Ch. 2 Probability Theory

(October 26, 2018)

The concept of chance and uncertainty are as old as civilization itself. People have always had to cope with uncertainty about the weather, their food supply, and other aspects of their environment, and have strives to **reduce this uncertainty** and their effects.

In this chapter we discuss general probability models. These models are used to describe events that occur "by chances" or "randomly". Most of us have some idea of what probability is. Surprisingly enough, however, it is difficult to define probability in a way that applies to every situation where we use the term and in a way which is agreeable to most people.

1 Descriptive Study of Data

1.1 Histograms and Their Numerical Characteristics

By descriptive study of data we refer to the summarization and exposition (tabulation, grouping, graphical representation) of *observed* data as well as the derivation of numerical characteristics such as measures of location, dispersion and shape. Although the descriptive study of data is an important facet of modeling with real data itself, in the present study it is mainly used to motivate the need for probability theory and statistical inference proper.

In order to make the discussion more specific let us consider the after-tax personal income data of 12000 household for 1999-2000 in Taiwan. There data in row form con-

stitute 12000 numbers between $5000 and $100000. This presents us with a formidable task in attempting to understand how income is distributed among the 12000 households represented in the data. The purpose of descriptive statistics is to help us make some sense of such data. A natural way to proceed is to summarize the data by allocating the numbers into classes (intervals). The number of intervals is chosen a priori and it depends on the degree of summarization needed. Then we have the "Table of the personal income in Taiwan". The first column of the table shows the income intervals, the second column the second column shows the number of income falling into each interval and the third column the relative frequency for each interval. The relative frequency is calculated by dividing the number of observations in each interval by the total number of observations. The fourth column is the cumulative frequency. Summarizing the data in this Table enables us to get some idea of how income is distributed among various classes. If we plot the relative (cumulative) frequencies in a bar graph we get what is known as the *histogram* (cumulative). That is, a histogram is a graphical representation of the distribution of numerical data.

For further information on the distribution of income we could calculate various numerical characteristics describing the histogram's location, dispersion and shape. Such measures can be calculate directly in terms of the raw data. However, in the present case it is more convenient for expositional purpose to use the grouped data. The main reason for this is to introduce various concept which will be reinterpreted in the context of probability.

The *mean* as measure of *location* takes the form

$$\bar{z} = \sum_{i=1}^{n} \phi_i z_i,$$

where ϕ_i and z_i refer to the relatively frequency and the midpoint of interval i. The *mode* as a measure of location refers to the value of income that occurs most frequently in the data set. Another measure of location is the *median* referring to the value of income in the middle when income are arranged in an ascending order according to the size of income. The best way to calculate the median is to plot the *cumulative frequency graph.*

Another important feature of the histogram is the dispersion of the relative frequency around a measure of central tendency. The most frequently used measure of dispersion is the *variance* defined by

$$v^2 = \sum_{i=1}^{n} (z_i - \bar{z})^2 \phi_i,$$

which is a measure of dispersion around the mean; v is known as the *standard deviation*.

We can extend the concept of the variance to

$$m_k = \sum_{i=1}^{n} (z_i - \bar{z})^k \phi_i, \quad k = 3, 4, ...$$

defining what are known as *higher central moments*. These higher moments can be used to get a better idea of the shape of the histogram. For example, the standardized form of the third and fourth moments defined by

$$SK = \frac{m_3}{v^3} \quad and \quad K = \frac{m_4}{v^4},$$

known as the *skewness* and *kurtosis coefficients*, measure the asymmetry and the peakedness of the histogram, respectively. In the case of a symmetric histogram, $SK = 0$ and the less peaked the histogram the greater value of K.

1.2 Looking Ahead

The most important drawback of descriptive statistics is that the study of the observed data enables us to draw certain conclusions which relate *only to the data in hand*. The temptation in analyzing the above income data is to attempt to make generalizations beyond the data in hand, in particular about the distribution of income in Taiwan (not just 12000 households in Taiwan). This, however, is not possible in the descriptive statistics framework. In order to be able to generalize beyond the data in hand we need *to "model"* the distribution of income in the Taiwan and *not just describe* the observed data in hand. Such a general model is provided by probability theory to be considered in Section 2.

It turns out that the model provided by probability theory owns a lot to the earlier developed descriptive statistics. In particular, most of the concepts which form the basis of the probability theory were motivated by the descriptive statistic concept considered above. The concepts of measures of location, dispersion and shape, as well as the frequency curve, were transplanted into probability theory with renewed interpretations. The frequency curve when reinterpreted becomes a density function purporting to model observable real world phenomena, As for the various measures, they will now be reinterpreted in terms of the density function.

2 Probability

Why we need the probability theory in analyzing observed data ? In the descriptive study of data considered in the last section, it is emphasized that the results cannot be generalized outside the observed data under consideration. Any question relating to the population from which the observed data were drawn cannot be answered within the descriptive statistics framework. In order to be able to do that we need the theoretical framework offered by probability theory. In fact, probability theory develops a *mathematical model* which provides the logical foundation of statistical inference procedures for analyzing observed data.

In developing a mathematical model we must first identify the important features, relations and entities in the real world phenomena and then devise the concepts and choose the assumptions with which to project a generalized description of these phenomena; an idealized pictures of these phenomena. *The model as a consistent mathematical system has "a life of its own" and can be analyzed and studied without direct reference to real world phenomena.* (Thinks of analyzing the population, we do not have to refer to the information in the sample.)

2.1 Interpretations of Probability

Despite the fact that the concept of probability is such a common and natural part of our experience, no single scientific interpretation of the term "probability" is accepted by all statisticians. Indeed the true meaning of probability is still a highly controversial subject and is involved in many current philosophical discussion pertaining to the foundations of statistics. Each of these interpretations can be very useful in applying probability theory to practical problems.

2.1.1 The Classical Approach

In 1850s Laplace proposed what is known today as the *classical* definition of probability:

Definition.
If a random experiment can result in N mutually exclusive and equally likely outcomes and if N_A of these outcomes result in the occurrences of the event A, then the *probability of* A is defined by

$$P(A) = \frac{N_A}{N}. \qquad \blacksquare$$

The obvious limitations of the classical approach are:

(a). It is applicable to situations where there is only a $finite$ number of possible outcomes; and

(b). The "equally likely" condition renders the definition *circular*. □

Some important random experiments may give rise to s set of infinite outcomes. And the idea of "equally likely" is synonymous with "equally probable", thus probability is defined using the idea of probability !

2.1.2 The Frequency Approach

The most influential of the approaches suggested in an attempt to tackle the problems posed by the classical approach are so-called $frequency$ and *subjective* approaches to probability. The frequency approach had its origins in the writing of Poisson but it was not until the late 1920 that Von Mises put forward a systematic account of the approach. The basic argument of the frequency approach is that probability does not have to be restricted to situations of apparent symmetry (equally likely) since the notation of probability should be interpreted as stemming from the observable stability of empirical frequencies. For example, in the case of a fair coin we say that the probability of $A = \{H\}$ is $\frac{1}{2}$, not because there are two equally likely outcomes but because repeated series of large numbers of trials demonstrate that the empirical frequency of occurrence of A "converges" to the limit $\frac{1}{2}$ as the number of trials goes to infinity.

Definition.

Denote n_A be the number of occurrences of an event A in n trials, then if

$$\lim_{n\to\infty}\left(\frac{n_A}{n}\right) = P_A,$$

we say that $P(A) = P_A$. ■

2.1.3 The Subjective Approach

Despite the fact that the frequency approach seems to be an important improvement over the classical approach, there are some obvious objections to it. "How can we generate infinite sequences of trials ?" and "What happens to phenomena where repeated trials are not possible ?"

The subject approach to probability renders the notation of probability of a subject status by regarding it as "degree of belief" on behalf of individuals assessing the uncertainty of a particular situation. For example, if a weatherman says that the probability of rain the next day is 0.4, he is telling the audiences his subjective belief in the likelihood of rain the next day, given the weather conditions today. Different weather forecasters would give different probabilities from the same present conditions, which indicates the subjective nature of these probabilities.

2.2 The Axiomatic Approach

As was explained above, there is controversy in regard to the proper meaning and interpretation of some of the probabilities that are assigned to the outcomes of many experiments. However, once probabilities have been assigned to some simple outcomes in an experiment, there is complete agreement among all authorities that the mathematical theory of probability provided the appropriate methodology for further study of these probabilities. Almost all work in the mathematical theory have been related to the following two problems:

(a). methods for determining the probabilities of certain events from the specified probabilities of each possible outcome of an experiment and;

(b). methods for revising the probabilities of events when additional information is obtained.

By the 1920s there was a wealth of results and probability began to grow into a systematic body of knowledge. Although various people attempted a systematization of probability it was the work of the Russian mathematician Kolmogorov which provided to be the cornerstone for a systematic approach to probability theory. Kolmogorov managed to relate the concept of the probability to that of a measure in integration theory and exploited to the full analogies between set theory and the theory of functions on the one hand and the concept of a random variable on the other. In a monumental monograph in 1933 he proposed an axiomatization of probability theory establishing it once and for all as part of mathematical property. There is no doubt that this monograph provided to be the watershed for the later development of probability theory growing enormously in importance and applicability.

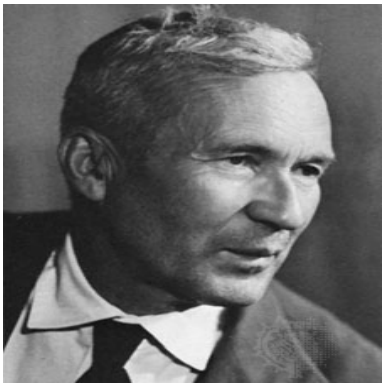

Figure (2-1).
Kolmogorov, A. (1908-1987)

The axiomatic approach to probability proceeds from a set of axioms (accepted without questioning as obvious), which are based on many centuries of human experience, and the subsequent development is built *deductively* using formal logical arguments, like any other part of mathematics such as geometry or linear algebra. In mathematics an axiomatic system is required to be *complete*, $non-redundant$ and *consistent*. By complete we mean that the set of axioms postulated should enable us to prove every other theorem in the theory in question using the axioms and mathematical logic. The notion of non-redundancy refers to the impossibility of deriving any

axiom of the system from the other axioms. Consistency refers to the non- contradictory nature of the axioms.

2.2.1 Random Experiment, Sample Space and Events

A probability model is by construction intended to be a description of a *chance mechanism* giving rise to observed data. The starting point of such a model is provided by the concept of a *random experiment* describing a simplistic and idealized process giving rise to observed data.

Definition. (Random Experiment)
A *random experiment*, denoted by $\mathcal{E}$, is an experiment which satisfies the following conditions:
(a). *all possible distinct outcomes* are known a priori;
(b). In any particular trial the *outcomes* is *not known a priori*; and
(c). It can be repeated under identical conditions. ■

The axiomatic approach to probability theory can be viewed as a formalization of the concept of a random experiment. In an attempt to formalize condition (a) (of the definition of random experiment) all possible distinct outcomes are known a priori, Kolmogorov devised the set $\mathcal{S}$ which included "all possible distinct outcome" and has to be postulated before the experiment is performed.

Definition. (Sample Space)
The *sample space*, denoted by $\mathcal{S}$, is defined to be the set of all possible outcomes of the experiment $\mathcal{E}$. The elements of $\mathcal{S}$, denoted as s, are called *elementary events*. That is, $s \in \mathcal{S}$. ■

Example.
Consider the random experiment $\mathcal{E}$ of tossing a fair coin twice and observing the faces

turning up. The sample space of $\mathcal{E}$ is

$$\mathcal{S} = \{(HT), (TH), (HH), (TT)\},$$

with $(HT), (TH), (HH)$ and (TT) being the elementary events belonging to $\mathcal{S}$. ■

The second ingredient of $\mathcal{E}$ related to (b) and in particular to the various form events can take. A moment of reflection suggested that there is no particular reason why we should be interested in elementary outcomes only. We might be interested in such events as A_1–"at least one H", A_2–"at most one H", and these are not elementary events; in particular

$$A_1 = \{(HT), (TH), (HH)\}$$

and

$$A_2 = \{(HT), (TH), (TT)\}$$

are combinations of elementary events. All such outcome are called *events* associated with the same sample space $\mathcal{S}$ and they are defined by *combining elementary events.*

Definition. (Events)
Any designated collection of sample outcomes, including individual outcome (i.e. elements of $\mathcal{S}$), the entire sample space, and the null set, constitutes an *event.*[1] ■

Given that $\mathcal{S}$ is a set with members the elementary events, this takes us immediately into the realm of set theory and events can be formally defined to be subsets of $\mathcal{S}$ formed by set theoretic operation ("$\cap$"-intersection, "$\cup$"-union, "$-$"-complementation) on the elementary events. For example,

$$A_1 = \{(HT)\} \cup \{(TH)\} \cup \{(HH)\} = \{\overline{(TT)}\} \subset \mathcal{S},$$
$$A_2 = \{(HT)\} \cup \{(TH)\} \cup \{(TT)\} = \{\overline{(HH)}\} \subset \mathcal{S}.$$

[1]Understanding the concept of an event is crucial for the discussion which follows. Intuitively an event is any proposition associated with $\mathcal{E}$ which may occur or not at each trial. We say that event A_1 occurs when any one of the elementary events it comprises occurs. Thus, when a trial is made only one elementary event is observed but a large number of events may have occurred. For example, if the elementary event (HT) occurs in a particular trial, A_1 and A_2 have occurred as well.

Two special events are $\mathcal{S}$ itself, called the *sure events* and the *impossible event* $\varnothing$ defined to contain no elements of $\mathcal{S}$, i.e. $\varnothing = \{\ \}$; the latter is defined for completeness.

2.2.2 Set of Subsets of Sample space, σ-Field

A third ingredient of $\mathcal{E}$ also associated with (b) which Kolmogorov had to formalized was the idea of uncertainty related to the outcome of any particular trial of $\mathcal{E}$. This he formalized in the notion of probabilities attributed to the various events associated with $\mathcal{E}$, such as $P(A_1)$, $P(A_2)$, expressing the "likelihood" of occurrences of these events. Although attributing probabilities to the elementary events presents no particular mathematical problem, going the same for events in general is not as straightforward. The difficulty arises because if A_1 and A_2 are events, $\overline{A_1} = \mathcal{S} - A_1$, $\overline{A_2} = \mathcal{S} - A_2$, $A_1 \cap A_2$, $A_1 \cup A_2$, etc., are also events because the *occurrence or non-occurrence* of A_1 and A_2 implies the occurrence or not of these events. This implies that for the attribution of probabilities to make sense we have to impose some mathematical structures on the set of all events, say $\mathcal{F}$, which reflects the fact that whichever way we combine these events, the end result is always an event. The temptation at this stage is to define $\mathcal{F}$ to be the set of all subsets of $\mathcal{S}$, called the *power set*. Surely, this covers all possibilities ! In the above example, the power set of $\mathcal{S}$ take the form

$$\begin{aligned}\mathcal{F} \;=\; & \{S, \varnothing, \{(HT)\}, \{(TH)\}, \{(HH)\}, \{(TT)\}, \{(HT),(TH)\}, \\ & \{(HT),(HH)\}, \{(HT),(TT)\}, \{(TH),(HH)\}, \{(TH),(TT)\}, \\ & \{(HH),(TT)\}, \{(HT),(TH),(HH)\}, \{(HT),(TH),(TT)\}, \\ & \{(TH),(HH),(TT)\}, \{(HT),(HH),(TT)\}\}.\end{aligned}$$

Definition. (Power Set)
The set of all the subset of X is called the power set of X, denoted 2^X. The power set of a set with n elements has 2^n elements, which accounts for its name and representation.■

Sometimes we are not interested in all the subsets of $\mathcal{S}$, we need to define a set independently of the power set by endowing it with a mathematical structure which ensures that no inconsistency arises. This is achieved by requiring that $\mathcal{F}$ in the following has a special mathematical structure. It is a σ-field related to $\mathcal{S}$. The idea

behind the following definition is to specify subset of power set that are large enough to be interesting, but whose characteristics may be more tractable. We typically do this by *choosing a base collection* of sets with known properties, and then specifying certain operations for creating new sets from existing ones. These operations permit an interesting diversity of class members to be generated, but important properties of the sets may be deducted from those of the base collection.

Definition. (σ-field):
Let $\mathcal{F}$ be a set of subsets of $\mathcal{S}$. $\mathcal{F}$ is called a $\boldsymbol{\sigma}$-field if:

(a). if $A \in \mathcal{F}$, then $\overline{A} \in \mathcal{F}$–closure under complementation;

(b). if $A_i \in \mathcal{F},\ i = 1, 2, ...,$ then $(\cup_{i=1}^{\infty} A_i) \in \mathcal{F}$–closure under countable union.
Note that (a) and (b) taken together implying the following:

(c). $\mathcal{S} \in \mathcal{F}$, because $A \cup \overline{A} = \mathcal{S}$;

(d). $\varnothing \in \mathcal{F}$ (from (c) $\overline{\mathcal{S}} = \varnothing \in \mathcal{F}$); and

(e). $A_i \in \mathcal{F},\ i = 1, 2, ...,$ then $(\cap_{i=1}^{\infty} A_i) \in \mathcal{F}$.[2] ■

These suggest that a σ-field is a **set of subsets of** $\mathcal{S}$ which is closed under complementation, countable unions and intersections. That is, any of these operations on the elements of $\mathcal{F}$ will give rise to an element of $\mathcal{F}$.

Example.
If we are interested in events with one of each H or T, there is no point in defining the σ-field to be the power set, and $\mathcal{F}_c$ below can do as well with fewer events to attributed probabilities to. Starting from the events of interest, $C = \{(HT), (TH)\}$, we construct the minimal σ-field generated by its elements. This can be achieved by extending C to include all the events generated by set theoretic operations (unions, intersections, complementations) on C. The the minimal σ-field generated by C is

$$\mathcal{F}_c = \{\{(HT), (TH)\}, \{(HH), (TT)\}, \mathcal{S}, \varnothing\},$$

[2] By applying De Morgan's laws: $\overline{\cap_{i=1}^{\infty} A_i} = \cup_{i=1}^{\infty} \overline{A}_i \in \mathcal{F}$, therefore $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$.

and we denote it by $\mathcal{F}_c = \sigma(C)$. ■

Definition. (Measurable Space)
The pair $(\mathcal{S}, \mathcal{F})$ is called a *measurable space* when $\mathcal{F}$ is a σ-field of $\mathcal{S}$.[3]

Exercise 1.
Construct the minimal σ-field, $\mathcal{F}_D = \sigma(D)$, generated by the event $D = \{(HT)\}$.[4] ■

2.2.3 Set of Subsets of $\mathbb{R}^1$, Borel-Field

Let us turn our attention to the various collections of events (σ-fields) that are relevant for econometrics.

Definition. (Borel-Field on $\mathbb{R}$):
The Borel σ-field $\mathcal{B}$ is the smallest collection of sets (called the Borel sets) that includes
(a). all the closed half-lines of $\mathbb{R}$;
(b). the complements $\overline{B}$ of any B in $\mathcal{B}$;
(c). the union $\cup_{n=1}^{\infty} B_i$ of any sequences $\{B_i\}$ of sets in $\mathcal{B}$. ■

The Borel set of $\mathbb{R}$ just defined are said to be *generated* by the all the closed half-lines sets of $\mathbb{R}$. The same Borel sets would be generated by all the open half-lines of $\mathbb{R}$, all open sets of $\mathbb{R}$, all the open intervals of $\mathbb{R}$, or all the closed intervals of $\mathbb{R}$. The Borel sets are a "rich" collection of events for which probabilities can be defined. To see how the Borel set contains almost every conceivable subset of $\mathbb{R}$ from the closed half-lines, consider the following example.

[3] From Wikipedia: A measure on X is a function which assigns a real number to subsets of X; this can be thought of as making precise a notion of "size" or "volume" for sets. One might like to assign such a size to every subset of X, but the axiom of choice implies that when the size under consideration is standard length for subsets of the real line, then there exist sets known as *vitali* sets for which no size exists. For this reason, one considers instead a smaller collection of privileged subsets of X whose measure is defined; these sets constitute the σ-algebra. Elements of the σ-algebra are called measurable sets. An ordered pair (X, Σ), where X is a set and Σ is a σ-algebra over X, is called a measurable space.

[4] Answers: $\mathcal{F}_D = \{\{(HT)\}, \{(TH), (HH), (TT)\}, \mathcal{S}.\varnothing\}$

Example.
Let $\mathcal{S}$ be the real line $\mathbb{R} = \{x : -\infty < x < \infty\}$ and the set of events of interest be

$$J = \{B_x : x \in \mathbb{R}\},$$

where $B_x = \{z : z \leq x\} = (-\infty, x]$. How can we construct a σ-field, $\sigma(J)$ on $\mathbb{R}$ from the events B_x?

By definition $B_x \in \sigma(J)$, then
(a). Taking complements of B_x: $\bar{B}_x = \{z : \ z > x, \ z \in \mathbb{R}\} = (x, \infty) \in \sigma(J)$;
(b). Taking countable unions of B_x: $\cup_{n=1}^{\infty}(-\infty, x - (1/n)] = (-\infty, x) \in \sigma(J)$;
(c). Taking complements of (b): $(\overline{-\infty, x}) = [x, \infty) \in \sigma(J)$;
(d). From (a), for $y > x, [y, \infty) \in \sigma(J)$;
(e). From (d), $(\overline{-\infty, x] \cup [y, \infty)} = (x, y) \in \sigma(J)$;
(f). $\cap_{n=1}^{\infty}(x - (1/n), x] = \{x\} \in \sigma(J)$. ■

This shows not only that $\sigma(J)$ is a σ-field but it includes almost every conceivable subset of $\mathbb{R}$, that is, it coincides with the σ-field generated by any set of subsets of $\mathbb{R}$, which we denote by $\mathcal{B}$, i.e. $\sigma(J) = \mathcal{B}$, or the *Borel Field* on $\mathbb{R}$.

Example.
The Borel sigma-field $\mathcal{B}$ of subsets of $\mathbb{R}$ is defined as a smallest sigma-field over the system of all open subsets of $\mathbb{R}$. Prove that $\mathcal{B}$ is also a smallest sigma-field over the system of all half-closed intervals $(a, b]$ with $-\infty < a < b < \infty$.

Let $\mathcal{B}_1$ be the smallest sigma-field that contains all open intervals and let $\mathcal{B}_2$ be the smallest sigma-field that contains all half-closed intervals. Then $(a, b] = \{\cap(a, b+1/n) : n \in \mathbb{N}\} \in \mathcal{B}_2$ which is a countable intersection of open intervals. So $(a, b]$ is also in $\mathcal{B}_1$, as $\mathcal{B}_1$ contains all open intervals and is a sigma-field. Therefore $\mathcal{B}_2 \subset \mathcal{B}_1$.

On the other hand, if (a, b) is an open interval, then $(a, b) = \{\cup(a, b - 1/n] : n \in \mathbb{N}\} \in \mathcal{B}_1$ which is a countable union of half-closed intervals, so (a, b) is also in $\mathcal{B}_2$. Again we conclude that $\mathcal{B}_1 \subset \mathcal{B}_2$. We conclude that $\mathcal{B}_1 = \mathcal{B}_2$. ■

2.2.4 Probability Measure, Probability Space

Having solved the technical problem in attributing probabilities to events by postulating the existence of a σ- field $\mathcal{F}$ associated with the sample space $\mathcal{S}$, Kolmogorov went on to formalize the concept of probability itself.

Definition. (Probability Measure)
Let $(\mathcal{S}, \mathcal{F})$ be a measurable space. A mapping $\mathcal{P}(\cdot) : \mathcal{F} \to [0, 1]$ is a *probability measure* on $(\mathcal{S}, \mathcal{F})$ provided that

(a). $\mathcal{P}(A) \geq 0$ for $\forall A \in \mathcal{F}$.

(b). $\mathcal{P}(\mathcal{S}) = 1$; and

(c). For any disjointed sequence $\{A_i\}$ of sets in $\mathcal{F}$ (i.e., $A_i \cap A_j = \varnothing$ for all $i \neq j$), $\mathcal{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$.[5] ■

To summarize the arguments so far, Kolmogorov formalized the condition (a) and (b) of the random experiment $\mathcal{E}$ in the form of the trinity $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$ comprising the set of all outcomes $\mathcal{S}$–the sample space, a σ-field $\mathcal{F}$ of events related to $\mathcal{S}$ and a probability function $\mathcal{P}(\cdot)$ assigning probability to events in $\mathcal{F}$. For the coin example, if we choose $\mathcal{F}$ (The first is H and the second is T)$= \{\{(HT)\}, \{(TH), (HH), (TT)\}, \varnothing, \mathcal{S}\}$ to be the σ-field of interest, $\mathcal{P}(\cdot)$ is defined by

$$\mathcal{P}(\mathcal{S}) = 1, \quad \mathcal{P}(\varnothing) = 0, \quad \mathcal{P}(\{(HT)\}) = \frac{1}{4}, \quad \mathcal{P}(\{(TH), (HH), (TT)\}) = \frac{3}{4}.$$

Because of its importance the trinity $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$ is given a name.

[5]The first two axioms seem rather self-evident and are satisfied by both the classical as well as frequency definitions of probability. Hence, in some sense, the axiomatic definitions of probability "overcome" the deficiencies of the other definitions by making the interpretation of probability dispensable for mathematical model to be built. The third axioms is less obvious, stating that the probability of the union of unrelated events must be equal to the addition of their separate probabilities. For example, since $\{(HT)\} \cap \{(HH)\} = \varnothing$,

$$\begin{aligned} \mathcal{P}(\{(HT)\} \cup \{(HH)\}) &= \mathcal{P}(\{(HT)\}) + \mathcal{P}(\{(HH)\}) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

Again this coincides with the "frequency interpretation" result.

Definition. (Probability Space):
A sample space $\mathcal{S}$ endowed with a σ-field $\mathcal{F}$ and a probability measure $\mathcal{P}(\cdot)$ is called a *probability space*. That is we call the triple $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ a probability space. ■

Finally, as far as condition (c) of $\mathcal{E}$ is concerned, yet to be formalized, it will prove of paramount importance in the context of the limit theorems in Chapter 4.

2.3 Conditional Probability

So far we have considered probabilities of events on the assumption that no information is available relating to the outcome of a particular trial. Sometimes, however, additional information is available in the form of the *known* occurrence of some event A. For example, in the case of tossing a fair coin twice we might know that in the first trial it was heads. What difference does this information make to the original triple $(\mathcal{S}, \mathcal{F}, \mathcal{P})$?

Firstly, knowing that the first trial was a head, the set of all possible outcomes now becomes

$$\mathcal{S}_A = \{(HT), (HH)\},$$

since $(TH), (TT)$ are no longer possible. Secondly, the σ-field taken to become[6]

$$\mathcal{F}_A = \{\varnothing, \{(HT)\}, \{(HH)\}, \mathcal{S}_A\}.$$

Thirdly the probability measure function become

$$\mathcal{P}_A(\mathcal{S}_A) = 1, \;\; \mathcal{P}_A(\varnothing) = 0, \;\; \mathcal{P}_A(\{(HT)\}) = \frac{1}{2}, \;\; \mathcal{P}_A(\{(HH)\}) = \frac{1}{2}.$$

Thus, knowing that event A-one H has occurred in the first trial transformed the original probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ to the *conditional probability space* $(\mathcal{S}_A, \mathcal{F}_A, \mathcal{P}_A)$. The question that naturally arises is to what extent we can derive the above conditional probabilities without having to transform the original probability space. The following formula provides us with a way to calculate the conditional probability.

[6]Since $\mathcal{S}_A = \{(HT), (HH)\}$, therefore $\overline{(HT)} = (HH)$ and $\overline{(HH)} = (HT)$.

Definition. (Conditional Probability)
Let A_1 and A be any two events defined on $\mathcal{S}$ such that $\mathcal{P}(A) > 0$. The conditional probability of A_1, assuming that A has already occurred, is written as

$$\mathcal{P}_A(A_1) = \mathcal{P}(A_1|A) = \frac{\mathcal{P}(A_1 \cap A)}{\mathcal{P}(A)}.$$ ■

Example.
Let $A_1 = \{(HT)\}$ and $A = \{(HT), (HH)\}$, then since $\mathcal{P}(A_1) = \frac{1}{4}$, $\mathcal{P}(A) = \frac{1}{2}$, $\mathcal{P}(A_1 \cap A) = \mathcal{P}(\{(HT)\}) = \frac{1}{4}$,

$$\mathcal{P}_A(A_1) = \mathcal{P}(A_1|A) = \frac{1/4}{1/2} = \frac{1}{2},$$

as above. ■

Using the above rule of conditional probability we can deduce that

$$\begin{aligned} \mathcal{P}(A_1 \cap A_2) &= \mathcal{P}(A_1|A_2) \cdot \mathcal{P}(A_2) \\ &= \mathcal{P}(A_2|A_1) \cdot \mathcal{P}(A_1) \quad for\ A_1, A_2 \in \mathcal{F}. \end{aligned}$$

This is called the *multiplication rule*.

Moreover, when knowing that A_2 has occurred does not change the original probability of A_1, i.e.

$$\mathcal{P}(A_1|A_2) = \mathcal{P}(A_1),$$

we say that A_1 and A_2 are *independent*. The following definition gives a necessary and sufficient condition for two events to be independent.

Definition. (Independence)
Two events A_1 and A_2 are said to be independent if $\mathcal{P}(A_1 \cap A_2) = \mathcal{P}(A_1) \cdot \mathcal{P}(A_2)$.[7] ■

Independence is very different from *mutual exclusiveness* in the sense that $A_1 \cap A_2 = \varnothing$ but $\mathcal{P}(A_1|A_2) \neq \mathcal{P}(A_1)$ and vice versa can both arise. Independence is a probabilistic statement which ensures that the occurrence of one event does not

[7] To see this, recall that $\mathcal{P}(A_1|A_2) = \frac{\mathcal{P}(A_1 \cap A_2)}{\mathcal{P}(A_2)} = \mathcal{P}(A_1)$ when they are independent.

influence the occurrence (or non-occurrence) of the other event. On the other hand, mutual exclusiveness is a statement which refers to the events (set) themselves not the associated probability. Two events are said to be mutually exclusive when they cannot occur together.

3 Random Variable and Its Distribution

The model based on $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ does not provide us with a flexible enough framework because the sample space are not expressed in numbers as in the real world outcome. The basic idea underlying the construction of $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ was to set up a framework for studying probability of events as a prelude to analyse problem involving uncertainty. One facet of $\mathcal{E}$ which can help us suggest a more flexible probabilities space is the fact when the experiment is performed the outcome is often considered in relation to some *quantifiable attribute*; i.e. an attribute which can be repressed in numbers. It turns out that assigning numbers to *qualitative* outcome make possible a much more flexible formulation of probability theory. This suggests that if we could find a consistent way to assign numbers to outcomes we might be able to change $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ to something more easily handled. The concept of a *random variable* is designed to just that *without changing the underlying probabilistic structure of* $(\mathcal{S}, \mathcal{F}, \mathcal{P})$.

3.1 The Concept of a Random Variable

Let us consider the possibility of defining a function $X(\cdot)$ which maps $\mathcal{S}$ directly into the real line $\mathbb{R}$, that is,

$$X(\cdot) : \mathcal{S} \to \mathbb{R}_X,$$

assigning a real number x_1 to each s_1 in $\mathcal{S}$ by $x_1 = X(s_1)$, $x_1 \in \mathbb{R}, s_1 \in \mathcal{S}$. The question arises as to whether every function from $\mathcal{S}$ to $\mathbb{R}_x$ will provided us with a consistent way of attaching numbers to elementary events; consistent in the sense of preserving the event structure of the probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P})$. The answer, unsurprisingly, is not. This is because, although X is a function defined on $\mathcal{S}$, probabilities are assigned to events in $\mathcal{F}$ and the issue we have to face is how to define the value taken by X for *the different elements* of $\mathcal{S}$ in a way which preserves the event structures of $\mathcal{F}$. What we require from $X^{-1}(\cdot)$ or (X) is to provide us with a correspondence between $\mathbb{R}_x$ and $\mathcal{S}$ which reflects the event structures of $\mathcal{F}$, that is, it preserves union, intersections and complements. In other word for each subset N of $\mathbb{R}_x$, the inverse image $X^{-1}(N)$ must be an event in $\mathcal{F}$. This prompts us to define a random variable X to be any function satisfying this *event preserving condition* in relation to some σ-field defined on $\mathbb{R}_x$; for generality we always take the Borel field $\mathcal{B}$ on $\mathbb{R}$.

Definition. (Random Variable)
A random variable X is a real valued function from $\mathcal{S}$ to $\mathbb{R}$ which satisfies the condition that for each Borel set $B \in \mathcal{B}$ on $\mathbb{R}$, the set $X^{-1}(B) = \{s : X(s) \in B, s \in \mathcal{S}\}$ is an event in $\mathcal{F}$. ■

Example.
Define the function X—"the number of Heads in these two trial", then $X(\{HH\}) = 2$, $X(\{TH\}) = 1$, $X(\{HT\}) = 1$, and $X(\{TT\}) = 0$. Further we see that $X^{-1}(2) = \{(HH)\}$, $X^{-1}(1) = \{(TH), (HT)\}$ and $X^{-1}(0) = \{(TT)\}$. In fact, it can be shown that the σ-field related to the random variables, X, so defined is

$$\mathcal{F}_X = \{S, \varnothing, \{(HH)\}, \{(TT)\}, \{(TH), (HT)\}, \{(HH), (TT)\},$$
$$\{(HT), (TH), (HH)\}, \{(HT), (TH), (TT)\}\}.$$

We can verify that $X^{-1}(\{0\}) \cup \{1\}) = \{(HT), (TH), (TT)\} \in \mathcal{F}_X$, $X^{-1}(\{0\}) \cup \{2\}) = \{(HH), (TT)\} \in \mathcal{F}_X$ and $X^{-1}(\{1\}) \cup \{2\}) = \{(HT), (TH), (HH)\} \in \mathcal{F}_X$. ■

Example.
Consider the random variable Y—"the number of Head in the first trial", then $Y(\{HH\}) = Y(\{HT\}) = 1$, and $Y(\{TT\}) = Y(\{TH\}) = 0$. However Y does not preserve the event structures of $\mathcal{F}_X$ since $Y^{-1}(\{0\}) = \{(TH), (TT)\}$ is not an event in $\mathcal{F}_X$ and so does $Y^{-1}(\{1\}) = \{(HH), (HT)\}$. ■

From the two examples above, we see that the question "$X(\cdot) : \mathcal{S} \to \mathbb{R}_X$ is a random variable ?" does not make any sense unless some σ-field $\mathcal{F}$ is also specified. In the case of the function X–number of heads, in the coin-tossing example we see that it is a random variable relative to the σ-field $\mathcal{F}_X$. This, however, does not preclude Y from being a random variable with respect to some other σ-field $\mathcal{F}_Y$; for instance $\mathcal{F}_Y = \{\mathcal{S}, \varnothing, \{(HH), (HT)\}, \{(TH), (TT)\}\}$. Intuition suggests that for any real value function $X(\cdot) : \mathcal{S} \to \mathbb{R}$ we *should be able* to define a σ-field $\mathcal{F}_X$ on $\mathcal{S}$ such that X is a random variable. The concept of a σ-field generated by a random variable enables us to concentrate on particular aspects of an experiment without having to consider everything associated with the experiment at the same time. Hence, when we choose to

define a random variable and the associated σ-field we make an implicit choice about the features of the random experiment we are interested in.

How do we decide that some function $X(\cdot) : \mathcal{S} \to \mathbb{R}$ is a random variables relative to a given σ-field $\mathcal{F}$? From the discussion of the σ-field $\sigma(J)$ generated by the set $J = \{B_x : x \in \mathbb{R}\}$ where $B_x = (-\infty, x]$ we know that $\mathcal{B} = \sigma(J)$ and if $X(\cdot)$ is such that

$$X^{-1}((-\infty, x]) = \{s : X(s) \in (-\infty, x], s \in \mathcal{S}\} \in \mathcal{F} \; for \; all \; (-\infty, x] \in \mathcal{B},$$

then

$$X^{-1}(B) = \{s : X(s) \in B, s \in \mathcal{S}\} \in \mathcal{F} \; for \; all \; B \in \mathcal{B}.$$

In other words, when we want to establish that X is a random variables we have to look no further than the half-closed interval $(-\infty, x]$, and the σ-field $\sigma(J)$ they generate, whatever the range $\mathbb{R}_x$. Let us use the shorthand notation $\{X(s) \leq x\}$ instead of $\{s : X(s) \in (-\infty, x], s \in \mathcal{S}\}$ to the above "numbers of Heads " example,

$$\begin{aligned} X^{-1}((-\infty, x]) &= \{s : X(s) \leq x\} \\ &= \begin{cases} \varnothing & x < 0, \\ \{(TT)\} & x = 0, \\ \{(TT)(TH)(HT)\} & x = 1, \\ \{(TT)(TH)(HT)(HH)\} & x = 2. \end{cases} \end{aligned}$$

we can see that $X^{-1}((-\infty, x]) \in \mathcal{F}_X$ for all $x \in \mathbb{R}$ and thus $X(\cdot)$ is a random variables with respect to $\mathcal{F}_X$.[8] We thus have the following equivalent definitions of random variable.

Definition. (Random Variable)
A real valued function X defined on the probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ is called a random variable if the set $\{s : X(s) \leq x\} \in \mathcal{F}$ for every $x \in \mathbb{R}$. ■

A random variable X relative to $\mathcal{F}$ maps $\mathcal{S}$ into a subset of the real line, and the Borel field $\mathcal{B}$ on $\mathbb{R}$ plays now the role of $\mathcal{F}$. In order to complete the model we need to assign probabilities to the elements B of $\mathcal{B}$. Common sense suggests that the assignment of the probabilities to the event $B \in \mathcal{B}$ must be consistent with the

[8] To see this, since it is known that $\{(TT)\} \in \mathcal{F}_X$, therefore $\overline{\{(TT)\}} = \{(TH)(HT)(HH)\} \in \mathcal{F}_X$. Hence $\{(TH)(HT)(HH)\} \cap \{(TT)(TH)(HT)\} = \{(TH)(HT)\} \in \mathcal{F}_X$. Finally, $\overline{\{(TH)(HT)\}} = \{(TT), (HH)\} \in \mathcal{F}_X$.

probabilities assigned to the corresponding events in $\mathcal{F}$. Formally, we need to define a set function on the Borel-field $P_X(\cdot) : \mathcal{B} \to [0,1]$ such that

$$P_X(B) = \mathcal{P}(X^{-1}(B)) \equiv \mathcal{P}(s : X(s) \in B, s \in S) \text{ for all } B \in \mathcal{B}.$$

In the above "number of Heads" example,
$P_X(\{0\}) = 1/4 = \mathcal{P}(\{TT\})$, $P_X(\{1\}) = 1/2 = \mathcal{P}(\{HT\}) + \mathcal{P}(\{TH\})$, $P_X(\{2\}) = 1/4 = \mathcal{P}(\{HH\})$ and $P_x(\{0\} \cup \{1\}) = 3/4$.

The question which arises is whether, in order to define the set function $P_X(\cdot)$, we need to consider all the elements of the Borel field $\mathcal{B}$. The answer is that we *do not need* to do that because, as argued above, any such element of $\mathcal{B}$ can be expressed in terms of the semi-closed intervals $(-\infty, x]$. This implies that by choosing such semi-closed intervals "intelligently", we can define $P_X(\cdot)$ with the minimum of effort. For example, we may define:

$$P_X((-\infty, x]) = \begin{cases} 0 & x < 0, \\ \frac{1}{4} & x = 0, \\ \frac{3}{4} & x = 1, \\ 1 & x = 2. \end{cases}$$

As we can see, the semi-closed intervals were chosen to divide the real line at the points corresponding to the value taken by X. This way of defining the semi-closed intervals is clearly non-unique but will prove very convenient in the next subsection.

In fact, the event and probability structure of $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$ is preserved in the induced probability space $(\mathbb{R}, \mathcal{B}, P_X(\cdot))$. We traded $\mathcal{S}$, a set of arbitrary elements, for $\mathbb{R}$, the real line; $\mathcal{F}$ a σ-field of subset of $\mathcal{S}$ with $\mathcal{B}$, the Borel field on the real line; and $\mathcal{P}(\cdot)$ a set function defined on arbitrary sets with $P_X(\cdot)$, a set function on semi-closed intervals of the real line.

3.2 The Distribution and Density Functions

In the previous section the introduction of the concept of a random variable X, enables us to trade the probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$ for $(\mathbb{R}, \mathcal{B}, P_X(\cdot))$ which has a much more

convenient mathematical structure. The latter probability space, however, is not as yet simple enough because $P_X(\cdot)$ is still a set function albeit on real line intervals. In order to simplify it we need to transform it into a point function with which we are so familiar. Define a point function

$$F(\cdot) : \mathbb{R}_X \to [0, 1],$$

which is seemingly, only a function of x. In fact, however, this function will do exactly the same job as $P_X(\cdot)$. Heuristically, this is achieved by defining $F(\cdot)$ as a point function by

$$P_X((-\infty, x]) = F(x) - F(-\infty), \quad for\ all\ x \in \mathbb{R},$$

and assigning the value zero to $F(-\infty)$.

Definition. (Distribution Function):
Let X be a random variable defined on $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$. The point function $F(\cdot) : \mathbb{R}_X \to [0, 1]$ defined by

$$F(x) = P_X((-\infty, x]) = Pr(X \leq x), \quad for\ all\ x \in \mathbb{R}$$

is called the *distribution function* of X and satisfied the following properties:
(a). $F(x)$ is non-decreasing;
(b). $F(-\infty)$=$\lim_{x\to-\infty} F(x) = 0$ and $F(\infty)$=$\lim_{x\to\infty} F(x) = 1$,
(c). $F(x)$ is continuous from the right.[9] (i.e. $\lim_{h\to 0} F(x + h) = F(x),\ \forall x \in \mathbb{R}$.) ■

The great advantage of $F(\cdot)$ over $\mathcal{P}(\cdot)$ and $P_X(\cdot)$ is that the former is a point function and can be represented in the form of an algebraic formula; the kind of functions we are so familiar with in elementary mathematics.

Discrete random variables are conceptually easy to define: for experiments whose sample space are either finite or countably infinite. However, continuous random variables need to approached in a more roundabout fashion.

Definition. (Discrete Random Variable)
A random variable X is called discrete if its range $\mathbb{R}_X$ is some subsets of the set of

[9] A function $f : D \to \mathbb{R}$ is left-continuous at $x = a$ if $\lim_{x\to a^-} f(x) = f(a)$. It is right-continuous at $x = a$ if $\lim_{x\to a^+} f(x) = f(a)$.

integers $Z = \{0, \pm 1, \pm 2, ...\}$. ■

Definition. (Continuous Random Variable)
A random variable X is called continuous if its distribution function $F(x)$ is continuous for all $x \in \mathbb{R}$ and there exists a non-negative function $f_X(\cdot)$ on the real line such that

$$F(x) = \int_{-\infty}^{x} f_X(u)du, \ \ \forall x \in \mathbb{R}_X. \qquad \blacksquare$$

In defining the concept of a continuous *r.v.* we have introduced the function $f_X(x)$ which is directly related to $F(x)$.

Definition. (Probability Density Function)
Let $F(x)$ be the density function of the random variable X. The non-negative function $f(x)$ defined by

$$F(x) = \int_{-\infty}^{x} f_X(u)du, \ \ \forall x \in \mathbb{R}_X; \text{ (when X is a continuous random variable)}$$

or

$$F(x) = \sum_{u \le x} f_X(u), \ \ \forall x \in \mathbb{R}_X; \text{ (when X is a discrete random variable)}$$

is said to be the *probability density function* (*pdf*) of X. The *pdf* satisfies the following properties:
(a). $f_X(x) \ge 0$, $\forall x \in \mathbb{R}_X$;
(b). $\int_{-\infty}^{\infty} f_X(x)dx = 1$;
(c). $Prob(a < X < b) = \int_a^b f_X(x)dx$;
(d). $f_X(x) = \frac{d}{dx}F(x)$, at every point where the distribution function is continuous. ■

Although we can use the distribution function $F(x)$ as the fundamental concept of our probability model we prefer to adopt the density function $f_X(x)$ instead, because we gain in simplicity and added intuition. It enhance intuition to view density function as distribution probability mass over the range of X.

Example.

Let X be uniformly distributed in the interval $[a, b]$ and we write $X \sim U(a, b)$. The DF of X takes the form:

$$F(x) = \begin{cases} 0 & x < a, \\ \frac{x-a}{b-a} & a \leq x < b, \\ 1 & x \geq b. \end{cases}$$

The corresponding pdf of X is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & elsewhere. \end{cases}$$

3.3 Numerical Characteristics of Random Variables

Probability density function provide a global overview of a random variable's behavior. If X is discrete, $f_X(x)$, gives $P(X = x)$ for all x; if X is continuous, and A is an interval, or countable union of intervals, $P(X \in A) = \int_A f_X(x)dx$. Detail that explicit, though, is not always necessary-or even helpful. There are times when a more prudent strategy is to focus the information contained in $f_X(x)$ by summarizing certain of its features with a numerical characteristics. Furthermore, in modeling real phenomena using probability model of the form $\Phi = \{f(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ we need to be able to postulate such models having only a general quantitative description of the random variable in question at our disposal a priori. Such information comes in the form of these numerical characteristics of random variables such as the mean, the variance, the skewness and kurtosis coefficients and higher moments. Indeed, sometimes such numerical characteristics actually determine the type of probability density in Φ. Moreover, the analysis of density functions is usually undertaken in terms of these numerical characteristics. The search for these numerical characteristics is what the remainder of this section is primarily about.

3.3.1 Mathematical Expectation

The first feature of a *pdf* that we will examine is *central tendency*, a term referring to the "average" value of a random variable.

Definition. (Mean)
The mean (expected value or expectation) of X denoted by $E(X)$ is defined by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \; --for\ a\ continous\ r.v,$$

and

$$E(X) = \sum_{i} x_i f(x_i) \; --for\ a\ discrete\ r.v,$$

when the integral and sum exist. We always denote $E(X) = \mu$. ■

Example.
If $X \sim U(a, b)$, i.e. X is uniformly distributed r.v., then

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{a}^{b} x\left(\frac{1}{b-a}\right)dx = \frac{1}{2}\left(\frac{1}{b-a}\right)x^2\Big|_{a}^{b} = \frac{a+b}{2}.$$ ■

In the above example the mean of the r.v. X exists. The condition which guarantees the existence of $E(X)$ is that

$$E|X| = \int_{-\infty}^{\infty} |x|f(x)dx < \infty. \quad (since\ E(X) \leq E|X|)$$

Example.
One example where the mean does not exist is the cases of a Cauchy distributed r.v. with a pdf given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \; x \in \mathbb{R},$$

then the expectation of X would be

$$\begin{aligned} E|X| = \int_{-\infty}^{\infty} x f_X(x)dx &= \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}dx \\ &= \frac{\log(1+x^2)}{\pi}\Big|_{-\infty}^{\infty} \\ &= \infty - (-\infty), \end{aligned}$$

which is indeterminate. In this case we say that is $E(X)$ doesn't exist for the Cauchy distribution. ■

Results.

Some properties of the expectation are

(a). $E(c) = c$, if c is a constant.

(b). $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$ for any two r.v.'s X_1 and X_2 whose mean exist and a, b are real constants.

(c). Let X be a random variable such that $E|X| < \infty$, then for every $\varepsilon > 0$, $Pr(|X| \geq \varepsilon) \leq \frac{E|X|}{\varepsilon}$. This is the so called the *Markov Inequality.* ■

3.3.2 The Variance

The expected value is a good enough measure of central tendency, but it still leaves out some critical information about a random variable's behavior. Unless we are also provided with some indication of how "spread out" a random variable's probability density function, the expected value by itself can be misleading. One seemingly reasonable approach would be to average, in generalized sense, the squared deviations the values of X from their expected value.

Definition. (The Variance)

Related to the mean as a measure of location is the dispersion measure called the

variance and defined by

$$\begin{aligned} Var(X) &= E[X - E(X)]^2 \\ &= \int_{-\infty}^{\infty} (X-\mu)^2 f(x)dx \\ &= E(X^2) - \mu^2 \\ &\equiv \sigma^2. \end{aligned}$$ ■

Note that the square root of the variance is referred to as standard deviation.

Example.

Let $X \sim U(a,b)$, then

$$Var(X) = \int_{-\infty}^{\infty} \left[X - \left(\frac{a+b}{2}\right)\right]^2 \left(\frac{1}{b-a}\right) dx = \frac{(b-a)^2}{12}.$$ ■

Results.

Some properties of the variance are:

(a). $Var(c) = 0$ for any constant c.

(b). $Var(aX+b) = a^2\, Var(X)$, for constant a and b.

(c). $Pr(|X - E(X)| \geq \varepsilon) \leq [Var(X)]/\varepsilon^2$ for $\varepsilon > 0$. This is the so called the *Chebyshev's Inequality.* ■

3.3.3 Higher Moments

The quantities we have identified as the mean and the variance are actually special cases of what are referred to more generally as a random variable's moment. Specially, $E(X)$ is the *first moment about the origin* and σ^2, the *second moment about the mean.* As the terminology suggests, we will have occasion to define even "higher" moments of X.

Definition.

(a). r-th Row Moments is the moment of inertia from $x = 0$ defined by:

$$\mu'_r \equiv E(X^r) = \int_{-\infty}^{\infty} x^r f(x)dx, \quad r = 0, 1, 2, ..,$$

(b). r-th Central Moments is defined as the moment around $x = \mu$:

$$\mu_r \equiv E(X - \mu)^r = \int_{-\infty}^{\infty} (x - \mu)^r f(x)dx, \quad r = 0, 1, 2, ... \qquad \blacksquare$$

Example.

These higher moments are sometimes useful in providing us with further information relating to the distribution and density function of r.v.'s. In particular, the 3rd and 4th central moments, when standardized in the form:

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

and

$$\alpha_4 = \frac{\mu_4}{\sigma^4}$$

are referred to measure of *skewness* and *kurtosis* and provided us with measures of asymmetry and flatness of peak, respectively. $\blacksquare$

3.4 Some Univariate Distribution

3.4.1 The Normal Distribution

Finding probability distribution to describe empirical data is one of the most important contribution a statistician can make to the research scientist. By far the most widely used probability model in statistics is the *normal* distribution.

Definition.

A random variable X, $x \in \mathbb{R}$, is normally distributed if its probability density function is given by

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}_+. \tag{2-1}$$

We often express this by $X \sim N(\mu,\sigma^2)$.[10] ■

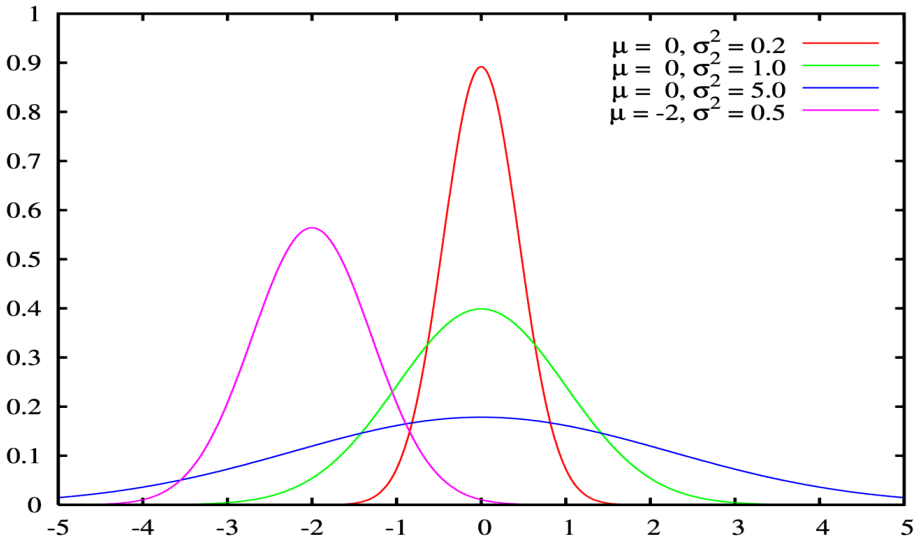

Figure (2-2). The Normal Distribution.

As far as the shape of the normal distribution and density function are concerned we note the following characteristics.

Results.

(a). The normal density is symmetric about μ, i.e.

$$f(\mu+k) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{k^2}{2\sigma^2}\right\} = f(\mu-k)$$
$$\Rightarrow Pr(\mu \le X \le \mu+k) = Pr(\mu-k \le X \le \mu),\quad k>0,$$

[10] The notation here is $\boldsymbol{\theta} = [\mu\ \ \sigma^2]'$ and $\boldsymbol{\Theta} = [\mathbb{R}\ \ \mathbb{R}_+]'$.

and for the distribution function,

$$F(-x) = 1 - F(x + 2\mu).$$

(b). The density function attains its maximum at $x = \mu$,

$$\frac{df(x)}{dx} = f(x)\left(\frac{-2(x-\mu)}{2\sigma^2}\right) = 0 \Rightarrow x = \mu, \text{ and } f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}.$$

(c). The density function has two points of inflection at $x = \mu + \pm\sigma$:

$$\frac{d^2f(x)}{dx^2} = \frac{\sigma^{-3}}{\sqrt{2\pi}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right] = 0 \Rightarrow x = \mu \pm \sigma. \blacksquare$$

A special case of normal distribution, which becomes the most useful of all probability distribution, is the one with $\mu = 0$ and $\sigma^2 = 1$.

Definition.
The density function of the random variable Z,

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

which does not depend on the unknown parameters μ, σ. This is called the standard normal distribution, which we write in this form, $Z \sim N(0, 1)$. ■

Exercise 2.
Show that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-z^2}{2}\right) dz = 1. \quad \blacksquare \tag{2-2}$$

3.4.2 Exponential Family of Distribution

A continuous random variable X has a gamma distribution with parameters η and λ, written

$$f(x) = \frac{\lambda^\eta}{\Gamma(\eta)} e^{-\lambda x} x^{\eta-1},\ x \le 0,\ \lambda > 0,\ \eta > 0.$$

Many familiar distributions are special cases, including the exponential ($\eta = 1$), and chi-squared ($\lambda = 1/2,\ \eta = n/2$).

3.5 The Notation of a Probability Model

In a continuous random variable, it is impossible to get the pdf $f(x)$ from the random experiment $\mathcal{E}$ directly (either it is costly or is impossible to know all X), we have to *assume* a (parametric) probability distribution to model a particular real phenomenon by previous experience in modeling similar phenomenon or by a preliminary study of the data.

By a parameterized probability model, we may transform the original uncertainty related to $\mathcal{E}$ to uncertainty related to *unknown parameters* $\boldsymbol{\theta}$ of $f(\cdot)$; in order to emphasize this we write the pdf as $f(x;\boldsymbol{\theta})$. We are now in a position to define our probability model in the form of parametric family of density function which we denote by

$$\Phi = \{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

Φ represents a set of density functions indexed by the unknown parameters $\boldsymbol{\theta}$ which are assumed to belong to a parameter space $\boldsymbol{\Theta}$.

Example.

The Pareto distribution:

$$\Phi = \left\{ f(x;\theta) = \frac{\theta}{x_0}\left(\frac{x_0}{x}\right)^{\theta+1},\ x > x_0,\ \theta \in \Theta \right\},$$

x_0–a known number, $\Theta \in \mathbb{R}_+$–the positive real line. For each value θ in Θ, $f(x;\theta)$ represents a different density. ■

When a particular parameter family of densities Φ is chosen, as the appropriate probability model for modeling a real phenomenon, we are in effect assuming that the observed data available were generated by the "*chance mechanism*" described by one of those density in $\boldsymbol{\Theta}$. The original uncertainty relating to the outcome of a particular trial of the experiment has now been transformed into the uncertainty relating to the choice of one $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$, say $\boldsymbol{\theta}^*$ which determines uniquely the one density, that is, $f(x;\boldsymbol{\theta}^*)$, which gives rise the observed data. The task of estimating $\boldsymbol{\theta}^*$ or testing some hypotheses about $\boldsymbol{\theta}^*$ using the observed data lies with *statistical inference* in next following chapters.

4 Random Vector and Its Distribution

The probability model formulated in the previous chapter was in the form of a parametric family of densities associated with a random variable $X : \Phi = \{f(x;\theta),\ \theta \in \Theta\}$. In practice, however, there are many observable phenomena where the outcome comes in the form of several quantitative attributes. For example, data on personal income might be related to number of children, social class, type of occupation, age class, etc. In order to be able to model such real phenomena we extend a single r.v.'s framework to one for multidimensional r.v.'s or random vectors, that is,

$$\mathbf{x} = (X_1, X_2, ..., X_k)',$$

where each X_i, $i = 1, 2, ..., k$ measures a particular quantifiable attribute of the random experiment's ($\mathcal{E}$) outcome.

4.1 Joint Distribution and Density functions

Consider the random experiment $\mathcal{E}$ of tossing a fair coin twice. Define the function $X_1(\cdot)$ to be the number of heads and $X_2(\cdot)$ to be the number of tails. The function $(X_1(\cdot), X_2(\cdot)) : \mathcal{S} \to \mathbb{R}^2$ is a two dimensional vector function which assigns to each elements s of $\mathcal{S}$, the pair of ordered numbers (x_1, x_2) where $x_1 = X_1(s), x_2 = X_2(s)$.

Definition.
A (bivariate) random vector $\mathbf{x}(\cdot)$ is a vector function

$$\mathbf{x}(\cdot) : \mathcal{S} \to \mathbb{R}^2,$$

such that for any two real numbers $(x_1, x_2) \equiv \boldsymbol{x}$, the event

$$\mathbf{x}^{-1}((-\infty, \boldsymbol{x}]) = \{s : -\infty < X_1(s) \leq x_1,\ -\infty < X_2(s) \leq x_2, s \in \mathcal{S}\} \in \mathcal{F}. \quad \blacksquare$$

The random vector induces a probability space $(\mathbb{R}^2, \mathcal{B}^2, P_{\mathbf{x}}(\cdot))$, where $\mathcal{B}^2$ $(\equiv \mathcal{B} \times \mathcal{B})$ are Borel subsets on the plane and $P_{\mathbf{x}}(\cdot)$ a probability set function defined over events

in $\mathcal{B}^2$, in a way which preserves the probability structure of the original probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$. This is achieved by attributing to each $\mathbf{B} \in \mathcal{B}^2$ the probability

$$P_{\mathbf{x}}(\mathbf{B}) = \mathcal{P}(\{s : (X_1(s), X_2(s)) \in \mathbf{B}\})$$

or

$$P_{\mathbf{x}}((-\infty, \boldsymbol{x}]) = Pr(X_1 \leq x_1,\ X_2 \leq x_2).$$

We can go a step further to reduce the step function $P_X(\cdot)$ to a point function $F(x_1, x_2)$, we call the joint (cumulative) distribution function.

Definition.

Let $\mathbf{x} \equiv (X_1, X_2)'$ be a random vector defined on $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$. The function defined by

$$F(\cdot, \cdot) : \mathbb{R}^2 \to [0, 1],$$

such that

$$F(\boldsymbol{x}) \equiv F(x_1, x_2) = P_X((-\infty, \boldsymbol{x}]) = Pr(X_1 \leq x_1, X_2 \leq x_2) \equiv Pr(\mathbf{x} \leq \boldsymbol{x})$$

is said to be the *joint distribution function* of $\mathbf{x}$. ■

Example.

In the coin-tossing example above, the random vector $\mathbf{x}(\cdot)$ takes the value $(1, 1)$, $(2, 0)$,$(0, 2)$ with probability $\frac{1}{2}, \frac{1}{4}$ and $\frac{1}{4}$, respectively. In order to derive the joint distribution (DF) we have to define all the events of the form $\{s : X_1(s) \leq x_1, X_2(s) \leq x_2, s \in S\}$ for all $(x_1, x_2) \in \mathbb{R}^2$:

$$\begin{aligned}
&\{s : X_1(s) \leq x_1, X_2(s) \leq x_2, s \in S\} \\
&= \begin{cases} \varnothing & x_1 < 0, x_2 < 0, \\ \{TT\} & x_1 = 0, x_2 = 2, \\ \{(TT), (TH), (HT)\} & x_1 = 1, x_2 = 2, \\ S & x_1 = 2, x_2 = 2. \end{cases}
\end{aligned}$$

The joint distribution function of X_1 and X_2 is given by

$$F(x_1,x_2)=\begin{cases}0 & x_1<0,x_2<0,\\ \frac{1}{4} & x_1=0,x_2=2,\\ \frac{3}{4} & x_1=1,x_2=2,\\ 1 & x_1=2,x_2=2.\end{cases}$$

■

Definition.

The joint distribution of X_1 and X_2 is called continuous if there exists a non-negative function $f(x_1,x_2)$ such that

$$F(x_1,x_2)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}f(u,v)du\,dv,$$

where the function $f(x_1,x_2)$ is called the joint density function of X_1 and X_2. ■

Results.

The joint density function $f(x_1,x_2)$ implies the following properties:

(a). $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x_1,x_2)dx_1\,dx_2=1$.

(b). $Pr(a<X_1\le b,c<X_2\le d)=\int_a^b\int_c^d f(x_1,x_2)dx_1\,dx_2$.

(c). $f(x_1,x_2)=\frac{\partial^2}{\partial x_1\partial x_2}F(x_1,x_2)$, if $f(\cdot)$ is continuous at (x_1,x_2). ■

Example. (Bivariate Normal Distribution)

A 2×1 random vector $\mathbf{x}=(X_1,X_2)'$ is said to follow a bivariate normal distribution, if their joint density function can be written as

$$f(x_1,x_2;\boldsymbol{\theta})=\frac{(1-\rho^2)^{-1/2}}{2\pi\sigma_1\sigma_2}$$
$$exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\},$$

$x_1,x_2\in\mathbb{R}$, and $\boldsymbol{\theta}=(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)\in\mathbb{R}^2\times\mathbb{R}_+^2\times[0,1]$. Here, $E(X_i)=\mu_i$, $Var(X_i)=\sigma_i^2$, $i=1,2$; and $Cov(X_1,X_2)=\rho$. ■

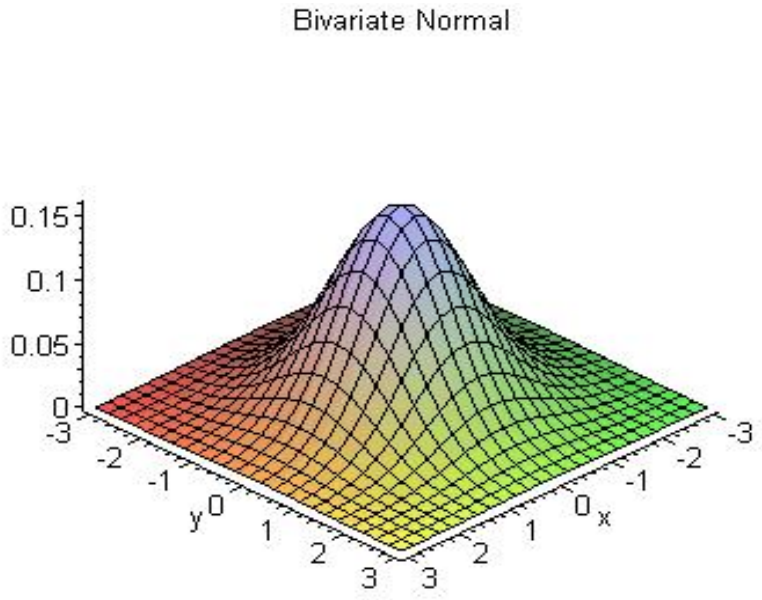

Figure (2-3). Bivariate Normal Distribution

The extension of the concept of a random variable X to that of a random vector $\mathbf{x} = (X_1, X_2, ..., X_n)'$ enables us to generalize the probability model

$$\Phi = \{f(\mathbf{x}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

to that of a parametric family of joint density functions

$$\Phi = \{f(x_1, x_2, ..., x_n; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

This is a very important generalization since in most applied disciplines, the real phenomena to be modeled are usually multidimensional in the sense that there is more than one quantifiable features to be considered.

Notation.

We are now in a right position to clarify our notations used in this handouts.

(A). For nonstochastic cases:

(a). a,x,y etc.: is an element, (1×1).

(b). $\mathbf{a}, \mathbf{x}, \mathbf{y}$: is a column vector, $(n \times 1)$.

(c). $\mathbf{A}, \mathbf{X}, \mathbf{Y}$: is a matrix, $(m \times n)$.

B. For stochastic cases:

(a). 1×1:

(i). X: random variable;

(ii). x: the value that X takes;

(iii). $f_X(x)$: the "probability" that the random variable X takes on the value x.

(b). $n \times 1$:

(i). $\mathbf{x} = (X_1, X_2, ..., X_n)'$: random vector;

(ii). $\boldsymbol{x} = (x_1, x_2, ..., x_n)'$: the value that $\mathbf{x}$ takes;

(iii). $f_{\mathbf{x}}(\boldsymbol{x})$:is the (joint) probability that $(X_1, X_2, ..., X_n)'$ takes on the values $\boldsymbol{x} = (x_1, x_2, ..., x_n)$.

(c). $\mathbf{X}$, $\mathbf{\Lambda}$ etc.: matrix (such as the regressors matrix and the variance- covariance matrix). ■

4.2 Marginal Distributions

Let $\mathbf{x} \equiv (X_1, X_2)'$ be a bivariate random vector defined on $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ with a joint distribution function $F(x_1, x_2)$. The question which naturally arises is whether we could separate X_1 and X_2 and consider them as individual random variables. The answer to this question leads us to the concept of a *marginal distribution*.

Definition. (Marginal Distribution)
The *marginal distribution functions* of X_1 and X_2 are defined by

$$F_1(x_1) = lim_{x_2 \to \infty} F(x_1, x_2).$$

and

$$F_2(x_2) = lim_{x_1 \to \infty} F(x_1, x_2).$$ ■

Having separated X_1 and X_2 we need to see whether they can be considered as single r.v.'s defined on the same probability space. In defining a random vector we imposed the condition that

$$\{s : X_1(s) \leq x_1, X_2(s) \leq x_2\} \in \mathcal{F}.$$

The definition of the marginal distribution we used the event

$$\{s : X_1(s) \leq x_1, X_2(s) \leq \infty\},$$

which we know belong to $\mathcal{F}$. This event, however, can be written as the intersection of two sets of the form

$$\{s : X_1(s) \leq x_1\} \cap \{s : X_2(s) \leq \infty\}$$

but the second set is $\mathcal{S}$ i.e. $\{s : X_2(s) \leq \infty\} = \mathcal{S}$,
which implies that

$$\{s : X_1(s) \leq x_1\} \cap \{s : X_2(s) \leq \infty\} = \{s : X_1(s) \leq x_1\},$$

which indeed belongs to $\mathcal{F}$ and it is the condition needed for X_1 to be a r.v. with a probability function $F_1(x_1)$; the same is true for X_2.

Definition. (Marginal Density Function)
The marginal density functions of X_1 and X_2 are defined (from standard iterated Riemann integrals on Cells) [11] by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

and

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

[11] By definition, $F_1(x_1) = F(x_1, \infty) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(u, v) dv du$. So $f(x_1) = \int_{-\infty}^{\infty} f(u, v) dv$.

Example.
For the random vector (X_1, X_2)=(no. of heads, no. of tails) above, the marginal density of $f_1(x_1)$ is "recovered" from

$$\begin{aligned} f_1(0) &= f(0,1) + f(0,1) + f(0,2) = 0 + 0 + 1/4 = 1/4, \\ f_1(1) &= f(1,0) + f(1,1) + f(1,2) = 0 + 1/2 + 0 = 1/2, \\ f_1(2) &= f(2,0) + f(2,1) + f(2,2) = 1/4 + 0 + 0 = 1/4. \end{aligned}$$

■

4.3 Independence of Random Variable

The concept of independent events that was introduced in section 2.3 leads quite naturally to a similar definition for independent random variables.

Definition.
Random variables X and Y are said to be *independent* if for all sets A and B,

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

■

A more workable characterization of this definition is provided in next Theorem.

Theorem.
Two random variables X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

for all x and y. *Independence* in terms of the distribution function takes the same form

$$F(x, y) = F(x) \cdot F(y).$$

■

It is quite obvious that knowing the joint density function of X_1 and X_2, we can derive their marginal density functions; the reverse, however, is not true in general. Knowledge of $f(x_1)$ and $f(x_2)$ is enough to derive $f(x_1, x_2)$ only when X_1 and X_2 are independent: $f(x_1, x_2) = f(x_1) \cdot f(x_2)$.

Result.

Let $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^l$ be a continuous function. If $\mathbf{z}$ and $\mathbf{y}$ be independent, then $\boldsymbol{g}(\mathbf{z})$ and $\boldsymbol{g}(\mathbf{y})$ are also independent.

Proof.

Let $\mathbf{A}_1 = [\mathbf{z} : \boldsymbol{g}(\boldsymbol{z}) \leq \mathbf{a}_1]$ and $\mathbf{A}_2 = [\mathbf{y} : \boldsymbol{g}(\boldsymbol{y}) \leq \mathbf{a}_2]$. Then $F_{\boldsymbol{g}(\mathbf{z})\boldsymbol{g}(\mathbf{y})}(\mathbf{a}_1, \mathbf{a}_2) \equiv P[\boldsymbol{g}(\mathbf{z}) \leq \mathbf{a}_1, \boldsymbol{g}(\mathbf{y}) \leq \mathbf{a}_2] = P[\mathbf{z} \in \mathbf{A}_1, \mathbf{y} \in \mathbf{A}_2] = P[\mathbf{z} \in \mathbf{A}_1] \cdot P[\mathbf{y} \in \mathbf{A}_2] = P[\boldsymbol{g}(\mathbf{z}) \leq \mathbf{a}_1] \cdot P[\boldsymbol{g}(\mathbf{y}) \leq \mathbf{a}_2] = F_{\boldsymbol{g}(\mathbf{z})}(\mathbf{a}_1) F_{\boldsymbol{g}(\mathbf{y})}(\mathbf{a}_2)$ for all $\mathbf{a}_1$, $\mathbf{a}_2 \in \mathbb{R}^l$. Hence $\boldsymbol{g}(\mathbf{z})$ and $\boldsymbol{g}(\mathbf{y})$ are independent. ■

4.4 Conditional Distributions

We have seen that the probability of independent event have their random variable counterparts. Another of these carryovers is the notion of conditional probability, or in what will be our terminology, a *conditional probability density function*. That is, in this section we consider the question of simplifying probability models Φ by *conditioning* with respect to some subsets of the r.v.'s.

In the context of the probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$ the conditional probability of **event** A_1 given **event** A_2 is defined by

$$\mathcal{P}(A_1|A_2) = \frac{\mathcal{P}(A_1 \cap A_2)}{\mathcal{P}(A_2)}, \quad \mathcal{P}(A_2) > 0; \quad A_1, A_2 \in \mathcal{F}.$$

Definition.

By using an analogous definition in term of distribution function, we define the condi-

tional density of X_1 given $X_2 = x_2$ to be

$$f_{X_1|X_2}(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \quad x_1 \in \mathbb{R}_{x_1}.$$

Similarly, the conditional density of X_2 given $X_1 = x_1$ is

$$f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \quad x_2 \in \mathbb{R}_{x_2},$$

provided $f_1(x_1) > 0$ and $f_2(x_2) > 0$.[12] ■

Example.

Let X and Y have joint density function $f(x, y) = y/10$, $(x, y) = (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)$. Then X has marginal density $f_X(1) = f(1, 1) + f(1, 2) + f(1, 3) = \frac{6}{10}$, $f_X(2) = f(2, 1) + f(2, 2) = \frac{3}{10}$, and $f_X(3) = f(3, 1) = \frac{1}{10}$. Therefore the conditional density function of Y given $x = 1$ is

$$f_{Y|X}(1|1) = \frac{f(1,1)}{f_X(1)} = \frac{1/10}{6/10} = \frac{1}{6},$$
$$f_{Y|X}(2|1) = \frac{f(1,2)}{f_X(1)} = \frac{2/10}{6/10} = \frac{2}{6},$$
$$f_{Y|X}(3|1) = \frac{f(1,3)}{f_X(1)} = \frac{3/10}{6/10} = \frac{3}{6}.$$

Similarly, the conditional density of Y given $X = 2$ is

$$f_{Y|X}(1|2) = \frac{f(2,1)}{f_X(2)} = \frac{1/10}{3/10} = \frac{1}{3},$$
$$f_{Y|X}(2|2) = \frac{f(2,2)}{f_X(2)} = \frac{2/10}{3/10} = \frac{2}{3}.$$ ■

Results.

(a) . The conditional density is a proper density function, i.e. for a given $X_2 = \breve{x}_2$,

(i). $f_{X_1|X_2}(x_1|\breve{x}_2) \geq 0$;

(ii). $\int_{-\infty}^{\infty} f_{X_1|X_2}(x_1/\breve{x}_2)dx_1 = 1$.

[12] However, the mathematical apparatus needed to bypass the problem that in a continuous random variable, X_1 and X_2, $f_1(x_1) = f_2(x_2) = 0$. This definition of conditional density does not make sense. See Billingsley (1979), p.354-407)

(b) . Knowledge of all these conditional densities is equivalent to knowledge of joint density, i.e.

$$\begin{aligned} f(x_1, x_2) &= f_{X_1|X_2}(x_1|x_2) \cdot f_2(x_2) \\ &= f_{X_2/X_1}(x_2|x_1) \cdot f_1(x_1), \quad (x_1, x_2) \in \mathbb{R}^2. \end{aligned}$$

(c) . An immediate implication of last equation is that if X_1 and X_2 are independent, then

$$f_{X_1|X_2}(x_1|x_2) = f_1(x_1), \qquad x_1 \in \mathbb{R}_{x_1}.$$ ■

Exercise 3.

Let $\mathbf{X} = (X_1, X_2, X_3)$ be a continuous random vector having joint density

$$f(x_1, x_2, x_3) = 6\exp(-x_1 - x_2 - x_3), \quad 0 < x_1 < x_2 < x_3.$$

Find marginal pdf of $f(x_2)$ and the conditional density of X_3 given $(X_1, X_2) = (x_1, x_2)$.■

5 The General Notation of Expectation

In section 3.3 we considered the notation of mathematical expectation in the context of the simple probability model

$$\Phi = \{f(x;\theta),\ \theta \in \Theta\}$$

as an useful characteristic of density functions of a single random variables. Since then we have generalized the probability model to

$$\mathbf{\Phi} = \{f(x_1, x_2, ..., x_k;\theta),\ \theta \in \Theta\}$$

and put forward a framework in the context of which joint density functions can be analyzed. These include marginalisation, conditioning and functions of random variables. The purpose of this section is to consider the notation of expectation in the context of this more general framework. For simplicity of exposition we consider the case where $k = 2$.

5.1 Expectation of a Marginal Random Variable

The expectation of a marginal random variable from a joint density is just as the definition

$$\begin{aligned} E(X_1) &= \int x_1 f_{X_1}(x_1)dx_1 = \int x_1 \left(\int f_{X_1\ X_2}(x_1\ x_2)dx_2 \right) dx_1 \\ &= \int\int x_1 f_{X_1\ X_2}(x_1\ x_2)dx_2 dx_1. \end{aligned}$$

We can generalize the above result to the expectation of a random vector.

Definition.

The expectation of the random vector $E(\mathbf{x})$ is just the vector that collecting all the expectation of marginal (individual) random variables, i.e.

$$E(\mathbf{x}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_k) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} = \boldsymbol{\mu} \qquad \blacksquare$$

5.2 Expectation of a Function of Random Variables

Let (X_1, X_2) be a bivariate random vector with $f_{\mathbf{x}}(x_1, x_2)$ their joint density function and let $h(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ be a Borel function. Define $Y = h(X_1, X_2)$ and consider its expectation. This can be defined in two equivalent ways:

(a).

$$E(Y) = \int_{-\infty}^{\infty} f_Y(y)dy;$$

(b).

$$E(Y) = E(h(X_1, X_2)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x_1, x_2)f(x_1, x_2)dx_1dx_2.$$

5.2.1 Forms of $h(X_1, X_2)$ of Particular Interest

A particular form of $h(X_1, X_2)$ we are interested is in the following.

Definition.

For $h(X_1, X_2) = (X_1 - E(X_1))^l(X_2 - E(X_2))^k$, where

$$\mu_1 = E(X_1) \quad and \quad \mu_2 = E(X_2).$$

Then

$$\begin{aligned} \mu_{lk} \equiv E(h(X_1, X_2)) &= E\left[(X_1 - E(X_1))^l(X_2 - E(X_2))^k\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(X_1 - \mu_1)^l(X_2 - \mu_2)^k f(x_1, x_2)dx_1dx_2 \end{aligned}$$

are called *joint central moment* of order $l+k$. ■

Two especially interesting joint central moments are the covariance and variance:

(a) *Covariance*: when $l=k=1$,

$$Cov(X_1,X_2)=E((X_1-\mu_1)(X_2-\mu_2))=E(X_1X_2)-E(X_1)\cdot E(X_2).$$

(b) *Variance*: when $l=2,k=0$ or $l=0,k=2$,

$$\begin{aligned} Var(X_1) &= E(X_1-\mu_1)^2, \text{ or} \\ Var(X_2) &= E(X_2-\mu_2)^2. \end{aligned}$$

Definition.

Using definition of covariance and variance we could define the correlation coefficient by

$$Corr(X_1,X_2)=\frac{Cov(X_1,X_2)}{\sqrt{[Var(X_1)\cdot Var(X_2)]}},$$

which has the properties that $-1\le Corr(X_1,X_2)\le 1$. ■

Theorem.

If X_1 and X_2 are independent then $Cov(X_1,X_2)=0$, and the converse is not true. ■

Result.

For a linear function $\sum_i a_iX_i$ the variance is of the form

$$Var\left(\sum_{i=1}^{n}a_iX_i\right)=\sum_{i=1}^{n}a_i^2Var(X_i)+\sum_{i\neq j}\sum a_ia_jCov(X_iX_j),$$

where a_i are real constant. ■

5.2.2 Properties of Expectation

(a). *Linearity*: $E[ah_1(X_1, X_2) + bh_2(X_1, X_2)] = aE(h_1(X_1, X_2)) + bE(h_2(X_1, X_2))$, where a and b are constant and $h_1(\cdot)$, $h_2(\cdot)$ are Borel functions from $\mathbb{R}^2$ to $\mathbb{R}$. In particular $E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i)$.

(b). If X_1 and X_2 are independent r.v.'s, for every Borel function $h_1(\cdot)$, $h_2(\cdot)$: $\mathbb{R} \to \mathbb{R}$,

$$E(h_1(X_1)h_2(X_2)) = E(h_1(X_1)) \cdot E(h_2(X_2)),$$

given that the above expectations exist.

5.3 Conditional Expectation

Conditional expectation is one of the most useful concepts in probability. It plays a very important role in extending the probability to time dependent random variables and linear regression models.

Definition.

The conditional expectation of X_1 given that X_2 takes a particular value $x_2 (X_2 = x_2)$ is defined by

$$E(X_1|X_2 = x_2) = \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1, x_2) dx_1,$$

and is a function of x_2. In general for any Borel function $h(\cdot)$ whose expectation exist

$$E(h(X_1)|X_2 = x_2) = \int_{-\infty}^{\infty} h(x_1) f_{X_1|X_2}(x_1, x_2) dx_1. \qquad \blacksquare$$

Example.

Let X and Y have joint density function $f(x, y) = y/10$, $(x, y) = (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)$. Then X has marginal density $f_X(1) = \frac{6}{10}$, $f_X(2) = \frac{3}{10}$, and $f_X(3) = \frac{1}{10}$. Therefore the conditional density function of Y given $x = 1$ is $f_{Y|X}(1|1) =$

$\frac{1}{6}$, $f_{Y|X}(2|1) = \frac{2}{6}$, $f_{Y|X}(3|1) = \frac{3}{6}$, and conditional density function of Y given $x = 2$ is $f_{Y|X}(1|2) = \frac{1}{3}$, $f_{Y|X}(2|2) = \frac{2}{3}$. Therefore the condition expectation of Y given $x = 1$ is

$$\begin{aligned} E(Y|x=1) &= 1 \cdot f_{Y|X}(1|1) + 2 \cdot f_{Y|X}(2|1) + 3 \cdot f_{Y|X}(3|1) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{2}{6} + 3 \cdot \frac{3}{6} \\ &= \frac{7}{3}, \end{aligned}$$

and the condition expectation of Y given $x = 2$ is

$$\begin{aligned} E(Y|x=1) &= 1 \cdot f_{Y|X}(1|2) + 2 \cdot f_{Y|X}(2|2) \\ &= 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} \\ &= \frac{5}{3}. \end{aligned}$$

■

5.3.1 Properties of the Conditional Expectation

Let X, X_1, and X_2 be random variables on $(\mathcal{S}, \mathcal{F}, \mathcal{P})$, then we have the following properties of conditional expectation.

(a). $E[a_1h(X_1) + a_2h(X_2)|X = x] = a_1E[h(X_1)|X = x] + a_2E[h(X_2)|X = x]$, a_1, a_2 is constants.

(b). If $X_1 \geq X_2$, $E(X_1|X = x) \geq E(X_2|X = x)$.

(c). $E[h(X_1, X_2)|X_2 = x_2] = E[h(X_1, x_2)|X_2 = x_2]$.

(d).** $E[h(X_1)|X_2 = x_2] = E[h(X_1)]$ if X_1 and X_2 are independent.

(e). $E[h(X_1)] = E_{X_2}\{E[h(X_1)|X_2 = x_2]\}$, this is so called *law of iterated expectation.*

(f). The conditional expectation $E(X_1|X_2 = x_2)$ is a non-stochastic function of x_2, i.e. $E(X_1|\cdot) : \mathbb{R}_{X_2} \to \mathbb{R}$. The graph $(x_2, E(X_1|X_2 = x_2))$ is called the regression curve.

(g). $E[h(X_1) \cdot g(X_2)|X_2 = x_2] = g(x_2)E[h(X_1)|X_2 = x_2]$.

5.3.2 Higher Conditional Moments

As is the case of ordinary expectation, we can define higher conditional moments.

Definition.

(a). *Raw conditional moments*:

$$E(X_1^r|X_2 = x_2) = \int_{-\infty}^{\infty} x_1^r f_{X_1|X_2}(x_1, x_2)dx_1, \quad r \geq 1,$$

(b). *Central conditional moments* :

$$E[(X_1 - E(X_1|X_2 = x_2))^r|X_2 = x_2], \quad r \geq 2.$$ ■

5.4 Conditional Variance

Of particular interest of the higher conditional moments in last section is the conditional variance, sometimes called *skedasticity*.

A conditional variance is the variance of the *conditional distribution*:

$$\begin{aligned} Var(Y|x) &= E\{[Y - E(Y|x)]^2|x\} \\ &= \int_y [Y - E(Y|x)]^2 f(y|x)dy, \quad if\ y\ is\ continuous \\ &= \sum_y [Y - E(Y|x)]^2 f(y|x), \quad if\ y\ is\ discrete. \end{aligned}$$

The computation can be simplified by using

$$\begin{aligned} Var(Y|x) &= \{[Y - E(Y|x)]^2|x\} \\ &= E\{Y^2 - 2Y \cdot E(Y|x) + [E(Y|x)]^2|x\} \\ &= E(Y^2|x) - 2[E(Y|x)]^2 + [E(Y|x)]^2 \\ &\quad (Since\ \ E\{2Y \cdot E(Y|x)|x\} = 2[E(Y|x)]^2) \\ &= E(Y^2|x) - [E(Y|x)]^2. \end{aligned}$$

Theorem. (Decomposition of Variance):

$$Var(Y) = Var_x[E(Y|x)] + E_x[Var(Y|x)].$$

Proof.
From definition

$$\begin{aligned} Var_x[E(Y|x)] &= E_x\{E(Y|x) - E_x[E(Y|x)]\}^2 \\ &= E_x\{E(Y|x) - E(Y)\}^2 \quad (Since\ E_x[E(Y|x)] = E(Y)) \\ &= E_x\{[E(Y|x)]^2 - 2E(Y|x)E(Y) + [E(Y)]^2 \\ &= E_x[E(Y|x)]^2 - 2E_X E(Y|x)E(Y) + E_x[E(Y)]^2 \\ &= E_x[E(Y|x)]^2 - [E(Y)]^2, \end{aligned}$$

and

$$\begin{aligned} E_x[Var(Y|x)] &= E_x\{E(Y^2|x) - [E(Y|x)]^2\} \\ &= E(Y^2) - E_x[E(Y|x)]^2. \end{aligned}$$

Therefore

$$\begin{aligned} Var(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \{E_x[E(Y|x)]^2 - [E(Y)]^2\} + \{E(Y^2) - E_x[E(Y|x)]^2\} \\ &= Var_x[E(Y|x)] + E_x[Var(Y|x)]. \end{aligned}$$

■

Exercise 7.
Show that in a bivariate normal distribution, $Var(X_1|X_2 = x_2) = \sigma_1^2(1-\rho^2)$. That is the conditional variance is free of the conditional variables—*homoskedastic*. ■

6 Function of Random Variables

One of the most important problems in probability theory and statistical inference is to derive the distribution of a function $h(X_1, X_2, ..., X_k)$ when the distribution of the random vector $\mathbf{x} = (X_1, X_2, ..., X_k)$ is known. In cases such as these it is inefficient to compute the *pdf* of the "new" random variables by returning to elementary principles. Easier methods are available. This problem is important for at least two reasons:

(a). it is often the case that in modeling observable phenomena we are primarily interested in function of random variables; and

(b). in statistical inference the quantities of primary interest are commonly functions of random variables.

It is no exaggeration to say that the whole of statistical inference is based on our ability to derive the distribution of various functions of r.v.'s.

6.1 (Single) Function of One Random Variable ("One ⇒ One" Transformation)

We now derive a general method which is often helpful for deriving the density function of the transformed random variable without first finding the distribution function.

Let X be a random variable on the probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$. By definition, $X(\cdot) : \mathcal{S} \to \mathbb{R}$ is a real valued function. Suppose that $h(\cdot) : \mathbb{R} \to \mathbb{R}$, where h is a continuous function with at most a countable number of discontinuities. More formally we need $h(\cdot)$ to be a Borel function.

Definition.
A function $h(\cdot) : \mathbb{R}_x \to \mathbb{R}$ is said to be a *Borel function* if any $a \in \mathbb{R}$ and $x \in \mathbb{R}_x$, the set $B_h = \{h(x) \leq a\}$ is a Borel set, i.e. $B_h \in \mathcal{B}$, where $\mathcal{B}$ is the Borel field on $\mathbb{R}$. ■

The above definition is to require that $h(\cdot)$ is a Borel function that is an obvious condition to impose given that we need $h(X)$ to be a random variable itself.

Having ensured that the function $h(\cdot)$ of the r.v. X is itself a r.v. $Y = h(X)$, we want to derive the distribution of Y when the distribution of X is known. In general, the distribution of Y is defined as

$$F(y) = P(s : Y(s) \leq y) = P(s : X(s) \in h^{-1}((-\infty, y])).$$

Theorem.
Let X be a continuous r.v. and $Y = h(X)$ where $h(X)$ is differentiable for all $x \in \mathbb{R}_x$ and $[dh(x)]/(dx) > 0$ or $[dh(x)]/(dx) < 0$ for all x. Then the density function of Y is given by

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right| \quad for\ a < y < b,$$

where $|\ |$ stands for the absolute value and a and b refer to the smallest and biggest value y can take, respectively. ■

Example.
Let $X \sim (N(\mu, \sigma^2)$ and $Y = (X - \mu)/\sigma$, which implies that $[dh(x)]/(dx) = 1/\sigma > 0$ for all $x \in \mathbb{R}$ since $\sigma > 0$ by definition; $h^{-1}(y) = \sigma y + \mu$ and $[dh^{-1}(y)]/(dy) = \sigma$. Thus since

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

therefore,

$$\begin{aligned} f_Y(y) &= \frac{1}{\sigma\sqrt{2\pi}} exp\left\{ -\frac{1}{2} \left(\frac{\sigma y + \mu - \mu}{\sigma} \right)^2 \right\} \cdot (\sigma) \\ &= \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2} y^2 \right), \end{aligned}$$

i.e. $Y \sim N(0, 1)$ the standard normal distribution. ■

In cases where the conditions of the theorem above are not satisfied we need to derive the distribution from the relationship

$$F_Y(y) = Pr(h(x) \leq y) = Pr(X \in h^{-1}((-\infty, x])).$$

Exercise 4.
Let $X \sim N(\mu, \sigma^2)$. Find pdf of Y, where $Y = X^2$. ■

6.2 (Single) Function of Several Random Variables ("k ⇒ one" Transformation)

As in the case of a simple r.v., for a Borel function $h(\cdot) : \mathbb{R}^n \to \mathbb{R}$ and a random vector $\mathbf{x} = (X_1, X_2, ..., X_n)$, $h(\mathbf{x})$ is a random variable. Three commonly used functions of random variables (take two random variables as example) are:
(a). The distribution of $X_1 + X_2$,
(b). The distribution of X_1/X_2,
(c). The distribution of Y=min(X_1, X_2).

Exercise 5.
Let $X_i \sim U(-1, 1), i = 1, 2, 3$ and $Y = X_1 + X_2 + X_3$. Find pdf of Y. ■

6.3 Functions of Several Random Variables ("k ⇒ k" Transformation)

After considering various simple functions of r.v.'s separately, let us consider them together.

Theorem.
Let $\mathbf{x} = (X_1, X_2, ..., X_k)'$ be a random vector with a joint probability density function

$f_{\mathbf{x}}(x_1, x_2, ..., x_n)$ and define the *n-to-n* transformation:

$$\begin{aligned} Y_1 &= h_1(X_1, X_2, ..., X_k) \\ Y_2 &= h_2(X_1, X_2, ..., X_k) \\ &\vdots \\ Y_k &= h_k(X_1, X_2, ..., X_k), \end{aligned}$$

whose inverse take the form $h_i^{-1}(\cdot) = g_i(\cdot),\ i = 1, 2, ..., n$, that is,

$$\begin{aligned} X_1 &= g_1(Y_1, Y_2, ..., Y_n) \\ X_2 &= g_2(Y_1, Y_2, ..., Y_n) \\ &\vdots \\ X_n &= g_n(Y_1, Y_2, ..., Y_n). \end{aligned}$$

Assume:
(a) $h_i(\cdot)$ and $g_i(\cdot)$ are continuous;
(b) the partial derivatives $\partial X_i/\partial Y_i$, $i, j = 1, 2, ..., k$ exist and are continuous; and
(c) the Jocobian of the inverse transformation

$$\mathbf{J} = det\left(\frac{\partial(X_1, X_2, ..., X_k)'}{\partial(Y_1, Y_2, ..., Y_k)}\right) = det\left(\frac{\partial(h_1^{-1}(\mathbf{y}), h_2^{-1}(\mathbf{y}), ..., h_n^{-1}(\mathbf{y}))'}{\partial(Y_1, Y_2, ...Y_k)}\right) \neq 0.$$

Then

$$f(y_1, y_2, ..., y_k) = f(g_1(y_1, y_2, ..., y_k)), ..., g_n(y_1, y_2, ..., y_k)) \cdot |\mathbf{J}|\,. \qquad \blacksquare$$

Example.
Let $\mathbf{x} = (X_1, X_2)'$, where X_1 and X_2 are independent random variables that have the standard normal distribution. Here, the density of $\mathbf{x}$ is the product of the density function of X_1 and X_2. Thus

$$f_{\mathbf{x}}(x_1, x_2) = \frac{1}{2\pi}\exp\left[-\frac{1}{2}(x_1^2 + x_2^2)\right], \quad -\infty < x_1, x_2 < \infty.$$

Let $\mathbf{y} = [Y_1, Y_2]'$ be defined as $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - 2X_2$. In this case, $h_1(\mathbf{x}) = X_1 + X_2 = Y_1$, $h_2(\mathbf{x}) = X_1 - 2X_2 = Y_2$, $h_1^{-1}(\mathbf{y}) = 2Y_1 + Y_2 = X_1$, $h_2^{-1}(\mathbf{y}) = \frac{1}{3}(Y_1 - Y_2) = X_2$, and

$$\mathbf{J} = det\left[\frac{\partial(X_1, X_2)'}{\partial(Y_1, Y_2)}\right] = det\left[\frac{\partial(h_1^{-1}(\mathbf{y}), h_2^{-1}(\mathbf{y}))'}{\partial(Y_1, Y_2)}\right] = det\begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} \end{bmatrix} = -\frac{1}{3}.$$

Hence the density of $\mathbf{y}$ is

$$\begin{aligned} f_{\mathbf{y}}(y_1, y_2) &= \frac{1}{2\pi} \exp\left[-\frac{1}{2}\left(\frac{2y_1 + y_2}{3}\right)^2 - \frac{1}{2}\left(\frac{y_1 - y_2}{3}\right)^2\right] \times \frac{1}{3} \\ &= \frac{1}{6\pi} \exp\left[-\frac{1}{18}(5y_1^2 + 2y_1y_2 + 2y_2^2)\right], \quad -\infty < y_1, y_2 < \infty. \end{aligned}$$ ■

Exercise 6.

Let $X_i \sim N(0,1),\ i = 1, 2$ be two independent r.v.'s and $Y_1 = h_1(X_1, X_2) = X_1 + X_2$, $Y_2 = h_2(X_1, X_2) = \frac{X_1}{X_2}$. Find joint pdf of $f(Y_1, Y_2)$ and marginal density of $f_1(y_1)$ and $f_2(y_2)$. ■

6.4 Functions of Normally Distributed Random Variables**

The above example on functions of random variables show clearly that deriving the distribution of $h(X_1, ..., X_n)$ when $f(x_1, ..., x_n)$ is known is not an easy exercise. An important generalization of the results involving normal random variables will be summarized below for reference.

6.4.1 Univariate Normal

Definition.

We say that $X \sim N(\mu, \sigma^2)$, if the probability density of Y is

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]. \tag{2-3}$$

■

Results.

(a). $Z = \left[\frac{1}{\sigma} \cdot (X - \mu)\right] \sim N(0,1)$.–the standard normal distribution.

(b). If $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, ..., k$ are independent r.v.'s then

$$\left(\sum_{i=1}^{k} X_i\right) \sim N\left(\sum_{i=1}^{k} \mu_i,\ \sum_{i=1}^{n} \sigma_i^2\right). -- \textit{Reproductive Property}.$$

■

6.4.2 Chi-Square Distribution

Definition.

We say that $Y \sim \chi^2(k)$ if the density function of Y is

$$f_Y(y;k) = \frac{1}{2^{(k/2)}\Gamma(k/2)} y^{(k/2)-1} e^{-(y/2)}, \quad y > 0, \quad k = 1, 2, ..$$ ■

It is easy to see that $E(Y) = k$ and $Var(Y) = 2k$.

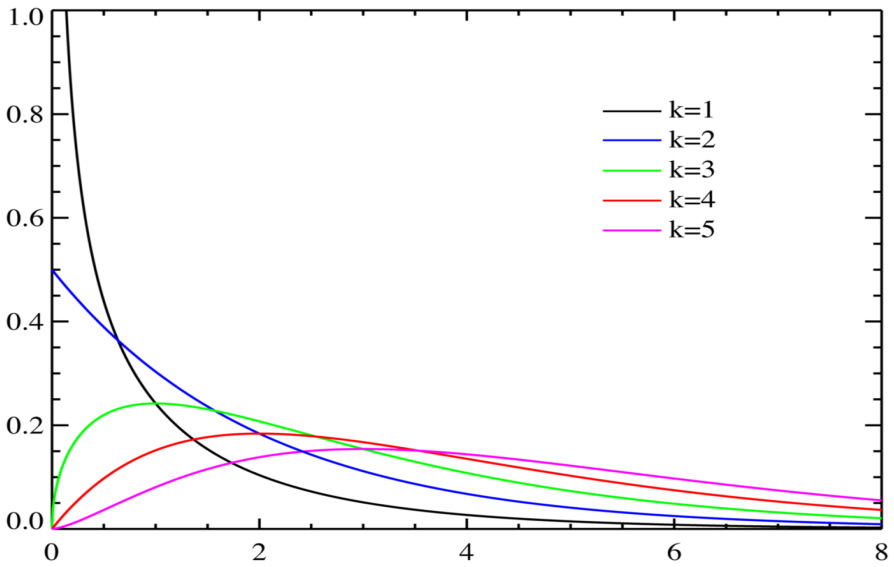

Figure (2-4). $\chi^2(k)$ Distribution

Results.

(a). If $X_i \sim N(0,1)$, $i = 1, 2, ..., k$ are independent r.v.'s then

$$\left(\sum_{i=1}^{k} X_i^2\right) \sim \chi^2(k) - -\text{chi-square with k degree of freedom.}$$

(b). If $Y_1, Y_2, ..., Y_k$ are independent r.v.'s $Y_i \sim \chi^2(k_i),\ i = 1, 2, ..., k$, then

$$\left(\sum_{i=1}^{k} Y_i\right) \sim \chi^2(k_1 + k_2 + \cdots + k_k). \qquad \blacksquare$$

6.4.3 Student t-Distribution

Definition.

We say that $W \sim t(k)$ if the density function of W is

$$f_W(w;k) = \frac{1}{\sqrt{(k\pi)}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\left(1 + \frac{w^2}{k}\right)^{[(k+1)/2]}}\ k > 0,\ w \in \mathbb{R}. \qquad \blacksquare$$

It is easy to see that

$$E(W) = 0, \quad Var(W) = \frac{k}{k-2},\ k > 2,\ \mu_4' = 3 + \frac{6}{k-4},\ k \geq 4.$$

These moments show that for a large n the t-distribution is very close to the standard normal.

Result.

If $X_1 \sim N(0,1)$ and $X_2 \sim \chi^2(k)$ are independent r.v.'s then

$$\frac{X_1}{\sqrt{(X_2/k)}} \sim t(k). \qquad \blacksquare$$

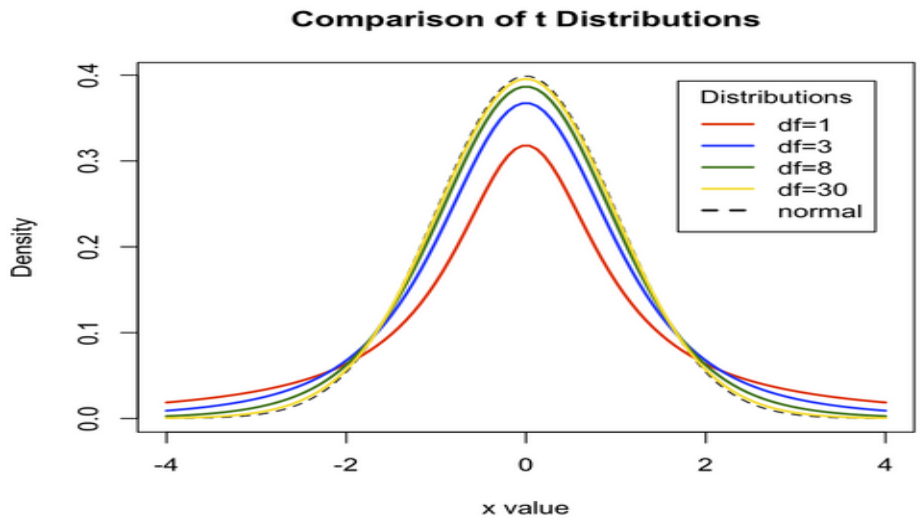

Figure (2-5). $t(k)$ Distribution

6.4.4 F-Distribution

Definition.

We say that $U \sim F(n_1, n_2)$ if the density function of U is

$$f_U(u; n_1, n_2) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\Gamma\left(\frac{n_1}{2}\right)\cdot\Gamma\left(\frac{n_2}{2}\right)} \frac{u^{\frac{1}{2}(n_1-2)}}{\left(1+\frac{n_1}{n_2}u\right)^{\frac{1}{2}(n_1+n_2)}}, \quad u > 0. \qquad \blacksquare$$

It is easy to see that

$$E(U) = \frac{n_2}{n_2-2}, \quad n > 2, \quad Var(U) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \quad n_2 > 4.$$

Result.

If $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$ are independent r.v.'s then

$$\frac{(X_1/n_1)}{(X_2/n_2)} \sim F(n_1, n_2). \qquad \blacksquare$$

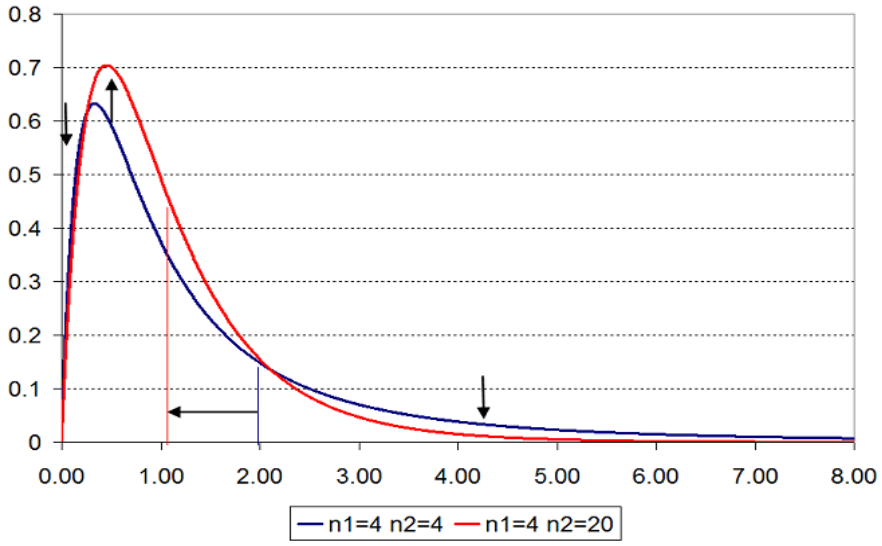

Figure (2-6). $F(n_1, n_2)$ Distribution

7 Generating Functions

7.1 The Moment-Generating Function

Finding moments of a random variable, particularly the higher moments is conceptually straightforward but can be difficult to accomplish in practice. Fortunately, an alternative method is available. For some densities, we can find a *moment-generating function*. Moment-generating functions can also be extremely useful in deriving the distribution of a *sum* of independent random variables. Such problems are important in statistics (for example, estimators from a linear function of the sample) and typically difficult.

Definition. (Moment-Generating Function):
Let X be a continuous random variable with a cumulative distribution function $F_X(x)$. We define the kth raw moment of X by $E(X^k)$, and the moment-generating function of X by

$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} dF_X(x) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx, \tag{2-4}$$

where t is a scalar. The convergence of the integral in (3) depends on the choice of the scalar t.[13] ■

In applied mathematics courses, the moment-generating function is called the Laplace transform of the density function of X. The moment-generating function *does not* have any obvious meaning by itself, but we shall see that it is very useful for doing distribution theory. The most basic property of the moment-generating function is that it generates the moments.

Theorem.
Let $M^{(k)}(t) = \frac{d^k}{dt^k} M(t)$. If $M(t)$ exists, then $E(X^k)$ is finite for all k, and

$$E(X) = M'(0), \quad E(X^2) = M''(0),$$

and in general,

$$E(X^k) = M^{(k)}(0). \qquad \blacksquare$$

[13] See Khuri, p.251 for example.

Proof.
We assume that X is a continuous random variable with range S. Note that

$$\frac{\partial^k}{\partial t^k} e^{tx} = x^k e^{tx}.$$

Using the method of *differentation under the integral sign*[14] we have

$$M^{(k)}(t) = \frac{d^k}{dt^k} \int_S e^{tx} f(x) dx = \int_S \frac{\partial^k}{\partial t^k} (e^{tx} f(x)) dx = \int_S x^k e^{tx} f(x) dx = E(X^k e^{tx}).$$

Therefore,

$$M^{(k)}(0) = E(X^k e^0) = E(X^k). \qquad \blacksquare$$

To generate moments is just *only one of many uses* we shall have for moment-generating functions. Most of the other uses depend on the following important two theorems.

Theorem. (Uniqueness Theorem):
Let X and Y be random variables that have the same moment-generating functions. Then X and Y have the same distribution.

Proof.
See Arnold p. 115. $\blacksquare$

This theorem implies that a random variable's moment generating function completely determines its distribution. We can now describe the distribution of a random variable in one of four ways: by defining $P(X \in A)$ for all A, by giving the distribution function of X, by giving the density of X, or by giving the moment-generating function of X.

Theorem.

[14] See Khuri, p.301.

(a). Let X be a random variable with moment-generating function $M_X(t)$. Let $U = aX + b$ for constant a and b. Then U has moment-generating function

$$M_U(t) = e^{bt} M_X(at).$$

(b). Let $X_1, ..., X_n$ be independent random variables such that X_i has moment-generating unction $M_i(t)$. Let $V = \sum_i X_i$ and $W = \sum_i a_i X_i + b$. Then V and W have moment-generating functions

$$M_V(t) = \prod_i M_i(t) \quad and \quad M_W(t) = e^{bt} \prod_i M_i(a_i t).$$

(c). Let $X_1, ..., X_n$ be independent random variables with the same distribution and with common moment-generating unction $M(t)$. Then $V = \sum_i X_i$ has moment-generating function

$$M_V(t) = [M(t)]^n.$$

Proof.

(a). $M_U(t) = E(e^{tU}) = E(e^{t(aX+b)}) = E(e^{taX+tb}) = e^{tb}E(e^{(at)X}) = e^{tb}M_X(at)$.

(b). V is a special case of W. So we have

$$\begin{aligned} M_W(t) &= E(e^{tW}) = E\left(e^{t(\sum a_i X_i + b)}\right) = e^{tb} E\left(\prod^{ta_i X_i}\right) \\ &= e^{tb} \prod E(e^{(a_i)X_i}) = e^{tb} \prod M_i(a_i t). \end{aligned}$$

(c). This result follows directly from part (b). ■

From the foregoing theorem, we see that it is often easy to find the moment-generating function of a linear combination W from the individual moment-generating functions. If we can recognize the moment-generating function, then we have found the distribution of W, because of the *uniqueness theorem*. As we shall see in chapter 4, when the moment-generating function approach to finding an *induced* distribution works, it is the easiest method to use.

Finally, a function that is typically easier to differentiate that $M(t)$ is called the *cumulant-generating function*

Definition. (Cumulant-Generating Function)
Let X be a continuous random variable with a moment generating function $M(t)$. The cumulant-generating function $\varphi(t)$ of X is defined as

$$\varphi(t) = \ln(M(t)). \qquad \blacksquare$$

Theorem.
The mean and variance of X are respectively given by

$$\mu = \varphi'(0) \ \ and \ \ \sigma^2 = \varphi''(0).$$

Proof.
The first derivative $\varphi'(t) = M'(t)/M(t)$, and the second derivative $\varphi''(t) = [M''(t)M(t) - (M'(t))^2]/(M(t))^2$. Now, $M(0) = 1$, $M'(0) = \mu$, and $M''(0) = E(X^2)$. Therefore, $\varphi'(0) = \mu/1 = \mu$, and $\varphi''(0) = [E(X^2) \cdot 1 - \mu^2]/1^2 = E(X^2) - \mu^2 = \sigma^2$. $\blacksquare$

7.2 The Joint Moment-Generating Function

The moment-generating function of a random variable can be extended to the joint moment-generating function for a random variables.

Definition. (Joint Moment-Generating Function):
Let $\mathbf{x} = (X_1, ..., X_n)'$ be a random vector. Let $\mathbf{t} = (t_1, ..., t_n)'$. The joint moment-generating function of $\mathbf{x}$ is

$$M_{1\times 1}(\mathbf{t}) = M(t_1, ..., t_n) = E\left(\exp\left[\sum_i t_i X_i\right]\right) = E(\exp(\mathbf{t}'\mathbf{x})). \qquad \blacksquare$$

We first state how the moment-generating function generates the moment of $\mathbf{x}$.

Theorem.
Let

$$M_i(\mathbf{t}) = \frac{\partial}{\partial t_i} M(\mathbf{t}), \quad M_{ii}(\mathbf{t}) = \frac{\partial^2}{\partial t_i^2} M(\mathbf{t}), \quad M_{ij}(\mathbf{t}) = \frac{\partial^2}{\partial t_i \partial t_j} M(\mathbf{t}).$$

If the moment-generating function $M(\mathbf{t})$ of $\mathbf{x}$ exists, then

$$E(X_i) = M_i(\mathbf{0}), \quad E(X_i^2) = M_{ii}(\mathbf{0}), \text{ and } E(X_i X_j) = M_{ij}(\mathbf{0}).$$

Proof.
Note that

$$\frac{\partial}{\partial t_i} \exp\left(\sum_i X_i t_i\right) = X_i \exp\left(\sum_i t_i X_i\right) = X_i \exp(\mathbf{t}'\mathbf{x}).$$

Using the method of *differentation under the integral sign*[15] we have

$$M_i(\mathbf{t}) = \frac{\partial}{\partial t_i} \int_S e^{\mathbf{t}'\mathbf{x}} f(\mathbf{x}) d\mathbf{x} = \int_S \frac{\partial}{\partial t_i} (e^{\mathbf{t}'\mathbf{x}} f(\mathbf{x})) d\mathbf{x} = \int_S X_i e^{\mathbf{t}'\mathbf{x}} f(\mathbf{x}) d\mathbf{x} = E(X_i e^{\mathbf{t}'\mathbf{x}}).$$

Therefore,

$$M_i(\mathbf{0}) = E(X_i e^{\mathbf{0}}) = E(X_i).$$

Similarly,

$$\frac{\partial^2}{\partial t_i \partial t_j} \exp\left(\sum_i X_i t_i\right) = X_i X_j \exp\left(\sum_i t_i X_i\right) = X_i X_j \exp(\mathbf{t}'\mathbf{x}).$$

Using the method of *differentation under the integral sign*[16] we have

$$\begin{aligned} M_{ij}(\mathbf{t}) &= \frac{\partial^2}{\partial t_i \partial t_j} \int_S e^{\mathbf{t}'\mathbf{x}} f(\mathbf{x}) d\mathbf{x} = \int_S \frac{\partial^2}{\partial t_i \partial t_j} (e^{\mathbf{t}'\mathbf{x}} f(\mathbf{x})) d\mathbf{x} \\ &= \int_S X_i X_j e^{\mathbf{t}'\mathbf{x}} f(\mathbf{x}) d\mathbf{x} \\ &= E(X_i X_j e^{\mathbf{t}'\mathbf{x}}). \end{aligned}$$

Therefore,

$$M_{ij}(\mathbf{0}) = E(X_i X_j e^{\mathbf{0}}) = E(X_i X_j). \qquad \blacksquare$$

[15] See Khuri, p.301.
[16] See Khuri, p.301

The following definition is a direct extension of cumulant-generating function of a single random variable.

Definition. (Joint Cumulant-Generating Function):
Let $\mathbf{x}$ be a continuous random variable with a moment generating function $M(\mathbf{t})$. The cumulant-generating function $\varphi(\mathbf{t})$ of $\mathbf{X}$ is defined as

$$\varphi(\mathbf{t}) = \ln(M(\mathbf{t})). \quad \blacksquare$$

Theorem.
Let $\varphi(\mathbf{t}) = \ln(M(\mathbf{t}))$. Then

$$E(X_i) = \varphi_i(\mathbf{0}), \quad Var(X_i) = \varphi_{ii}(\mathbf{0}), \quad Cov(X_iX_j) = \varphi_{ij}(\mathbf{0}),$$

where φ_{ij} are defined analogously to M_{ij}. $\blacksquare$

The following theorem states that a random vector's joint moment generating function completely determines its distribution.

Theorem. (Uniqueness Theorem):
Let $\mathbf{x}$ and $\mathbf{y}$ be n-dimensional random vectors. If $\mathbf{x}$ and $\mathbf{y}$ have the same joint moment-generating function, the $\mathbf{x}$ and $\mathbf{y}$ have the same distribution. $\blacksquare$

Theorem. (Marginal Moment-Generating Function):
Let $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$, where $\mathbf{y}$ and $\mathbf{z}$ are random vectors. Let $\mathbf{t} = (\mathbf{u}', \mathbf{v}')'$, where $\mathbf{u}$ and $\mathbf{v}$ have the same dimension as $\mathbf{y}$ and $\mathbf{z}$. Suppose that $\mathbf{x}$ has moment-generating function $M_{\mathbf{x}}(\mathbf{t})$. Then

(a). $\mathbf{y}$ and $\mathbf{z}$ have marginal moment-generating functions

$$M_{\mathbf{y}}(\mathbf{u}) = M_{\mathbf{x}}(\mathbf{u}, \mathbf{0}) \quad \textit{and} \quad M_{\mathbf{z}}(\mathbf{v}) = M_{\mathbf{x}}(\mathbf{0}, \mathbf{v}).$$

(b). $\mathbf{y}$ and $\mathbf{z}$ are independent if and only if

$$M_{\mathbf{x}}(\mathbf{u}, \mathbf{v}) = M_{\mathbf{y}}(\mathbf{u}) M_{\mathbf{z}}(\mathbf{v}).$$

Proof.

(a).

$$M_{\mathbf{x}}(\mathbf{t}) = M_{\mathbf{x}}[(\mathbf{u}', \mathbf{v}')'] = E[\exp(\mathbf{t}'\mathbf{x})] = E[\exp(\mathbf{u}'\mathbf{y} + \mathbf{v}'\mathbf{z})].$$

Therefore, by definition

$$M_{\mathbf{y}}(\mathbf{u}) = E[\exp(\mathbf{u}'\mathbf{y})] = E[\exp(\mathbf{u}'\mathbf{y} + \mathbf{v}'\mathbf{0})] = M_{\mathbf{x}}(\mathbf{u}, \mathbf{0}).$$

The proof for $M_{\mathbf{z}}(\mathbf{v})$ is similar.

(b). Suppose that $\mathbf{y}$ and $\mathbf{z}$ are independent. Then

$$\begin{aligned} M_{\mathbf{x}}[(\mathbf{u}', \mathbf{v}')'] &= E[\exp(\mathbf{u}'\mathbf{y} + \mathbf{v}'\mathbf{z})] = E\left[\exp(\mathbf{u}'\mathbf{y}) \cdot \exp(\mathbf{v}'\mathbf{z})\right] \\ &= \left[E(\exp(\mathbf{u}'\mathbf{y}))\right]\left[E(\exp(\mathbf{v}'\mathbf{z}))\right] \\ &= M_{\mathbf{y}}(\mathbf{u}) M_{\mathbf{z}}(\mathbf{v}). \end{aligned}$$

■

Part(a) of this theorem says that to find the marginal moment-generating function of a subset of $\mathbf{x}$, we just put zeros in the moment-generating function for the t's associated with the variables that are not in the subset. Note that to find the marginal density function from the joint density function, we have to integrate out the variables that we do not want. Obviously, it is considerably easier to substitute zeros into a function than to integrate out variable, so that we shall often use the joint moment-generating function to find marginal distribution.

Part (b) of this theorem says that $\mathbf{y}$ and $\mathbf{z}$ are independent if and only if the marginal moment-generating function factor into the product of the marginal moment generating functions. We have already seen that $\mathbf{y}$ and $\mathbf{z}$ are independent if and only if the probability function factors into the product of the marginal probability functions and that they are independent if and only if the joint density function factors into the

product of marginal density functions.

7.3 The Characteristic Function

We have seen that the moment generating function $M_X(t)$ for a random variable X can be used to obtain the moments of X. It may be recalled, however, that $M_X(t)$ may not be defined for all values of t. To generate all the moments of X, it is sufficient for $M_X(t)$ to be defined in a neighborhood of $t = 0$. Some well-known distributions do not have moment generating functions, such as the Cauchy distribution. Another function that generate the moments of a random variable in a manner similar to $M_X(t)$, but is *defined for all values of* t and for all random variables, is the characteristic function.

Definition. (Characteristic Function):
The characteristic function of a continuous random variable X, denoted by $C_X(t)$,[17] is

$$\begin{aligned} C_X(t) &= E[e^{itX}] \\ &= \int_{-\infty}^{\infty} e^{itx} f(x)dx \\ &= E[\cos(Xt)] + iE[\sin(Xt)], \end{aligned}$$

where $f(x)$ is the probability density function of X, and i is the complex number $\sqrt{-1}$. ■

Because

$$-1 \leq \cos(Xt) \leq 1, \quad -1 \leq \sin(Xt) \leq 1.$$

Therefore

$$\begin{aligned} \int_{-\infty}^{\infty} |\cos(Xt) f(x)dx \leq \int_{-\infty}^{\infty} f(x)dx &= 1 \\ \int_{-\infty}^{\infty} |\sin(Xt) f(x)dx \leq \int_{-\infty}^{\infty} f(x)dx &= 1, \end{aligned}$$

[17] By Demoivre's theorem, $e^{ia} = \cos(a) + i\sin(a)$.

that is, $E[\cos(Xt)]$ and $E[\sin(Xt)]$ are both finite, and hence $C_X(t)$ is finite for all t for any random variable X. In applied mathematics, the characteristic function is often called the **Fourier** transform of the density function.

The characteristic function and the moment generating function, when the latter exists, are related according to the formula

$$C_X(t) = M_X(it).$$

Furthermore, it can be shown that if X has finite moment, then they can be obtained by repeatedly differentiating $C_X(t)$ and evaluating the derivatives at zero, as is shown in the following theorem.

Theorem.

Let

$$C_X^{(k)}(t) = \frac{d^k}{dt^k} C_X(t),$$

then

$$E(X^k) = \frac{1}{i^k} C_X^{(k)}(0).$$

Proof.

We assume that X is a continuous random variable with range S. Note that

$$\frac{\partial^k}{\partial t^k} e^{itx} = i^k x^k e^{tx}.$$

Using the method of differentiation under the integral sign[18] we have

$$\begin{aligned} C_X^{(k)}(t) &= \frac{d^k}{dt^k} \int_S e^{itx} f(x)dx = \int_S \frac{\partial^k}{\partial t^k}(e^{itx} f(x))dx \\ &= \int_S i^k x^k e^{tx} f(x)dx = i^k E(X^k e^{tx}). \end{aligned}$$

Therefore,

$$\frac{1}{i^k} C_X^{(k)}(0) = E(X^k e^0) = E(X^k). \qquad \blacksquare$$

[18] See Khuri, p.301.

Theorem. (Uniqueness Theorem):
If X and Y have the same characteristic function, then they have the same distribution.■

Result.
Let Y and Z be independent. Then if $X = Y + Z$, then

$$C_X(t) = C_Y(t) \cdot C_Z(t). \tag{2-5}$$

Proof.

$$\begin{aligned} C_X(t) &= E(\exp(itX)) = E(\exp(it(Y+Z))) \\ &= E(\exp(itY)\exp(itZ)) \\ &= E((\exp(itY))E(\exp(itZ)) \end{aligned}$$

by independence. Hence $C_X(t) = C_Y(t) \cdot C_Z(t)$. ■

The definition of characteristic function can be extended to a random vector.

Definition.
The characteristic function of a random $k \times 1$ random vector $\mathbf{x}$, denoted by $C_{\mathbf{x}}(\boldsymbol{t})$, is

$$C_{\mathbf{x}}(\boldsymbol{t}) = E[\exp(i\boldsymbol{t}'\mathbf{x})],$$

where $\boldsymbol{t}$ is $k \times 1$ real vector. ■

Theorem.

(a). Let $X \sim N(\mu, \sigma^2)$. Then $C_X(t) = \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right)$.

(b). Let $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $C_{\mathbf{x}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}'\boldsymbol{\mu} - \frac{\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}}{2}\right)$. ■

Example. (The Characteric Function of the Standard Normal Distribution)
The characteristic function of the standard normal distribution with the density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty,$$

is

$$\begin{aligned} C_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{itx} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2 - 2itx)} dx \\ &= \frac{e^{-t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-it)^2} dx \\ &= e^{-t^2/2}. \end{aligned}$$

■

8 Multivariate Normal distribution

The multivariate normal distribution is by far the most important distribution in statistical inference for a variety of reasons including the fact that some of the statistics based on sampling from such a distribution have tractable distributions themselves. Before we consider the multivariate normal distribution, however, let us introduce some notations and various simple results related to random vectors and their distributions in general.

8.1 The First Two Moments of A Multivariate Distribution

Let $\mathbf{x} \equiv (X_1, X_2, ..., X_k)'$ be an $k \times 1$ random vector defined on the probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$. The mean vector $E(\mathbf{x})$ is defined by

$$E(\mathbf{x}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ . \\ . \\ . \\ E(X_k) \end{bmatrix} \equiv \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_k \end{bmatrix} \equiv \boldsymbol{\mu}, \ an\ k \times 1\ vector,$$

and the covariance matrix $Cov(\mathbf{x})$ by

$$\begin{aligned} Cov(\mathbf{x}) &= E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' \\ &= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & . & . & . & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & . & . & . & Cov(X_2, X_k) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ Cov(X_k, X_1) & Cov(X_k, X_2) & . & . & . & Var(X_k) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & . & . & . & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & . & . & . & \sigma_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{k1} & \sigma_{k2} & . & . & . & \sigma_k^2 \end{bmatrix} = E(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}' \equiv \boldsymbol{\Sigma}, \end{aligned}$$

where $\boldsymbol{\Sigma}$ is an $k \times k$ symmetric positive definite matrix.

By dividing σ_{ij} by $\sigma_i\sigma_j$, we obtain the *correlation matrix*:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & . & . & . & \rho_{1k} \\ \rho_{21} & 1 & . & . & . & \rho_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho_{k1} & \rho_{k2} & . & . & . & 1 \end{bmatrix},$$

where $\rho_{ij} = \sigma_{ij}/\sigma_i\sigma_j$.

Results.

If $\mathbf{x}$ has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then for $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b}$,

(a). $E(\mathbf{z}) = \mathbf{A}E(\mathbf{x}) + \mathbf{b} = \boldsymbol{A\mu} + \mathbf{b}$;

(b).

$$\begin{aligned} Cov(\mathbf{z}) &= E[(\mathbf{A}\mathbf{x} + \mathbf{b} - (\boldsymbol{A\mu} + \mathbf{b}))(\mathbf{A}\mathbf{x} + \mathbf{b} - (\boldsymbol{A\mu} + \mathbf{b}))'] \\ &= \mathbf{A}E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'. \end{aligned}$$ ■

8.2 The Multivariate Normal Distribution

Definition.

Let $\mathbf{x} \equiv (X_1, X_2, ..., X_k)'$ be an $k \times 1$ random vector with $E(\mathbf{x}) = \boldsymbol{\mu}$ and $Cov(\mathbf{x}) = \boldsymbol{\Sigma}$. If the joint density of $\mathbf{x}$ is in the form of

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-1/2)(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2-6}$$

then we say that $\mathbf{x}$ follows a multivariate normal distribution, denoted as $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.■

Exercise 8.

Let $\mathbf{R}$ be the correlation matrix of $\mathbf{x}$, shows that the density function of a multivariate normal $\mathbf{x}$ can also expressed as

$$f(\boldsymbol{x}) = (2\pi)^{-k/2}(\sigma_1\sigma_2 \cdots \sigma_k)^{-1}|\mathbf{R}|^{-1/2}\exp(-1/2)\boldsymbol{\epsilon}'\mathbf{R}^{-1}\boldsymbol{\epsilon}, \tag{2-7}$$

where $\epsilon_i = (x_i - \mu_i)/\sigma_i$, $i = 1, 2, ..., k$. ■

Three special cases of multivariate normal distribution are of interest.

(a). If all the variables are uncorrelated, then $\sigma_{ij} = 0$ for $i \neq j$. Thus $\mathbf{R} = \mathbf{I}$, and the density in (2-7) becomes

$$\begin{aligned} f(\boldsymbol{x}) &= f(x_1, x_2, ..., x_k) = (2\pi)^{-k/2}(\sigma_1\sigma_2 \cdots \sigma_k)^{-1} \exp(-1/2)\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \qquad (2\text{-}8) \\ &= f(x_1)f(x_2) \cdots f(x_k) = \prod_{i=1}^{k} f(x_i), \end{aligned}$$

where $f(x_i) \sim N(\mu_i, \sigma_i)$. That is, *if normally distributed variables are uncorrelated, then they are independent.*

(b). If $\boldsymbol{\mu} = \mathbf{0}$ and $\sigma_i^2 = \sigma^2$ then $X_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $\epsilon_i = x_i/\sigma$, and the density in (2-8) becomes

$$f(\boldsymbol{x}) = (2\pi)^{-k/2}(\sigma^2)^{-k/2} \exp(-1/2\sigma^2)\boldsymbol{x}'\boldsymbol{x}. \qquad (2\text{-}9)$$

(c). If $\sigma^2 = 1$ then (2-9) becomes

$$f(\boldsymbol{x}) = (2\pi)^{-k/2} \exp(-1/2)\boldsymbol{x}'\boldsymbol{x}. \qquad (2\text{-}10)$$

This is the multivariate standard normal distribution.

8.2.1 Marginal and Conditional Normal Distributions

In this section we find that the marginal and conditional distribution derived from a multivariate normal distribution are also normal.

Theorem.

Let $\mathbf{x}_1$ be any subset of the random vector $\mathbf{x}$, including a single variable, and let $\mathbf{x}_2$ be the remaining variables. Partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ likewise so that

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad and \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then

(a). If $(\mathbf{x}_1, \mathbf{x}_2)'$ have a joint multivariate normal distribution, then the marginal distribution are also normal, i.e.

$$\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

and

$$\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}).$$

(b). The conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2 = \boldsymbol{x}_2$ is normal as well:

$$\mathbf{x}_1|(\mathbf{x}_2 = \boldsymbol{x}_2) \sim N(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2})$$

where

$$\begin{aligned} \boldsymbol{\mu}_{1.2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \\ \boldsymbol{\Sigma}_{11.2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

■

8.2.2 Linear Function of a Normal Random Vector

Any linear function of a vector of joint normally distributed variables is also normally distributed.

Theorem.
If $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Ax} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

Proof.
Consider $\mathbf{y} = \mathbf{Ax}$, where $\mathbf{A}$ is $q \times k$. Let $\boldsymbol{t}$ be $q \times 1$, then

$$\begin{aligned} C_{\mathbf{y}}(\boldsymbol{t}) &= E(\exp(i\boldsymbol{t}'\mathbf{y})) = E[\exp(i\boldsymbol{t}'\mathbf{Ax})] \\ &= E[\exp(i\boldsymbol{\lambda}'\mathbf{x})] \\ &= C_{\mathbf{x}}(\boldsymbol{\lambda}), \end{aligned}$$

where $\boldsymbol{\lambda} = \mathbf{A}'\boldsymbol{t}$. Hence if $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$\begin{aligned} C_{\mathbf{y}}(\boldsymbol{t}) = C_{\mathbf{x}}(\boldsymbol{\lambda}) &= \exp\left(i\boldsymbol{\lambda}'\boldsymbol{\mu} - \frac{\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}}{2}\right) \\ &= \exp\left(i\boldsymbol{t}'\mathbf{A}\boldsymbol{\mu} - \frac{\boldsymbol{t}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\boldsymbol{t}}{2}\right), \end{aligned}$$

so that $\mathbf{y} = \mathbf{A}\mathbf{x} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. ■

Corollary.
If $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. ■

8.2.3 Quadratic Forms Related to the Normal Distribution

The multivariate normal vector in a quadratic form is very important in deriving the distribution of a test statistics for statistical inference in a later chapter. We collect two useful results here.

Theorem. (Distribution of an idempotent quadratic form)
Let $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, then for $\mathbf{A}$ an idempotent symmetric matrix, we have $(\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma})'\mathbf{A}(\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma}) \sim \chi^2(tr\ A)$.

Proof.
Let $\mathbf{z} = \frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma}$. It is easy to see that $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. Let $\mathbf{A}$ be arranged in a diagonal matrix $\boldsymbol{\Lambda}$ and an orthogonal matrix $\mathbf{C}$ such that $\mathbf{A} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$ and that $\mathbf{C}'\mathbf{C} = \mathbf{I}$. Then the quadratic form

$$q = \left(\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma}\right)' \mathbf{A} \left(\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma}\right) = \mathbf{z}'\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'\mathbf{z} = \mathbf{y}'\boldsymbol{\Lambda}\mathbf{y},$$

where $\mathbf{y} = \mathbf{C}'\mathbf{z}$. Because $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$, and $\mathbf{C}'\mathbf{C} = \mathbf{I}$. It follows that

$$\mathbf{y} = \mathbf{C}'\mathbf{z} \sim N(\mathbf{0}, \mathbf{C}'\mathbf{I}\mathbf{C}) \equiv N(\mathbf{0}, \mathbf{I}).$$

Hence the quadratic form

$$q = \mathbf{y}'\boldsymbol{\Lambda}\mathbf{y} = \sum_{i=1}^{k} \lambda_i y_i^2.$$

Because $\mathbf{A}$ is idempotent, λ_i is always 0 or 1, then

$$q = \sum_{i=1}^{J} y_i^2 \sim \chi^2(J),$$

where J is the trace of $\mathbf{A}$. ■

Example.
Let $X_i \sim i.i.d\ N(0,1)$, then $\sum_{i=1}^{k}(X_i - \bar{X})^2 = \mathbf{x}'\mathbf{M}_0\mathbf{x} \sim \chi^2(k-1)$, where $\mathbf{x} = [X_1, X_2, ..., X_k]'$. ■

Finally we will consider the general case $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We are interested in the distribution of

$$q = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix and hence is positive definite.

Theorem. (Distribution of a full rank quadratic form)
Let $\mathbf{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(k)$.

Proof.
Since $\boldsymbol{\Sigma}$ is positive definite, it has a square root. Define the symmetric matrix $\boldsymbol{\Sigma}^{1/2}$ so that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$. Then

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}.$$

It follows straightly that

$$\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I}) \equiv \mathbf{z}.$$

Therefore

$$\mathbf{z}'\mathbf{z} = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(k). \quad ■$$

8.2.4 Independence of Quadratic Form

The normal family of distribution (chi-squared, t and F distribution) can be derived as a functions of independent quadratic forms. Here we establish the condition for independence.

Theorem.
If $\mathbf{x} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{x}'\mathbf{B}\mathbf{x}$ are two idempotent quadratic form in $\mathbf{x}$, $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{x}'\mathbf{B}\mathbf{x}$ are independent if $\mathbf{AB} = \mathbf{0}$.

Proof.
Since $\mathbf{A}$ and $\mathbf{B}$ are both symmetric and idempotent, $\mathbf{A} = \mathbf{A}'\mathbf{A}$ and $\mathbf{B} = \mathbf{B}'\mathbf{B}$. The quadratic forms are therefore

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{x}_1'\mathbf{x}_1 \quad \textit{where} \;\; \mathbf{x}_1 = \mathbf{A}\mathbf{x}$$

and

$$\mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{x}'\mathbf{B}'\mathbf{B}\mathbf{x} = \mathbf{x}_2'\mathbf{x}_2 \quad \textit{where} \;\; \mathbf{x}_2 = \mathbf{B}\mathbf{x}.$$

Both vectors have zero mean vectors, so the covariance matrix of $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$E(\mathbf{x}_1\mathbf{x}_2') = \mathbf{A}\mathbf{I}\mathbf{B}' = \mathbf{A}\mathbf{B} = \mathbf{0}.$$

Since $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$ are linear functions of a normally distributed random vector, they are, in turn, normally distributed. Their zero covariance matrix implies that they are statistically independent using the fact that continuous functions of two independent random vector are also independent. ■

Example.
The F distribution is the ratio of two independent chi-squared variables, each divided by its respective degree of freedom. Let $\mathbf{A}$ and $\mathbf{B}$ be two idempotent matrices with rank r_A and r_B and let $\mathbf{AB} = \mathbf{0}$. Then

$$\frac{\mathbf{x}'\mathbf{A}\mathbf{x}/r_A}{\mathbf{x}'\mathbf{B}\mathbf{x}/r_B} \sim F[r_A, r_B]. \qquad ■$$

Theorem.
A linear function, $\mathbf{Lx}$, and an idempotent quadratic form $\mathbf{x}'\mathbf{Ax}$, in a standard normal vector are statistically independent if $\mathbf{LA} = \mathbf{0}$.

Example.
A particular case of t-ratio is

$$t(n-1) = \frac{\sqrt{n}\bar{X}}{s} = \frac{\mathbf{j}'\mathbf{x}}{s},$$

where $\mathbf{j} = \frac{1}{\sqrt{n}}\mathbf{i}$ and $s^2 = \frac{\mathbf{x}'\mathbf{M}_i\mathbf{x}}{n-1}$. It suffices to shows that $\mathbf{M}_0 \cdot \mathbf{j} = \mathbf{0}$.

Linton Hall, MSU.

End of this Chapter