   (c) Show that $P(X = Y) = 0$

   (d) Suppose that $\lambda = \mu$. Show that the conditional distribution of Y given that $X + Y = u$ is uniform in $[0, u]$. Hence, find $\mathbf{E}[X|X + Y]$ and $\mathbf{E}[Y|X + Y]$.

9. Let X and Y be independent and identically distributed r.v.'s with $\mathbf{E}[X] < \infty$. Then show that

$$\mathbf{E}[X/X + Y] = \mathbf{E}[Y/X + Y] = \frac{X + Y}{2}$$

   Now if $\{X_i\}$ is a sequence if i.i.d. r.v's. using the first part establish that :

$$\mathbf{E}[X_1/S_n] = \frac{S_n}{n}$$

   where $S_n = \sum_{i=1}^n X_i$

10. Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (i.i.d) random variables for which $E[X_1^{-1}]$ exists. Let $S_n = \sum_{i=1}^n X_i$. Show that for $m < n$

$$E[\frac{S_m}{S_n}] = \frac{m}{n}$$

   Find $E[S_m|S_n]$ and $E[S_n|S_m]$ assuming that $E[X_i] = 1$.

11. Let X and Y be independent r.v's with the following distributions:

   i) X is $N(0, \sigma^2)$.

   ii) Y takes the value 1 with prob. p and -1 with prob q=1-p

   Define $Z = X + Y$ and $W = XY$. Then:

   i) Find the probability density of Z and its characteristic function.

   ii) Find the conditional density of Z given Y.

   iii) Show that W is Gaussian. Define $U = W + X$. Show that U is not Gaussian. Find $\mathbf{E}[W]$ and $var(W)$. Are W and X correlated ? Independent ?

   This problem shows the importance of the condition of 'jointly Gaussian' in connection with linear combinations of Gaussian r.v's.

12. Let $X \sim N(0,1)$ and let $a > 0$. Show that the r.v. Y defined by:

$$\begin{aligned} Y &= X \ \ if \ \ |X| < a \\ &= -X \ \ if \ \ |X \geq a \end{aligned}$$

has a $N(0,1)$ distribution. Find an expression for $\rho(a) = cov(X,Y)$ in terms of the normal density function. Are the pair (X,Y) jointly Gaussian?

13. Let X and θ be independent r.v's with the following distributions:

$$\begin{aligned} p_X(x) &= 0 \ \ x < 0 \\ &= xe^{-\frac{x^2}{2}} \ \ x \geq 0 \end{aligned}$$

θ is uniformly distributed in $[0, 2\pi]$.

Define

$$Z_t = X\cos(2\pi t + \theta)$$

Then show that for each t, Z_t is Gaussian.

14. Let X and Y be two r.v's with joint density given by:

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

Let $Z = \sqrt{X^2+Y^2}$.

Find $\mathbf{E}[X|Z]$.

15. Show that a sequence of *Markovian* events have the Markov property in the reverse direction i.e. $\mathbb{P}(A_n/A_{n+1}, A_{n+2}, ..) = \mathbb{P}(A_n/A_{n+1})$ and hence or otherwise establish the conditional independence property $\mathbb{P}(A_{n+1} \bigcap A_{n-1}/A_n) = \mathbb{P}(A_{n+1}/A_n)\mathbb{P}(A_{n-1}/A_n)$.

16. Let X be a Poisson r.v. i.e. X takes integer values $0, 1, 2, ...$ with $\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$; $\lambda > 0$. Compute the mean and variance of X. Find the characteristic function of X. Let $\{X_i\}$ be independent Poisson r.v's. with parameters $\{\lambda_i\}$. Then show that $Y = \sum_{i=1}^n X_i$ is a Poisson r.v.

17. Let $\{X_i\}$ be i.i.d. r.v's. Find the distribution of $Y_n = \max\{X_1, X_2, ..., X_n\}$ and $Z_n = \min\{X_1, X_2, ..., X_n\}$ in terms of the distribution function $F(x)$ of X_1.

    Now define $\psi_n = Y_n - Z_n$. The r.v. ψ_n is called the range of $\{X_i\}_{i=1}^n$. Show that the distribution $F_\psi(z) = \mathbb{P}(\psi_n \leq z)$ is given by:

$$F_\psi(z) = n\int_{-\infty}^{\infty} \left(F(x+z) - F(x)\right)^{n-1} dF(x)$$

18. Let X and Y be independent r.v's such that $Z = X + Y$ is Gaussian. Show that X and Y are Gaussian r.v's. You may assume that they are identically distributed.

    This is a part converse to the statement that sums of jointly Gaussian r.v's are Gaussian. This shows that if a r.v. is Gaussian then we can always find two (or even n) independent Gaussian r.v's such that the given r.v. is the sum of the two r.v's. (or n r.v's). Similarly for the case of the Poisson r.v's . This is related to an important notion of *infinitely divisible* properties of such distributions. In fact it can be shown that if a distribution is infinitely divisible then its characteristic function is a product of a Gaussian characteristic function and a characteristic function of the Poisson type.

19. Let $X(\omega)$ and $Y(\omega)$ be jointly distributed, non-negative random variables. Show that:

$$\mathbf{P}(X+Y>z)=\mathbf{P}(X>z)+\mathbf{P}(X+Y>z\geq X)$$

and

$$\int_0^\infty \mathbf{P}(X+Y>z\geq X)dz=E[Y]$$

20. Let X and Y be two jointly distributed r.v's with joint density:

$$p_{XY}(x,y)=\frac{1}{x}, \qquad 0\leq y\leq x\leq 1$$

Find

(a) $p_X(x)$ the marginal density of X

(b) Show that , given $\{X=x\}$, the r.v. Y is uniformly distributed on $[0,x]$ i.e.the conditional density:

$$p_{Y|X}(y|x)=\frac{1}{x} \quad y\in[0,x]$$

(c) Find $E[Y|X=x]$

(d) Calculate $\mathbf{P}(X^2+Y^2<1|X=x)$ and hence show that $\mathbf{P}(X^2+Y^2\leq 1)=\ln(1+\sqrt{2})$

21. Suppose that X and Y are jointly Gaussian r.v's with $E[X]=E[Y]=m$ and $E[(X-m)^2]=E[(Y-m)^2]=\sigma^2$. Suppose the correlation coefficient is ρ. Show that $W=X-Y$ is independent of $V=X+Y$.

Define $M=\frac{X+Y}{2}$ and $\Sigma^2=(X-M)^2+(Y-M)^2$. Show that M and Σ^2 are independent. M and Σ^2 are the sample mean and variance.

Note: $(X-M)^2$ and $(Y-M)^2$ are not Gaussian.

This result can be extended to the sample mean and variance of any collection $\{X_i\}_{i=1}^N$ of Gaussian r.v's with common mean and variance where the sample mean $M_N=\frac{1}{N}\sum_{k=1}^N X_k$ and $\sigma_N^2=\frac{1}{N-1}\sum_{k=1}^N(X_k-M_N)^2$.

22. Let X and Y be jointly distributed 0 mean random variables with finite variance. Show that:

$$var(Y)=\mathbf{E}[var(Y|X)]+var(\mathbf{E}[Y|X])$$

23. Let X be an integrable r.v.. Let $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$. Show that:

$$\mathbf{E}[X|X^+] = X^+ - \frac{\mathbf{E}[X^-]}{\mathbb{P}(X^+ = 0)} \mathbb{I}_{[X^+=0]}$$

and

$$\mathbf{E}[X|X^-] = -X^- + \frac{\mathbf{E}[X^+]}{\mathbb{P}(X^- = 0)} \mathbb{I}_{[X^-=0]}$$

24. a) Show that a random variable is standard normal if and only if

$$\mathbf{E}[f'(X) - Xf(X)] = 0$$

for all continuous and piecewise differentiable functions $f(.) : \Re \to \Re$ with $E[f'(Z)] < \infty$ for $Z \sim N(0, 1)$.

b) Let $h : \Re \to \Re$ be a piecewise continuous function. Show that there exists a function f which satisfies the so-called Stein equation:

$$h(x) - \Phi h = f'(x) - xf(x)$$

where $\Phi h = E_N[h]$ where $E_N[.]$ denotes the expectation w.r.t. the standard normal distribution.

This problem is related to defining a metric in the space of probability distributions. It allows us to "measure" how "different" a given distribution is from a normal distribution by choosing $h(x) = \mathbf{1}_{[X \leq x]}$.

25. Show that Jensen's inequality holds for conditional expectations.

26. Let X be $N(m, \sigma^2)$. Show that the rate function associated with X is given by:

$$I(a) = \frac{(a - m)^2}{2\sigma^2}$$

Using this result compute the tail distributions $\mathbb{P}(|X - m| \geq k\sigma)$ for $k = 1, 2, 3, 5, 10$ and compare the results with the corresponding results using Chebychev's inequality and the exact result.

Let $S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ where $\{X_i\}$ are i.i.d. $N(0, 1)$ r.v's. Find $\mathbb{P}(|S_n| > \varepsilon))$ and find the smallest n (in terms of ε) such that the required probability is less than 10^{-3}.

Bibliography

[1] P. Brémaud; *An introduction to probabilistic modeling*, Undergraduate Texts in Mathematics, Springer-Verlag, N.Y., 1987

[2] W. Feller; *An introduction to probability theory and its applications Vol. 1*, 2nd. edition, J. Wiley and Sons, N.Y.,1961.

[3] R. E. Mortensen; *Random signals and systems*, J. Wiley, N.Y. 1987

[4] A. Papoulis; *Probability, random variables and stochastic processes*, McGraw Hill, 1965

[5] B. Picinbono; *Random signals and systems*, Prentice-Hall, Englewood Cliffs, 1993

[6] E. Wong; *Introduction to random processes*, Springer-Verlag, N.Y., 1983

At an advanced level the following two books are noteworthy from the point of view of engineering applications.

[1] E. Wong and B. Hajek; *Stochastic processes in engineering systems*, 2nd. Edition,

[2] A. N. Shiryayev; *Probability*, Graduate Texts in Mathematics, Springer-Verlag, N.Y. 1984