# Chapter 1

# INTRODUCTION

## 1.1 Set Theory Digression

A **set** is defined as any collection of objects, which are called **points** or **elements**. The biggest possible collection of points under consideration is called the **space**, **universe**, or **universal set**. For Probability Theory the space is called the **sample space**.

A set $A$ is called a **subset** of $B$ (we write $A \subseteq B$ or $B \supseteq A$) if every element of $A$ is also an element of $B$. $A$ is called a **proper subset** of $B$ (we write $A \subset B$ or $B \supset A$) if every element of $A$ is also an element of $B$ and there is at least one element of $B$ which does not belong to $A$.

Two sets $A$ and $B$ are called **equivalent sets** or **equal sets** (we write $A = B$) if $A \subseteq B$ and $B \subseteq A$.

If a set has no points, it will be called the **empty** or **null** set and denoted by $\phi$.

The **complement** of a set $A$ with respect to the space $\Omega$, denoted by $\bar{A}$, $A^c$, or $\Omega - A$, is the set of all points that are in but not in $A$.

The **intersection** of two sets $A$ and $B$ is a set that consists of the common elements of the two sets and it is denoted by $A \cap B$ or $AB$.

The **union** of two sets $A$ and $B$ is a set that consists of all points that are in $A$ or $B$ or both (but only once) and it is denoted by $A \cup B$.

The **set difference** of two sets $A$ and $B$ is a set that consists of all points in

$A$ that are not in $B$ and it is denoted by $A - B$.

## Properties of Set Operations

**Commutative**: $A \cup B = B \cup A$ and $A \cap B = B \cap A$.

**Associative**: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$.

**Distributive**: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

$(A^c)^c = \overline{(\bar{A})} = A$ i.e. the complement of the $A$-complement is $A$.

If $A$ subset of $\Omega$ (the space) then: $A \cap \Omega = A$, $A \cup \Omega = \Omega$, $A \cap \phi = \phi$, $A \cup \phi = A$, $A \cap \overline{A} = \phi$, $A \cup \overline{A} = \Omega$, $A \cap A = A$, and $A \cup A = A$.

De Morgan Law: $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$, and $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$.

**Disjoint or mutually exclusive** sets are the sets that their intersection is the empty set, i.e. $A$ and $B$ are mutually exclusive if $A \cap B = \phi$. Subsets $A_1, A_2, ....$ are mutually exclusive if $A_i \cap A_j = \phi$ for any $i \neq j$.

Uncertainty or variability are prevalent in many situations and it is the purpose of the probability theory to understand and quantify this notion. The basic situation is an experiment whose outcome is unknown before it takes place e.g., a) coin tossing, b) throwing a die, c) choosing at random a number from $\mathbb{N}$, d) choosing at random a number from $(0, 1)$.

The **sample space** is the collection or totality of all possible outcomes of a conceptual experiment. An **event** is a subset of the sample space. The class of all events associated with a given experiment is defined to be the **event space**.

Let us describe the sample space $S$, i.e. the set of all possible relevant outcomes of the above experiments, e.g., $S = \{H, T\}, S = \{1, 2, 3, 4, 5, 6\}$. In both of these examples we have a finite sample space. In example c) the sample space is a countable infinity whereas in d) it is an uncountable infinity.

**Classical or a priori Probability**: If a random experiment can result in $N$ mutually exclusive and equally likely outcomes and if $N(A)$ of these outcomes have an attribute $A$, then the **probability** of $A$ is the fraction $N(A)/N$ i.e. $P(A) = N(A)/N$,

where $N = N(A) + N(\overline{A})$.

EXAMPLE: Consider the drawing an ace (event $A$) from a deck of 52 cards. What is $P(A)$?

We have that $N(A) = 4$ and $N(\overline{A}) = 48$. Then $N = N(A) + N(\overline{A}) = 4 + 48 = 52$ and $P(A) = \frac{N(A)}{N} = \frac{4}{52}$

**Frequency or a posteriori Probability**: Is the ratio of the number $\alpha$ that an event $A$ has occurred out of $n$ trials, i.e. $P(A) = \alpha/n$.

EXAMPLE: Assume that we flip a coin 1000 times and we observe 450 heads. Then the a posteriori probability is $P(A) = \alpha/n = 450/1000 = 0.45$ (this is also the relative frequency). Notice that the a priori probability is in this case 0.5.

**Subjective Probability**: This is based on intuition or judgment.

We shall be concerned with a priori probabilities. These probabilities involve, many times, the counting of possible outcomes.

### *1.1.1 Some Counting Problems*

Some more sophisticated discrete problems require counting techniques. For example:

a) What is the probability of getting four of a kind in a five card poker?

b) What is the probability that two people in a classroom have the same birthday?

The sample space in both cases, although discrete, can be quite large and it not feasible to write out all possible outcomes.

**1. Duplication is permissible and Order is important (Multiple Choice Arrangement)**, i.e. the element $AA$ is permitted and $AB$ is a different element from $BA$. In this case where we want to arrange $n$ objects in $x$ places the possible outcomes is given from: $M_x^n = n^x$.

EXAMPLE: Find all possible combinations of the letters A, B, C, and D when duplication is allowed and order is important.

The result according to the formula is: $n = 4$, and $x = 2$, consequently the

possible number of combinations is $M_2^4 = 4^2 = 16$. To find the result we can also use a tree diagram.

**2. Duplication is not permissible and Order is important (Permutation Arrangement)**, i.e. the element $AA$ is **not** permitted and $AB$ is a different element from $BA$. In this case where we want to permute $n$ objects in $x$ places the possible outcomes is given from:

$$P_x^n \quad or \quad P(n,x) = n \times (n-1) \times ..(n-x+1) = \frac{n!}{(n-x)!}.$$

EXAMPLE: Find all possible permutations of the letters A, B, C, and D when duplication is not allowed and order is important.

The result according to the formula is: $n = 4$, and $x = 2$, consequently the possible number of combinations is $P_2^4 = \frac{4!}{(4-2)!} = \frac{2*3*4}{2} = 12$.

**3. Duplication is not permissible and Order is not important (Combination Arrangement)**, i.e. the element $AA$ is **not** permitted and $AB$ is **not** a different element from $BA$. In this case where we want the combinations of $n$ objects in $x$ places the possible outcomes is given from:

$$C_x^n \quad or \quad C(n,x) = \frac{P(n,x)}{x!} = \frac{n!}{(n-x)!x!} = \binom{n}{x}$$

EXAMPLE: Find all possible combinations of the letters A, B, C, and D when duplication is not allowed and order is not important.

The result according to the formula is: $n = 4$, and $x = 2$, consequently the possible number of combinations is $C_2^4 = \frac{4!}{2!*(4-2)!} = \frac{2*3*4}{2*2} = 6$.

Let us now define probability rigorously.

### *1.1.2 Definition of Probability*

Consider a collection of sets $A_\alpha$ with index $\alpha \in \Gamma$, which is denoted by $\{A_\alpha : \alpha \in \Gamma\}$. We can define for an index $\Gamma$ of arbitrary cardinality (the cardinal number of a set is the number of elements of this set):

$$\bigcup_{\alpha\in\Gamma} A_\alpha = \{x \in S : x \in A_\alpha \text{ for some } \alpha \in \Gamma\}$$

$\underset{\alpha\in\Gamma}{\cap} A_\alpha = \{x \in S : x \in A_\alpha \text{ for all } \alpha \in \Gamma\}$

A collection is exhaustive if $\cup_{\alpha\in\Gamma} A_\alpha = S$ (partition), and is pairwise exclusive or disjoint if $A_\alpha \cap A_\beta = \phi$, $\alpha \neq \beta$.

To define probabilities we need some further structure. This is because in uncountable cases we can not just define probability for all subsets of $S$, as there are some sets on the real line whose probability can not be determined, i.e., they are unmeasurable. We shall define probability on a family of subsets of $S$, of which we require the following structure.

**Definition 1** *Let be $\mathcal{A}$ a non-empty class of subsets of $S$. $\mathcal{A}$ is an algebra if*

*1. $A^c \in \mathcal{A}$, whenever $A \in \mathcal{A}$*

*2. $A_1 \cup A_2 \in \mathcal{A}$, whenever $A_1, A_2 \in \mathcal{A}$.*

*$\mathcal{A}$ is a $\sigma$-algebra if also*

*2′. $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$, whenever $A_n \in \mathcal{A}$, n=1,2,3,...*

Note that since $\mathcal{A}$ is non-empty, (1) and (2) $\Rightarrow \phi \in \mathcal{A}$ and $S \in \mathcal{A}$. Note also that $\cap_{n=1}^{\infty} A_n \in \mathbf{A}$. The largest $\sigma$-algebra is the set of all subsets of $S$, denoted by $\mathcal{P}(S)$, and the smallest is $\{\phi, S\}$. We can generate a $\sigma$-algebra from any collection of subsets by adding to the set the complements and the unions of its elements. For example let $S = \mathbb{R}$, and

$$\mathcal{B} = \{[a,b], (a,b], [a,b), (a,b),\ a,b \in \mathbb{R}\},$$

and let $\mathcal{A} = \sigma(\mathcal{B})$ consists of all intervals and countable unions of intervals and complements thereof. This is called the Borel $\sigma$-algebra and is the usual $\sigma$-algebra we work when $S = \mathbb{R}$. The $\sigma$-algebra $\mathcal{A} \subset \mathcal{P}(\mathbb{R})$, i.e., there are sets in $\mathcal{P}(\mathbb{R})$ not in $\mathcal{A}$. These are some pretty nasty ones like the Cantor set. We can alternatively construct the Borel $\sigma$-algebra by considering $\mathcal{J}$ the set of all intervals of the form $(-\infty, x]$, $x \in \mathbb{R}$. We can prove that $\sigma(\mathcal{J}) = \sigma(\mathcal{B})$. We can now give the definition of probability measure which is due to Kolmogorov.

**Definition 2** *Given a sample space $S$ and a $\sigma$-algebra $(S, \mathcal{A})$, a probability measure is a mapping from $\mathcal{A} \longrightarrow \mathbb{R}$ such that*

1. $P(A) \geq 0$ *for all* $A \in \mathcal{A}$
2. $P(S) = 1$
3. *if* $A_1, A_2, ...$ *are pairwise disjoint, i.e.,* $A_i \cap A_j = \phi$ *for all* $i \neq j$*, then*

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

In such a way we have a probability space $(S, \mathcal{A}, P)$. When $S$ is discrete we usually take $\mathcal{A} = \mathcal{P}(S)$. When $S = \mathbb{R}$ or some subinterval thereof, we take $\mathcal{A} = \sigma(\mathcal{B})$.

$P$ is a matter of choice and will depend on the problem. In many discrete cases, the problem can usually be written such that outcomes are equally likely.

$$P(\{x\}) = 1/n, \quad n = \sharp(S).$$

In continuous cases, $P$ is usually like Lebesgue measure, i.e.,

$$P((a, b)) \propto b - a.$$

Properties of $P$

1. $P(\phi) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$
4. $P(B \cap A^c) = P(B) - P(B \cap A)$
5. If $A \subset B \Rightarrow P(A) \leq P(B)$
6. $P(B \cup A) = +P(A) + P(B) - P(A \cap B)$ More generally, for events $A_1, A_2, ...A_n \in \mathcal{A}$ we have:

$$P\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} P[A_i] - \sum\sum_{i<j} P[A_i A_j] + \sum\sum\sum_{i<j<k} P[A_i A_j A_k] - .. + (-1)^{n+1} P[A_1..A_n].$$

For $n = 3$ the above formula is:

$$P\left[A_1 \bigcup A_2 \bigcup A_3\right] = P[A_1] + P[A_2] + P[A_3] - P[A_1 A_2] - P[A_1 A_3] - P[A_2 A_3] + P[A_1 A_2 A_2].$$

7. $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$

Proofs involve manipulating sets to obtain disjoint sets and then apply the axioms.

## 1.2 Conditional Probability and Independence

In many statistical applications we have variables $X$ and $Y$ (or events $A$ and $B$) and want to explain or predict $Y$ or $A$ from $X$ or $B$, we are interested not only in marginal probabilities but in conditional ones as well, i.e., we want to incorporate some information in our predictions. Let $A$ and $B$ be two events in $\mathcal{A}$ and a probability function $P(.)$. The **conditional probability** of $A$ given event $B$, is denoted by $P[A|B]$ and is defined as follows:

**Definition 3** *The probability of an event $A$ given an event $B$, denoted by $P(A|B)$, is given by*

$$P([A|B) = \frac{P(A \cap B)}{P(B)} \quad if \quad P(B) > 0$$

*and is left undefined if $P(B) = 0$.*

From the above formula is evident $P[AB] = P[A|B]P[B] = P[B|A]P[A]$ if both $P[A]$ and $P[B]$ are nonzero. Notice that when speaking of conditional probabilities we are conditioning on some given event $B$; that is, we are assuming that the experiment has resulted in some outcome in $B$. $B$, in effect then becomes our "new" sample space. All probability properties of the previous section apply to conditional probabilities as well, i.e. $P(\cdot|B)$ is a probability measure. In particular:

1. $P(A|B) \geq 0$
2. $P(S|B) = 1$
3. $P(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B)$ for any pairwise disjoint events$\{A_i\}_{i=1}^{\infty}$.

Note that if $A$ and $B$ are mutually exclusive events, $P(A|B) = 0$. When $A \subseteq B$, $P(A|B) = \frac{P(A)}{P(B)} \geq P(A)$ with strict inequality unless $P(B) = 1$. When $B \subseteq A$, $P(A|B) = 1$.

However, there is an additional property (Law) called the **Law of Total Probabilities** which states that:

LAW OF TOTAL PROBABILITY:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

For a given probability space $(\Omega, \mathcal{A}, P[.])$, if $B_1, B_2, ..., B_n$ is a collection of mutually exclusive events in $\mathcal{A}$ satisfying $\bigcup_{i=1}^{n} B_i = \Omega$ and $P[B_i] > 0$ for $i = 1, 2, ..., n$ then for every $A \in \mathcal{A}$,

$$P[A] = \sum_{i=1}^{n} P[A|B_i]P[B_i]$$

Another important theorem in probability is the so called **Bayes' Theorem** which states:

BAYES RULE: Given a probability space $(\Omega, \mathcal{A}, P[.])$, if $B_1, B_2, ..., B_n$ is a collection of mutually exclusive events in $\mathcal{A}$ satisfying $\bigcup_{i=1}^{n} B_i = \Omega$ and $P[B_i] > 0$ for $i = 1, 2, ..., n$ then for every $A \in \mathcal{A}$ for which $P[A] > 0$ we have:

$$P[B_j|A] = \frac{P[A|B_j]P[B_j]}{\sum_{i=1}^{n} P[A|B_i]P[B_i]}$$

Notice that for events $A$ and $B \in \mathcal{A}$ which satisfy $P[A] > 0$ and $P[B] > 0$ we have:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

This follows from the definition of conditional independence and the law of total probability. The probability $P(B)$ is a prior probability and $P(A|B)$ frequently is a likelihood, while $P(B|A)$ is the posterior.

Finally the **Multiplication Rule** states:

Given a probability space $(\Omega, \mathcal{A}, P[.])$, if $A_1, A_2, ..., A_n$ are events in $\mathcal{A}$ for which $P[A_1 A_2 .....A_{n-1}] > 0$ then:

$$P[A_1A_2.....A_n] = P[A_1]P[A_2|A_1]P[A_3|A_1A_2].....P[A_n|A_1A_2....A_{n-1}]$$

EXAMPLE: A plant has two machines. Machine A produces 60% of the total output with a fraction defective of 0.02. Machine B the rest output with a fraction defective of 0.04. If a single unit of output is observed to be defective, what is the probability that this unit was produced by machine A?

If $A$ is the event that the unit was produced by machine A, $B$ the event that the unit was produced by machine B and $D$ the event that the unit is defective. Then we ask what is $P[A|D]$. But $P[A|D] = \frac{P[AD]}{P[D]}$. Now $P[AD] = P[D|A]P[A] = 0.02 * 0.6 = 0.012$. Also $P[D] = P[D|A]P[A] + P[D|B]P[B] = 0.012 + 0.04 * 0.4 = 0.028$. Consequently, $P[A|D] = 0.571$. Notice that $P[B|D] = 1 - P[A|D] = 0.429$. We can also use a tree diagram to evaluate $P[AD]$ and $P[BD]$.

EXAMPLE: A marketing manager believes the market demand potential of a new product to be high with a probability of 0.30, or average with probability of 0.50, or to be low with a probability of 0.20. From a sample of 20 employees, 14 indicated a very favorable reception to the new product. In the past such an employee response (14 out of 20 favorable) has occurred with the following probabilities: if the actual demand is high, the probability of favorable reception is 0.80; if the actual demand is average, the probability of favorable reception is 0.55; and if the actual demand is low, the probability of the favorable reception is 0.30. Thus given a favorable reception, what is the probability of actual high demand?

Again what we ask is $P[H|F] = \frac{P[HF]}{P[F]}$. Now $P[F] = P[H]P[F|H]+P[A]P[F|A]+P[L]P[F|L] = 0.24+0.275+0.06 = 0.575$. Also $P[HF] = P[F|H]P[H] = 0.24$. Hence $P[H|F] = \frac{0.24}{0.575} = 0.4174$

EXAMPLE: There are five boxes and they are numbered 1 to 5. Each box contains 10 balls. Box $i$ has $i$ defective balls and $10-i$ non-defective balls, $i = 1, 2, .., 5$. Consider the following random experiment: First a box is selected at random, and then a ball is selected at random from the selected box. 1) What is the probability

that a defective ball will be selected? 2) If we have already selected the ball and noted that is defective, what is the probability that it came from the box 5?

Let $A$ denote the event that a defective ball is selected and $B_i$ the event that box i is selected, $i = 1, 2, .., 5$. Note that $P[B_i] = 1/5$, for $i = 1, 2, .., 5$, and $P[A|B_i] = i/10$. Question 1) is what is $P[A]$? Using the theorem of total probabilities we have:

$P[A] = \sum_{i=1}^{5} P[A|B_i]P[B_i] = \sum_{i=1}^{5} \frac{i}{5}\frac{1}{5} = 3/10$. Notice that the total number of defective balls is 15 out of 50. Hence in this case we can say that $P[A] = \frac{15}{50} = 3/10$. This is true as the probabilities of choosing each of the 5 boxes is the same. Question 2) asks what is $P[B_5|A]$. Since box 5 contains more defective balls than box 4, which contains more defective balls than box 3 and so on, we expect to find that $P[B_5|A] > P[B_4|A] > P[B_3|A] > P[B_2|A] > P[B_1|A]$. We apply Bayes' theorem:

$$P[B_5|A] = \frac{P[A|B_5]P[B_5]}{\sum_{i=1}^{5} P[A|B_i]P[B_i]} = \frac{\frac{1}{2}\frac{1}{5}}{\frac{3}{10}} = \frac{1}{3}$$

Similarly $P[B_j|A] = \frac{P[A|B_j]P[B_j]}{\sum_{i=1}^{5} P[A|B_i]P[B_i]} = \frac{\frac{j}{10}\frac{1}{5}}{\frac{3}{10}} = \frac{j}{15}$ for $j = 1, 2, ..., 5$. Notice that unconditionally all $B_i's$ were equally likely.

Let $A$ and $B$ be two events in $\mathcal{A}$ and a probability function $P(.)$. Events $A$ and $B$ are defined **independent** if and only if one of the following conditions is satisfied:

(i) $P[AB] = P[A]P[B]$.

(ii) $P[A|B] = P[A]$ if $P[B] > 0$.

(iii) $P[B|A] = P[B]$ if $P[A] > 0$.

These are equivalent definitions except that $(i)$ does not really require $P(A)$, $P(B) > 0$. **Notice** that the property of two events $A$ and $B$ and the property that $A$ and $B$ are mutually exclusive are distinct, though related properties. We know that if $A$ and $B$ are mutually exclusive then $P[AB] = 0$. Now if these events are

also independent then $P[AB] = P[A]P[B]$, and consequently $P[A]P[B] = 0$, which means that either $P[A] = 0$ or $P[B] = 0$. Hence two mutually exclusive events are independent if $P[A] = 0$ or $P[B] = 0$. On the other hand if $P[A] \neq 0$ and $P[B] \neq 0$, then if $A$ and $B$ are independent can not be mutually exclusive and oppositely if they are mutually exclusive can not be independent. Also notice that independence is not transitive, i.e., $A$ independent of $B$ and $B$ independent of $C$ does not imply that $A$ is independent of $C$.

EXAMPLE: Consider tossing two dice. Let $A$ denote the event of an odd total, $B$ the event of an ace on the first die, and $C$ the event of a total of seven. We ask the following:

(i) Are $A$ and $B$ independent?

(ii) Are $A$ and $C$ independent?

(iii) Are $B$ and $C$ independent?

(i) $P[A|B] = 1/2$, $P[A] = 1/2$ hence $P[A|B] = P[A]$ and consequently $A$ and $B$ are independent.

(ii) $P[A|C] = 1 \neq P[A] = 1/2$ hence $A$ and $C$ are not independent.

(iii) $P[C|B] = 1/6 = P[C] = 1/6$ hence $B$ and $C$ are independent.

Notice that although $A$ and $B$ are independent and $C$ and $B$ are independent $A$ and $C$ are not independent.

Let us extend the independence of two events to several ones:

For a given probability space $(\Omega, \mathcal{A}, P[.])$, let $A_1, A_2, ..., A_n$ be n events in $\mathcal{A}$. Events $A_1, A_2, ..., A_n$ are defined to be **independent** if and only if:

$P[A_iA_j] = P[A_i]P[A_j]$ for $i \neq j$

$P[A_iA_jA_k] = P[A_i]P[A_j]P[A_k]$ for $i \neq j, i \neq k, k \neq j$

and so on

$P[\bigcap_{i=1}^{n} A_i] = \prod_{i=1}^{n} P[A_i]$

Notice that pairwise independence does not imply independence, as the following example shows.

EXAMPLE: Consider tossing two dice. Let $A_1$ denote the event of an odd face in the first die, $A_2$ the event of an odd face in the second die, and $A_3$ the event of an odd total. Then we have: $P[A_1]P[A_2] = \frac{1}{2}\frac{1}{2} = P[A_1A_2]$, $P[A_1]P[A_3] = \frac{1}{2}\frac{1}{2} = P[A_3|A_1]P[A_1] = P[A_1A_3]$, and $P[A_2A_3] = \frac{1}{4} = P[A_2]P[A_3]$ hence $A_1, A_2, A_3$ are pairwise independent. However notice that $P[A_1A_2A_3] = 0 \neq \frac{1}{8} = P[A_1]P[A_2]P[A_3]$. Hence $A_1, A_2, A_3$ are **not** independent.

# Chapter 2

# RANDOM VARIABLES, DISTRIBUTION FUNCTIONS, AND DENSITIES

The probability space $(S, \mathcal{A}, P)$ is not particularly easy to work with. In practice, we often need to work with spaces with some structure (metric spaces). It is convenient therefore to work with a cardinalization of $S$ by using the notion of random variable.

Formally, a random variable $X$ is just a mapping from the sample space to the real line, i.e.,

$$X : S \longrightarrow \mathbb{R},$$

with a certain property, it is a measurable mapping, i.e.

$$\mathcal{A}_X = \{A \subset S : X(A) \in \mathcal{B}\} = \left\{X^{-1}(B) : B \in \mathcal{B}\right\} \subseteq \mathcal{A},$$

where $\mathcal{B}$ is a sigma-algebra on $\mathbb{R}$, for any $B$ in $\mathcal{B}$ the inverse image belongs to $\mathcal{A}$. The probability measure $P_X$ can then be defined by

$$P_X\left(X \in B\right) = P\left(X^{-1}\left(B\right)\right).$$

It is straightforward to show that $\mathcal{A}_X$ is a $\sigma$-algebra whenever $\mathcal{B}$ is. Therefore, $P_X$ is a probability measure obeying Kolmogorov's axioms. Hence we have transferred $(S, \mathcal{A}, P) \longrightarrow (\mathbb{R}, \mathcal{B}, P_X)$, where $\mathcal{B}$ is the Borel $\sigma$-algebra when $X(S) = \mathbb{R}$ or any uncountable set, and $\mathcal{B}$ is $\mathcal{P}\left(X\left(S\right)\right)$ when $X\left(S\right)$ is finite. The function $X(.)$ must be such that the set $A_r$, defined by $A_r = \{\omega : X(\omega) \leq r\}$, belongs to $\mathcal{A}$ for every real number $r$, as elements of $\mathcal{B}$ are left-closed intervals of $\mathbb{R}$.

The important part of the definition is that in terms of a random experiment, $S$ is the totality of outcomes of that random experiment, and the function, or random variable, $X(.)$ with domain $S$ makes some real number correspond to each outcome of the experiment. The fact that we also require the collection of $\omega's$ for which $X(\omega) \leq r$ to be an event (i.e. an element of $\mathcal{A}$) for each real number $r$ is not much of a restriction since the use of random variables is, in our case, to describe only events.

EXAMPLE: Consider the experiment of tossing a single coin. Let the random variable $X$ denote the number of heads. In this case $S = \{head, tail\}$, and $X(\omega) = 1$ if $\omega = head$, and $X(\omega) = 0$ if $\omega = tail$. So the random variable $X$ associates a real number with each outcome of the experiment. To show that $X$ satisfies the definition we should show that $\{\omega : X(\omega) \leq r\}$, belongs to $\mathcal{A}$ for every real number $r$. $\mathcal{A} = \{\phi, \{head\}, \{tail\}, S\}$. Now if $r < 0$, $\{\omega : X(\omega) \leq r\} = \phi$, if $0 \leq r < 1$ then $\{\omega : X(\omega) \leq r\} = \{tail\}$, and if $r \geq 1$ then $\{\omega : X(\omega) \leq r\} = \{head, tail\} = S$. Hence, for each $r$ the set $\{\omega : X(\omega) \leq r\}$ belongs to $\mathcal{A}$ and consequently $X(.)$ is a random variable.

In the above example the random variable is described in terms of the random experiment as opposed to its functional form, which is the usual case.

We can now work with $(\mathbb{R}, \mathcal{B}, P_X)$, which has metric structure and algebra. For example, we toss two die in which case the sample space is

$$S = \{(1,1), (1,2), ..., (6,6)\}.$$

We can define two random variables: the Sum and Product:

$$\begin{aligned} X(S) &= \{2,3,4,5,6,7,8,9,10,11,12\} \\ X(S) &= \{1,2,3,4,5,6,8,9,10,...,36\} \end{aligned}$$

The simplest form of random variables are the indicators $I_A$

$$I_A(s) = \begin{cases} 1 & if \quad s \in A \\ 0 & if \quad s \notin A \end{cases}$$

This has associated sigma algebra in $S$

$$\{\phi, S, A, A^c\}$$

Finally, we give formal definition of a continuous real-valued random variable.

**Definition 4** *A random variable is continuous if its probability measure $P_X$ is absolutely continuous with respect to Lebesgue measure, i.e., $P_X(A) = 0$ whenever $\lambda(A) = 0$.*

### *2.0.1 Distribution Functions*

Associated with each random variable there is the distribution function

$$F_X(x) = P_X(X \leq x)$$

defined for all $x \in \mathbb{R}$. This function effectively replaces $P_X$. Note that we can reconstruct $P_X$ from $F_X$.

EXAMPLE. $S = \{H, T\}$, $X(H) = 1$, $X(T) = 0$, $(p = 1/2)$.

If $x < 0$, $F_X(x) = 0$

If $0 \leq x < 1$, $F_X(x) = 1/2$

If $x \geq 1$, $F_X(x) = 1$.

EXAMPLE. The logit c.d.f. is

$$F_X(x) = \frac{1}{1 + e^{-x}}$$

It is continuous everywhere and asymptotes to 0 and 1 at $\pm\infty$ respectively. Strictly increasing.

Note that the distribution function $F_X(x)$ of a continuous random variable is a continuous function. The distribution function of a discrete random variable is a step function.

**Theorem 5** *A function $F(\cdot)$ is a c.d.f. of a random variable $X$ if and only if the following three conditions hold*

*1.* $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$

*2.* $F$ *is a nondecreasing function in* $x$

*3.* $F$ *is right-continuous, i.e., for all* $x_0$, $\lim_{x\to x_0^+} F(x) = F(x_0)$

*4.* $F$ *is continuous except at a set of points of Lebesgue measure zero.*

### *2.0.2 Discrete Random Variables.*

As we have already said, a random variable $X$ will be defined to be **discrete** if the range of $X$ is countable. If a random variable $X$ is discrete, then its corresponding cumulative distribution function $F_X(.)$ will be defined to be **discrete**, i.e. a step function.

By the range of $X$ being countable we mean that there exists a finite or denumerable set of real numbers, say $x_1, x_2, ...x_n, ...$, such that $X$ takes on values only in that set. If $X$ is discrete with distinct values $x_1, x_2, ...x_n, ...$, then $S = \bigcup_n \{\omega : X(\omega) = x_n\}$, and $\{X = x_i\} \cap \{X = x_j\} = \phi$ for $i \neq j$. Hence $1 = P[S] = \sum_n P[X = x_n]$ by the third axiom of probability.

If $X$ is a discrete random variable with distinct values $x_1, x_2, ...x_n, ...$, then the function, denoted by $f_X(.)$ and defined by

$$f_X(x) = \begin{cases} P[X = x] & if \quad x = x_j, \quad j = 1, 2, ..., n, ... \\ 0 & if \quad x \neq x_j \end{cases}$$

is defined to be the **discrete density function** of $X$.

Notice that the discrete density function tell us how likely or probable each of the values of a discrete random variable is. It also enables one to calculate the probability of events described in terms of the discrete random variable. Also notice that for any discrete random variable $X$, $F_X(.)$ can be obtained from $f_X(.)$, and vice versa

EXAMPLE: Consider the experiment of tossing a single die. Let $X$ denote the number of spots on the upper face. Then for this case we have:

$X$ takes any value from the set $\{1, 2, 3, 4, 5, 6\}$. So $X$ is a discrete random variable. The density function of $X$ is: $f_X(x) = P[X = x] = 1/6$ for any

$x \in \{1,2,3,4,5,6\}$ and 0 otherwise. The cumulative distribution function of $X$ is: $F_X(x) = P[X \leq x] = \sum_{n=1}^{[x]} P[X = n]$ where $[x]$ denotes the integer part of $x$.. Notice that $x$ can be any real number. However, the points of interest are the elements of $\{1,2,3,4,5,6\}$. Notice also that in this case $\Omega = \{1,2,3,4,5,6\}$ as well, and we do not need any reference to $\mathcal{A}$. ■

EXAMPLE: Consider the experiment of tossing two dice. Let $X$ denote the total of the upturned faces. Then for this case we have:

$\Omega = \{(1,1),(1,2),...(1,6),(2,1),(2,2),....(2,6),(3,1),.....,(6,6)\}$ a total of (using the Multiplication rule) $36 = 6^2$ elements. $X$ takes values from the set $\{2,3,4,5,6,7,8,9,10,11,12\}$. The density function is:

$$f_X(x) = P[X = x] = \begin{cases} 1/36 & for \quad x = 2 \quad or \quad x = 12 \\ 2/36 & for \quad x = 3 \quad or \quad x = 11 \\ 3/36 & for \quad x = 4 \quad or \quad x = 10 \\ 4/36 & for \quad x = 5 \quad or \quad x = 9 \\ 5/36 & for \quad x = 6 \quad or \quad x = 8 \\ 1/36 & for \quad x = 7 \\ 0 & for \quad any \quad other \quad x \end{cases}$$

The cumulative distribution function is:

$$F_X(x) = P[X \leq x] = \sum_{n=1}^{[x]} P[X = n] = \begin{cases} 0 & for \quad x < 2 \\ \frac{1}{36} & for \quad 2 \leq x < 3 \\ \frac{3}{36} & for \quad 3 \leq x < 4 \\ \frac{6}{36} & for \quad 4 \leq x < 5 \\ \frac{10}{36} & for \quad 5 \leq x < 6 \\ & ..... \\ \frac{35}{36} & for \quad 11 \leq x < 12 \\ 1 & for \quad 12 \leq x \end{cases}$$

Notice that, again, we do not need any reference to $\mathcal{A}$. ■

In fact we can speak of discrete density functions without reference to some random variable at all.

Any function $f(.)$ with domain the real line and counterdomain $[0,1]$ is defined to be a **discrete density function** if for some countable set $x_1, x_2, ...x_n, ....$ has the following properties:

i) $f(x_j) > 0$ for $j = 1, 2, ...$

ii) $f(x) = 0$ for $x \neq x_j$; $j = 1, 2, ...$

iii) $\sum f(x_j) = 1$, where the summation is over the points $x_1, x_2, ...x_n, ....$

*2.0.3 Continuous Random Variables*

A random variable $X$ is called **continuous** if there exist a function $f_X(.)$ such that $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ for every real number $x$. In such a case $F_X(x)$ is the **cumulative distribution** and the function $f_X(.)$ is the **density function**.

Notice that according to the above definition the density function is not uniquely determined. The idea is that if the a function change value if a few points its integral is unchanged. Furthermore, notice that $f_X(x) = dF_X(x)/dx$.

The notations for discrete and continuous density functions are the same, yet they have different interpretations. We know that for discrete random variables $f_X(x) = P[X = x]$, which is not true for continuous random variables. Furthermore, for discrete random variables $f_X(.)$ is a function with domain the real line and counterdomain the interval $[0,1]$, whereas, for continuous random variables $f_X(.)$ is a function with domain the real line and counterdomain the interval $[0, \infty)$. Note that for a continuous r.v.

$$P(X = x) \leq P(x - \varepsilon \leq X \leq x) = F_X(x) - F_X(x - \varepsilon) \to 0$$

as $\varepsilon \to 0$, by the continuity of $F_X(x)$. The set $\{X = x\}$ is an example of a set of measure (in this case the measure is $P$ or $P_X$) zero. In fact, any countable set is of measure zero under a distribution which is absolutely continuous with respect to Lebesgue measure. Because the probability of a singleton is zero

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b)$$

for any $a$, $b$.

EXAMPLE: Let $X$ be the random variable representing the length of a telephone conversation. One could model this experiment by assuming that the distribution of $X$ is given by $F_X(x) = (1 - e^{-\lambda x})$ where $\lambda$ is some positive number and the random variable can take values only from the interval $[0, \infty)$. The density function is $dF_X(x)/dx = f_X(x) = \lambda e^{-\lambda x}$. If we assume that telephone conversations are measured in minutes, $P[5 < X \leq 10] = \int_5^{10} f_X(x)dx = \int_5^{10} \lambda e^{-\lambda x}dx = e^{-5\lambda} - e^{-10\lambda}$, and for $\lambda = 1/5$ we have that $P[5 < X \leq 10] = e^{-1} - e^{-2} = 0.23$. ■

The example above indicates that the density functions of continuous random variables are used to calculate probabilities of events defined in terms of the corresponding continuous random variable $X$ i.e. $P[a < X \leq b] = \int_a^b f_X(x)dx$. Again we can give the definition of the density function without any reference to the random variable i.e. any function $f(.)$ with domain the real line and counterdomain $[0, \infty)$ is defined to be a **probability density function** $iff$

(i) $f(x) \geq 0$ for all $x$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$.

In practice when we refer to the certain distribution of a random variable, we state its density or cumulative distribution function. However, notice that not all random variables are either discrete or continuous.

# Chapter 3

# EXPECTATIONS AND MOMENTS OF RANDOM VARIABLES

An extremely useful concept in problems involving random variables or distributions is that of expectation.

### *3.0.4 Mean or Expectation*

Let $X$ be a random variable. The **mean** or the **expected value** of $X$, denoted by $E[X]$ or $\mu_X$, is defined by:

(i) $E[X] = \sum x_j P[X = x_j] = \sum x_j f_X(x_j)$

if $X$ is a discrete random variable with counterdomain the countable set $\{x_1, ..., x_j, ..\}$

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

if $X$ is a continuous random variable with density function $f_X(x)$ and if either $\left|\int_0^{\infty} x f_X(x) dx\right| < \infty$ or $\left|\int_{-\infty}^{0} x f_X(x) dx\right| < \infty$ or both.

(iii) $E[X] = \int_0^{\infty} [1 - F_X(x)] dx - \int_{-\infty}^{0} F_X(x) dx$

for an arbitrary random variable $X$.

(i) and (ii) are used in practice to find the mean for discrete and continuous random variables, respectively. (iii) is used for the mean of a random variable that is neither discrete nor continuous.

Notice that in the above definition we assume that the sum and the integrals exist. Also that the summation in (i) runs over the possible values of $j$ and the $j^{th}$ term is the value of the random variable multiplied by the probability that the random variable takes this value. Hence $E[X]$ is an average of the values that the

random variable takes on, where each value is weighted by the probability that the random variable takes this value. Values that are more probable receive more weight. The same is true in the integral form in (ii). There the value $x$ is multiplied by the approximate probability that $X$ equals the value $x$, i.e. $f_X(x)dx$, and then integrated over all values.

*Notice* that in the definition of a mean of a random variable, only density functions or cumulative distributions were used. Hence we have really defined the mean for these functions without reference to random variables. We then call the defined mean the mean of the cumulative distribution or the appropriate density function. Hence, we can speak of the mean of a distribution or density function as well as the mean of a random variable.

*Notice* that $E[X]$ is the center of gravity (or centroid) of the unit mass that is determined by the density function of $X$. So the mean of $X$ is a measure of where the values of the random variable are centered or located i.e. is a measure of central location.

EXAMPLE: Consider the experiment of tossing two dice. Let $X$ denote the total of the upturned faces. Then for this case we have:

$$E[X] = \sum_{i=2}^{12} i f_X(i) = 7$$

EXAMPLE: Consider a $X$ that can take only to possible values, 1 and -1, each with probability 0.5. Then the mean of $X$ is:

$$E[X] = 1 * 0.5 + (-1) * 0.5 = 0$$

Notice that the mean in this case is not one of the possible values of $X$.

EXAMPLE: Consider a continuous random variable $X$ with density function $f_X(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$. Then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{0}^{\infty} x \lambda e^{-\lambda x} dx = 1/\lambda$$

EXAMPLE: Consider a continuous random variable $X$ with density function $f_X(x) = x^{-2}$ for $x \in [1, \infty)$. Then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{1}^{\infty} x x^{-2} dx = \lim_{b \to \infty} \log b = \infty$$

so we say that the mean does not exist, or that it is infinite.

Median of $X$: When $F_X$ is continuous and strictly increasing, we can define the median of $X$, denoted $m(X)$, as being the unique solution to

$$F_X(m) = \frac{1}{2}.$$

Since in this case, $F_X^{-1}(\cdot)$ exists, we can alternatively write $m = F_X^{-1}(\frac{1}{2})$. For discrete r.v., there may be many m that satisfy this or may none. Suppose

$$X = \begin{cases} 0 & 1/3 \\ 1 & 1/3 \\ 2 & 1/3 \end{cases},$$

then there does not exist an $m$ with $F_X(m) = \frac{1}{2}$. Also, if

$$X = \begin{cases} 0 & 1/4 \\ 1 & 1/4 \\ 2 & 1/4 \\ 3 & 1/4 \end{cases},$$

then any $1 \leq m \leq 2$ is an adequate median.

Note that if $E(X^n)$ exists, then so does $E(X^{n-1})$ but not vice versa ($n > 0$). Also when the support is infinite, the expectation does not necessarily exist.

If $\int_0^\infty x f_X(x)dx = \infty$ but $\int_{-\infty}^0 x f_X(x)dx > -\infty$, then $E(X) = \infty$

If $\int_0^\infty x f_X(x)dx = \infty$ and $\int_{-\infty}^0 x f_X(x)dx = -\infty$, then $E(X)$is not defined.

EXAMPLE: [Cauchy] $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. This density function is symmetric about zero, and one is temted to say that $E(X) = 0$. But $\int_0^\infty x f_X(x)dx = \infty$ and $\int_{-\infty}^0 x f_X(x)dx = -\infty$, so $E(X)$ does not exist according to the above definition.

Now consider $Y = g(X)$, where g is a (piecewise) monotonic continuous function. Then

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_{-\infty}^{\infty} g(x) f_X(x)dx = E(g(x))$$

**Theorem 6** *Expectation has the following properties:*

*1. [Linearity]* $E\left(a_1 g_1(X) + a_2 g_2(X) + a_3\right) = a_1 E\left(g_1(X)\right) + a_2 E\left(g_2(X)\right) + a_3$

*2. [Monotonicity] If* $g_1(x) \geq g_2(x) \Rightarrow E\left(g_1(X)\right) \geq E\left(g_2(X)\right)$

*3. Jensen's inequality. If* $g(x)$ *is a weakly convex function, i.e.,* $g\left(\lambda x + (1-\lambda) y\right) \leq \lambda g\left(x\right) + (1-\lambda) g\left(y\right)$ *for all* $x$, $y$, *and all with* $0 \leq \lambda \leq 1$, *then*

$$E\left(g(X)\right) \geq g\left(E\left(X\right)\right).$$

**An Interpretation of Expectation**

We claim that $E\left(X\right)$ is the unique minimizer of $E\left(X-\theta\right)^2$ with respect to $\theta$, assuming that the second moment of $X$ is finite.

**Theorem 7** *Suppose that* $E\left(X^2\right)$ *exists and is finite, then* $E\left(X\right)$ *is the unique minimizer of* $E\left(X-\theta\right)^2$ *with respect to* $\theta$.

This Theorem says that the Expectation is the closest quantity to $\theta$, in mean square error.

*3.0.5 Variance*

Let $X$ be a random variable and $\mu_X$ be $E[X]$. The **variance** of X, denoted by $\sigma_X^2$ or $var[X]$, is defined by:

(i) $var[X] = \sum (x_j - \mu_X)^2 P[X = x_j] = \sum (x_j - \mu_X)^2 f_X(x_j)$

if $X$ is a discrete random variable with counterdomain the countable set $\{x_1, ..., x_j, ..\}$

(ii) $var[X] = \int_{-\infty}^{\infty} (x_j - \mu_X)^2 f_X(x) dx$

if $X$ is a continuous random variable with density function $f_X(x)$.

(iii) $var[X] = \int_0^{\infty} 2x[1 - F_X(x) + F_X(-x)]dx - \mu_X^2$

for an arbitrary random variable $X$.

The variances are defined only if the series in (i) is convergent or if the integrals in (ii) or (iii) exist. Again, the variance of a random variable is defined in terms of

the density function or cumulative distribution function of the random variable and consequently, variance can be defined in terms of these functions without reference to a random variable.

Notice that variance is a measure of spread since if the values of the random variable $X$ tend to be far from their mean, the variance of $X$ will be larger than the variance of a comparable random variable whose values tend to be near their mean. It is clear from (i), (ii) and (iii) that the variance is a nonnegative number.

If $X$ is a random variable with variance $\sigma_X^2$, then the **standard deviation** of $X$, denoted by $\sigma_X$, is defined as $\sqrt{var(X)}$

The standard deviation of a random variable, like the variance, is a measure of spread or dispersion of the values of a random variable. In many applications it is preferable to the variance since it will have the same measurement units as the random variable itself.

EXAMPLE: Consider the experiment of tossing two dice. Let $X$ denote the total of the upturned faces. Then for this case we have ($\mu_X = 7$):

$$var[X] = \sum_{i=2}^{12}(i - \mu_X)^2 f_X(i) = 210/36$$

EXAMPLE: Consider a $X$ that can take only to possible values, 1 and -1, each with probability 0.5. Then the variance of $X$ is ($\mu_X = 0$):

$$var[X] = 0.5 * 1^2 + 0.5 * (-1)^2 = 1$$

EXAMPLE: Consider a $X$ that can take only to possible values, 10 and -10, each with probability 0.5. Then we have:

$$\mu_X = E[X] = 10 * 0.5 + (-10) * 0.5 = 0$$

$$var[X] = 0.5 * 10^2 + 0.5 * (-10)^2 = 100$$

Notice that in examples 2 and 3 the two random variables have the same mean but different variance, larger being the variance of the random variable with values further away from the mean.

EXAMPLE: Consider a continuous random variable $X$ with density function $f_X(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$. Then ($\mu_X = 1/\lambda$):

$$var[X] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx = \int_{0}^{\infty} (x-1/\lambda)^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2}$$

EXAMPLE: Consider a continuous random variable $X$ with density function $f_X(x) = x^{-2}$ for $x \in [1, \infty)$. Then we know that the mean of $X$ does not exist. Consequently, we can not define the variance.

Notice that

$$Var\left(X\right) = E\left[\left(X - E(X)\right)^2\right] = E\left(X^2\right) - E^2\left(X\right)$$

and that

$$Var\left(aX+b\right) = a^2 Var\left(X\right), \quad SD = \sqrt{Var}, \quad SD\left(aX+b\right) = |a|SD(X),$$

i.e., $SD(X)$ changes proportionally. Variance/standard deviation measures dispersion, higher variance more spread out. Interquartile range: $F_X^{-1}(3/4) - F_X^{-1}(1/4)$, the range of middle half always exists and is an alternative measure of dispersion.

### *3.0.6 Higher Moments of a Random Variable*

If $X$ is a random variable, the **$\mathbf{r^{th}}$ raw moment** of $X$, denoted by $\mu_r^{/}$, is defined as:

$$\mu_r^{/} = E[X^r]$$

if this expectation exists. Notice that $\mu_r^{/} = E[X] = \mu_1^{/} = \mu_X$, the mean of $X$.

If $X$ is a random variable, the **$\mathbf{r^{th}}$ central moment** of $X$ **about** $\alpha$ is defined as $E[(X-\alpha)^r]$. If $\alpha = \mu_X$, we have the **$\mathbf{r^{th}}$ central moment** of $X$ about $\mu_X$, denoted by $\mu_r$, which is:

$$\mu_r = E[(X-\mu_X)^r]$$

We have measures defined in terms of quantiles to describe some of the characteristics of random variables or density functions. The **$\mathbf{q^{th}}$ quantile** of a random variable $X$ or of its corresponding distribution is denoted by $\xi_q$ and is defined as the smallest number $\xi$ satisfying $F_X(\xi) \geq q$. If $X$ is a continuous random variable, then the **$\mathbf{q^{th}}$ quantile** of $X$ is given as the smallest number $\xi$ satisfying $F_X(\xi) \geq q$.

The **median** of a random variable $X$, denoted by $med_X$ or $med(X)$, or $\xi_q$, is the $0.5^{th}$ quantile. Notice that if $X$ a continuous random variable the median of $X$ satisfies:

$$\int_{-\infty}^{med(X)} f_X(x)dx = \frac{1}{2} = \int_{med(X)}^{\infty} f_X(x)dx$$

so the median of $X$ is any number that has half the mass of $X$ to its right and the other half to its left. The median and the mean are measures of central location.

The third moment about the mean $\mu_3 = E\left(X - E\left(X\right)\right)^3$ is called a measure of asymmetry, or **skewness**. Symmetrical distributions can be shown to have $\mu_3 = 0$. Distributions can be skewed to the left or to the right. However, knowledge of the third moment gives no clue as to the shape of the distribution, i.e. it could be the case that $\mu_3 = 0$ but the distribution to be far from symmetrical. The ratio $\frac{\mu_3}{\sigma^3}$ is unitless and is call the **coefficient of skewness**. An alternative measure of skewness is provided by the ratio: (mean-median)/(standard deviation)

The fourth moment about the mean $\mu_4 = E\left(X - E\left(X\right)\right)^4$ is used as a measure of **kurtosis**, which is a degree of flatness of a density near the center. The **coefficient of kurtosis** is defined as $\frac{\mu_4}{\sigma^4} - 3$ and positive values are sometimes used to indicate that a density function is more peaked around its center than the normal (leptokurtic distributions). A positive value of the coefficient of kurtosis is indicative for a distribution which is flatter around its center than the standard normal (platykurtic distributions). This measure suffers from the same failing as the measure of skewness i.e. it does not always measure what it supposed to.

While a particular moment or a few of the moments may give little information about a distribution the entire set of moments will determine the distribution exactly. In applied statistics the first two moments are of great importance, but the third and forth are also useful.

### *3.0.7 Moment Generating Functions*

Finally we turn to the moment generating function (mgf) and characteristic Function (cf). The mgf is defined as

$$M_X(t) = E\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$$

for any real $t$, provided this integral exists in some neighborhood of 0. It is the Laplace transform of the function $f_X(\cdot)$ with argument $-t$. We have the useful inversion formula

$$f_X(x) = \int_{-\infty}^{\infty} M_X(t)\, e^{-tx} dt$$

The mgf is of limited use, since it does not exist for many r.v. the $cf$ is applicable more generally, since it always exists:

$$\varphi_X(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} f_X(x)dx = \int_{-\infty}^{\infty} \cos(tx)\, f_X(x)dx + i\int_{-\infty}^{\infty} \sin(tx)\, f_X(x)dx$$

This essentially is the Fourier transform of the function $f_X(\cdot)$ and there is a well defined inversion formula

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t)\, dt$$

If $X$ is symmetric about zero, the complex part of $cf$ is zero. Also,

$$\frac{d^r}{dt^r}\varphi_X(0) = E\left(i^r X^r e^{itX}\right) \downarrow_{t=0} = i^r E\left(X^r\right), \quad r = 1, 2, 3, ..$$

Thus the moments of $X$ are related to the derivative of the $cf$ at the origin.

If

$$c(t) = \int_{-\infty}^{\infty} \exp(itx)\, dF(x)$$

notice that

$$\frac{d^r c(t)}{dt^r} = \int_{-\infty}^{\infty} (ix)^r \exp(itx)\, dF(x)$$

and

$$\left.\frac{d^r c(t)}{dt^r}\right|_{t=0} = \int_{-\infty}^{\infty} (ix)^r\, dF(x) = (i)^r \mu_r^{/} \Rightarrow \mu_r^{/} = (-i)^r \left.\frac{d^r c(t)}{dt^r}\right|_{t=0}$$

the $r^{th}$ uncenterd moment. Now expanding $c(t)$ in powers of $t$ we get

$$c(t) = c(0) + \left.\frac{d^r c(t)}{dt^r}\right|_{t=0} t + ... + \left.\frac{d^r c(t)}{dt^r}\right|_{t=0} \frac{(t)^r}{r!} + ... = 1 + \mu_1^{/}(it) + ... + \mu_r^{/}\frac{(it)^r}{r!} + ...$$

The cummulants are defined as the coefficients $\kappa_1$, $\kappa_2, ..., \kappa_r$ of the identity in $it$

$$\begin{aligned} \exp\left(\kappa_1(it) + \kappa_2\frac{(it)^2}{2!} + ... + \kappa_r\frac{(it)^r}{r!} + ...\right) &= 1 + \mu_1^{/}(it) + ... + \mu_r^{/}\frac{(it)^r}{r!} + ... \\ &= c(t) = \int_{-\infty}^{\infty} \exp(itx)\, dF(x) \end{aligned}$$

The cumulant-moment connection:

Suppose $X$ is a random variable with $n$ moments $a_1, ...a_n$. Then $X$ has $n$ cumulants $k_1, ...k_n$ and

$$a_{r+1} = \sum_{j=0}^{r} \binom{r}{j} a_j k_{r+1-j} \text{ for } r = 0, ..., n-1.$$

Writing out for $r = 0, ...3$ produces:

$$\begin{aligned} a_1 &= k_1 \\ a_2 &= k_2 + a_1 k_1 \\ a_3 &= k_3 + 2a_1 k_2 + a_2 k_1 \\ a_4 &= k_4 + 3a_1 k_3 + 3a_2 k_2 + a_3 k_1. \end{aligned}$$

These recursive formulas can be used to calculate the $a's$ efficiently from the $k's$, and vice versa. When $X$ has mean 0, that is, when $a_1 = 0 = k_1$, $a_j$ becomes

$$\mu_j = E((X - E(X))^j),$$

so the above formulas simplify to:

$$\begin{aligned} \mu_2 &= k_2 \\ \mu_3 &= k_3 \\ \mu_4 &= k_4 + 3k_2^2. \end{aligned}$$

### *3.0.8 Expectations of Functions of Random Variablers*

**Product and Quotient**

Let $f(X,Y) = \frac{X}{Y}$, $E(X) = \mu_X$ and $E(Y) = \mu_Y$. Then, expanding $f(X,Y) = \frac{X}{Y}$ around $(\mu_X, \mu_Y)$, we have

$$f(X,Y) = \frac{\mu_X}{\mu_Y} + \frac{1}{\mu_Y}(X-\mu_X) - \frac{\mu_X}{(\mu_Y)^2}(Y-\mu_Y) + \frac{\mu_X}{(\mu_Y)^3}(Y-\mu_Y)^2 - \frac{1}{(\mu_Y)^2}(X-\mu_X)(Y-\mu_Y)$$

as $\frac{\partial f}{\partial X} = \frac{1}{Y}$, $\frac{\partial f}{\partial Y} = -\frac{X}{Y^2}$, $\frac{\partial^2 f}{\partial X^2} = 0$, $\frac{\partial^2 f}{\partial X \partial Y} = \frac{\partial^2 f}{\partial Y \partial X} = -\frac{1}{Y^2}$, and $\frac{\partial^2 f}{\partial Y^2} = 2\frac{X}{Y^3}$. Taking expectations we have

$$E\left(\frac{X}{Y}\right) = \frac{\mu_X}{\mu_Y} + \frac{\mu_X}{(\mu_Y)^3} Var(Y) - \frac{1}{(\mu_Y)^2} Cov(X,Y).$$

For the variance, take again the variance of the Taylor expansion and keeping only terms up to order 2 we have:

$$Var\left(\frac{X}{Y}\right) = \frac{(\mu_X)^2}{(\mu_Y)^2}\left[\frac{Var(X)}{(\mu_X)^2} + \frac{Var(Y)}{(\mu_Y)^2} - 2\frac{Cov(X,Y)}{\mu_X \mu_Y}\right].$$

# Chapter 4

# EXAMPLES OF PARAMETRIC UNIVARIATE DISTRIBUTIONS

A parametric family of density functions is a collection of density functions that are indexed by a quantity called parameter, e.g. let $f(x;\lambda) = \lambda e^{-\lambda x}$ for $x > 0$ and some $\lambda > 0$. $\lambda$ is the parameter, and as $\lambda$ ranges over the positive numbers, the collection $\{f(.;\lambda) : \lambda > 0\}$ is a parametric family of density functions.

### *4.0.9 Discrete Distributions*

**UNIFORM:**

Suppose that for $j = 1, 2, 3, ...., n$

$$P(X = x_j|\mathcal{X}) = \frac{1}{n}$$

where $\{x_1, x_2, ...x_n\} = \mathcal{X}$ is the support. Then

$$E(X) = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad Var(X) = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2.$$

The c.d.f. here is

$$P(X \leq x) = \frac{1}{n}\sum_{j=1}^{n} 1(x_j \leq x)$$

**Bernoulli**

A random variable whose outcome have been classified into two categories, called "success" and "failure", represented by the letters s and f, respectively, is called a Bernoulli trial. If a random variable $X$ is defined as 1 if a Bernoulli trial results in

success and 0 if the same Bernoulli trial results in failure, then $X$ has a Bernoulli distribution with parameter $p = P[success]$. The definition of this distribution is:

A random variable $X$ has a Bernoulli distribution if the discrete density of $X$ is given by:

$$f_X(x) = f_X(x;p) = \begin{cases} p^x(1-p)^{1-x} & for \quad x = 0,1 \\ 0 & otherwise \end{cases}$$

where $p = P[X = 1]$. For the above defined random variable $X$ we have that:

$$E[X] = p \qquad and \qquad var[X] = p(1-p)$$

**BINOMIAL:**

Consider a random experiment consisting of $n$ repeated independent Bernoulli trials with $p$ the probability of success at each individual trial. Let the random variable $X$ represent the number of successes in the $n$ repeated trials. Then $X$ follows a Binomial distribution. The definition of this distribution is:

A random variable $X$ has a **binomial** distribution, $X \sim Binomial(n,p)$, if the discrete density of $X$ is given by:

$$f_X(x) = f_X(x;n,p) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & for \quad x = 0,1,...,n \\ 0 & otherwise \end{cases}$$

where $p = P[X = 1]$ i.e. the probability of success in each independent Bernoulli trial and $n$ is the total number of trials. For the above defined random variable $X$ we have that:

$$E[X] = np \qquad and \qquad var[X] = np(1-p)$$

Mgf

$$M_X(t) = \left[pe^t + (1-p)\right]^n.$$

EXAMPLE: Consider a stock with value $S = 50$. Each period the stock moves up or down, independently, in discrete steps of 5. The probability of going up is

$p = 0.7$ and down $1 - p = 0.3$. What is the expected value and the variance of the value of the stock after 3 period?

If we call $X$ the random variable which is a success if the stock moves up and failure if the stock moves down. Then $P[X = success] = P[X = 1] = 0.7$, and $X \tilde{} Binomial(3, p)$. Now $X$ can take the values $0, 1, 2, 3$ i.e. no success, 1 success and 2 failures, etc.. The value of the stock in each case and the probabilities are:

$$S = 35, \text{ and } f_X(0) = \begin{pmatrix} 3 \\ 0 \end{pmatrix} p^0(1-p)^{3-0} = 1 * 0.3^3 = 0.027,$$

$$S = 45, \text{ and } f_X(1) = \begin{pmatrix} 3 \\ 1 \end{pmatrix} p^1(1-p)^{3-1} = 3 * 0.7 * 0.3^2 = 0.189,$$

$$S = 55, \text{ and } f_X(2) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} p^2(1-p)^{3-2} = 3 * 0.7^2 * 0.3 = 0.441,$$

$$S = 65 \text{ and } f_X(3) = \begin{pmatrix} 3 \\ 3 \end{pmatrix} p^3(1-p)^{3-3} = 1 * 0.7^3 = 0.343.$$

Hence the expected stock value is:

$E[S] = 35 * 0.027 + 45 * 0.189 + 55 * 0.441 + 65 * 0.343 = 56$, and $var[S] = (35 - 56)^2 * 0.027 + (-11)^2 * 0.189 + (-1)^2 * 0.441 + (9)^2 * 0.343$.

**Hypergeometric**

Let $X$ denote the number of defective balls in a sample of size $n$ when sampling is done **without** replacement from a box containing $M$ balls out of which $K$ are defective. The $X$ has a hypergeometric distribution. The definition of this distribution is:

A random variable $X$ has a **hypergeometric** distribution if the discrete den-

sity of $X$ is given by:

$$f_X(x) = f_X(x; M, K, n) = \begin{cases} \dfrac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} & for \quad x = 0, 1, ..., n \\ 0 & otherwise \end{cases}$$

where $M$ is a positive integer, $K$ is a nonnegative that is at most $M$, and $n$ is a positive integer that is at most $M$. For this distribution we have that:

$$E[X] = n\frac{K}{M} \qquad and \qquad var[X] = n\frac{K}{M}\frac{M-K}{M}\frac{M-n}{M-1}$$

Notice the difference of the binomial and the hypergeometric i.e. for the binomial distribution we have Bernoulli trials i.e. independent trials with fixed probability of success or failure, whereas in the hypergeometric in each trial the probability of success or failure changes depending on the result.

**Geometric**

Consider a sequence of independent Bernoulli trials with $p$ equal the probability of success on an individual trial. Let the random variable $X$ represent the number of trials required before the first success. Then $X$ has a geometric distribution. The definition of this distribution is: A random variable $X$ has a **geometric** distribution, $X \sim geometric(p)$ , if the discrete density of $X$ is given by:

$$f_X(x) = f_X(x; p) = \begin{cases} p(1-p)^x & for \quad x = 0, 1, ..., n \\ 0 & otherwise \end{cases}$$

where $p$ is the probability of success in each Bernoulli trial. For this distribution we have that:

$$E[X] = \frac{1-p}{p} \qquad and \qquad var[X] = \frac{1-p}{p^2}$$

It is worth noticing that the Binomial distribution $Binomial(n,p)$ can be approximated by a $Poisson(np)$ (see below). The approximation is more valid as $n \to \infty$, $p \to 0$, in such a way so that $np = constant$.

**POISSON:**

A random variable $X$ has a **Poisson** distribution, $X \sim Poisson(\lambda)$, if the discrete density of $X$ is given by:

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, 3, ...$$

In calculations with the Poisson distribution we may use the fact that

$$e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!} \quad for \quad any \quad t.$$

Employing the above we can prove that

$$E(X) = \lambda, \quad E(X(X-1)) = \lambda^2, \quad Var(X) = \lambda.$$

The Poisson distribution provides a realistic model for many random phenomena. Since the values of a Poisson random variable are nonnegative integers, any random phenomenon for which a count of some sort is of interest is a candidate for modeling in assuming a Poisson distribution. Such a count might be the number of fatal traffic accidents per week in a given place, the number of telephone calls per hour, arriving in a switchboard of a company, the number of pieces of information arriving per hour, etc.

EXAMPLE: It is known that the average number of daily changes in excess of 1%, for a specific stock Index, occurring in each six-month period is 5. What is the probability of having one such a change within the next 6 months? What is the probability of at least 3 changes within the same period?

We model the number of in excess of 1% changes, $X$, within the next 6 months as a Poisson process. We know that $E[X] = \lambda = 5$. Hence $f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-5}5^x}{x!}$,

for $x = 0, 1, 2, , ...$ Then $P[X = 1] = f_X(1) = \frac{e^{-5}5^1}{1!} = 0.0337$. Also $P[X \geq 3] = 1 - P[X < 3] =$

$$= 1 - P[X = 0] - P[X = 1] - P[X = 2] =$$
$$= 1 - \frac{e^{-5}5^0}{0!} - \frac{e^{-5}5^1}{1!} - \frac{e^{-5}5^2}{2!} = 0.875.$$

We can approximate the Binomial with Poisson. The approximation is better the smaller the $p$ and the larger the $n$.

*4.0.10 Continuous Distributions*

**UNIFORM ON** $[a, b]$.

A very simple distribution for a continuous random variable is the uniform distribution. Its density function is:

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & if \quad x \in [a,b] \\ 0 & otherwise \end{cases},$$

and

$$F(x|a,b) = \int_a^x f(z|a,b)\, dz = \frac{x-a}{b-a},$$

where $-\infty < a < b < \infty$. Then the random variable $X$ is defined to be **uniformly** distributed over the interval $[a, b]$. Now if $X$ is uniformly distributed over $[a, b]$ then

$$E(X) = \frac{a+b}{2}, \quad median = \frac{a+b}{2}, \quad Var(X) = \frac{(b-a)^2}{12}.$$

If $X \backsim U[a,b] \Longrightarrow X - a \backsim U[0, b-a] \Longrightarrow \frac{X-a}{b-a} \backsim U[0, b-a]$. Notice that if a random variable is uniformly distributed over one of the following intervals $[a, b)$, $(a, b]$, $(a, b)$ the density function, expected value and variance does not change.

**Exponential Distribution**

If a random variable $X$ has a density function given by:

$$f_X(x) = f_X(x;\lambda) = \lambda e^{-\lambda x} \qquad for \quad 0 \leq x < \infty$$

where $\lambda > 0$ then $X$ is defined to have an (negative) exponential distribution. Now this random variable $X$ we have

$$E[X] = \frac{1}{\lambda} \qquad and \qquad var[X] = \frac{1}{\lambda^2}$$

**Pareto-Levy or Stable Distributions**

The stable distributions are a natural generalization of the normal in that, as their name suggests, they are stable under addition, i.e. a sum of stable random variables is also a random variable of the same type. However, nonnormal stable distributions have more probability mass in the tail areas than the normal. In fact, the nonnormal stable distributions are so fat-tailed that their variance and all higher moments are infinite.

Closed form expressions for the density functions of stable random variables are available for only the cases of normal and Cauchy.

If a random variable $X$ has a density function given by:

$$f_X(x) = f_X(x; \gamma, \delta) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - \delta)^2} \qquad for \quad -\infty < x < \infty$$

where $-\infty < \delta < \infty$ and $0 < \gamma < \infty$, then $X$ is defined to have a **Cauchy** distribution. Notice that for this random variable even the mean is infinite.

**Normal or Gaussian:**

We say that $X \backsim N[\mu, \sigma^2]$ then

$$f\left(x|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$
$$E(X) = \mu, \quad Var(X) = \sigma^2.$$

The distribution is symmetric about $\mu$, it is also unimodal and positive everywhere. Notice

$$\frac{X - \mu}{\sigma} = Z \backsim N[0, 1]$$

is the standard normal distribution.

**Lognormal Distribution**

Let $X$ be a positive random variable, and let a new random variable $Y$ be defined as $Y = \log X$. If $Y$ has a normal distribution, then $X$ is said to have a lognormal distribution. The density function of a lognormal distribution is given by

$$f_X(x;\mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}} \qquad for \quad 0 < x < \infty$$

where $\mu$ and $\sigma^2$ are parameters such that $-\infty < \mu < \infty$ and $\sigma^2 > 0$. We haven

$$E[X] = e^{\mu+\frac{1}{2}\sigma^2} \qquad and \qquad var[X] = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$$

Notice that if $X$ is lognormally distributed then

$$E[\log X] = \mu \qquad and \qquad var[\log X] = \sigma^2$$

**Gamma-$\chi^2$**

$$f\left(x|\alpha,\beta\right) = \frac{1}{\Gamma\left(\alpha\right)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad 0 < x < \infty, \quad \alpha,\beta > 0$$

$\alpha$ is shape parameter, $\beta$ is a scale parameter. Here $\Gamma\left(\alpha\right) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt$ is the Gamma function, $\Gamma\left(n\right) = n!$. The $\chi^2_k$ is when $\alpha = k$, and $\beta = 1$.

Notice that we can approximate the Poisson and Binomial functions by the normal, in the sense that if a random variable $X$ is distributed as Poisson with parameter $\lambda$, then $\frac{X-\lambda}{\sqrt{\lambda}}$ is distributed approximately as standard normal. On the other hand if $Y \sim Binomial(n,p)$ then $\frac{Y-np}{\sqrt{np(1-p)}} \sim N(0,1)$.

The standard normal is an important distribution for another reason, as well. Assume that we have a sample of n independent random variables, $x_1, x_2, ..., x_n$, which are coming from the same distribution with mean $m$ and variance $s^2$, then we have the following:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - m}{s} \sim N(0,1)$$

This is the well known **Central Limit Theorem** for independent observations.

## 4.1 Multivariate Random Variables

We now consider the extension to multiple r.v., i.e.,

$$X = (X_1, X_2, .., X_k) \in \mathbb{R}^k$$

The joint pmf, $f_X(x)$, is a function with

$$P(X \in A) = \sum_{x \in A} f_X(x)$$

The joint pdf, $f_X(x)$, is a function with

$$P(X \in A) = \int_{x \in A} f_X(x)dx$$

This is a multivariate integral, and in general difficult to compute. If $A$ is a rectangle $A = [a_1, b_1] \times ... \times [a_k, b_k]$, then

$$\int_{x \in A} f_X(x)dx = \int_{a_k}^{b_k} ... \int_{a_1}^{b_1} f_X(x)dx_1..dx_k$$

The joint c.d.f. is defined similarly

$$F_X(x) = \sum_{z_1 \leq x_1, ..., z_k \leq x_k} f_X(z_1, z_2, ..., z_k)$$

$$F_X(x) = P(X_1 \leq x_1, ..., X_k \leq x_k) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_X(z_1, z_2, ..., z_k)dz_1..dz_k$$

The multivariate c.d.f. has similar coordinate-wise properties to a univariate c.d.f. For continuously differentiable c.d.f.'s

$$f_X(x) = \frac{\partial^k F_X(x)}{\partial x_1 \partial x_2..\partial x_k}$$

### *4.1.1 Conditional Distributions and Independence*

We defined conditional probability $P(A|B) = P(A \cap B)/P(B)$ for events with $P(B) \neq 0$. We now want to define conditional distributions of $Y|X$. In the discrete case there is no problem

$$f_{Y|X}(y|x) = P(Y = y|X = x) = \frac{f(y, x)}{f_X(x)}$$

when the event $\{X = x\}$ has nonzero probability. Likewise we can define

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) = \frac{\sum_{Y \leq y} f(y,x)}{f_X(x)}$$

Note that $f_{Y|X}(y|x)$ is a density function and $F_{Y|X}(y|x)$ is a c.d.f.

1) $f_{Y|X}(y|x) \geq 0$ for all $y$

2) $\sum_y f_{Y|X}(y|x) = \frac{\sum_y f(y,x)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1$

In the continuous case, it appears a bit anomalous to talk about the $P(y \in A|X = x)$, since $\{X = x\}$ itself has zero probability of occurring. Still, we define the conditional density function

$$f_{Y|X}(y|x) = \frac{f(y,x)}{f_X(x)}$$

in terms of the joint and marginal densities. It turns out that $f_{Y|X}(y|x)$ has the properties of p.d.f.

1) $f_{Y|X}(y|x) \geq 0$

2) $\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = \frac{\int_{-\infty}^{\infty} f(y,x)dy}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$

We can define Expectations within the conditional distribution

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \frac{\int_{-\infty}^{\infty} y f(y,x)dy}{\int_{-\infty}^{\infty} f(y,x)dy}$$

and higher moments of the conditional distribution

### *4.1.2 Independence*

We say that $Y$ and $X$ are independent (denoted by $\perp\!\!\!\perp$) if

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

for all events $A$, $B$, in the relevant sigma-algebras. This is equivalent to the cdf's version which is simpler to state and apply.

$$F_{YX}(y,x) = F(y)F(x)$$

In fact, we also work with the equivalent density version

$$\begin{aligned} f(y,x) &= f(y)f(x) \quad for \quad all \quad y,x \\ f_{Y|X}(y|x) &= f(y) \quad for \quad all \quad y \\ f_{X|Y}(x|y) &= f(x) \quad for \quad all \quad x \end{aligned}$$

If $Y \perp\!\!\!\perp X$, then $g(X) \perp\!\!\!\perp h(Y)$ for any measurable functions $g$, and $h$.

We can generalise the notion of independence to multiple random variables. Thus $Y$, $X$, and $Z$ are mutually independent if:

$$\begin{aligned} f(y,x,z) &= f(y)f(x)f(z) \\ f(y,x) &= f(y)f(x) \quad for \quad all \quad y,x \\ f(x,z) &= f(x)f(z) \quad for \quad all \quad x,z \\ f(y,z) &= f(y)f(z) \quad for \quad all \quad y,z \end{aligned}$$

for all $y, x, z$.

### *4.1.3 Examples of Multivariate Distributions*

**Multivariate Normal**

We say that $X\left(X_1, X_2, ..., X_k\right) \backsim MVN_k\left(\mu, \Sigma\right)$, when

$$f_X\left(x|\mu,\Sigma\right) = \frac{1}{\left(2\pi\right)^{k/2}\left[\det\left(\Sigma\right)\right]^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^{/}\Sigma^{-1}\left(x-\mu\right)\right)$$

where $\Sigma$ is a $k \times k$ covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & ... & \sigma_{1k} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & \sigma_{kk} \end{pmatrix}$$

and $\det\left(\Sigma\right)$ is the determinant of $\Sigma$.

**Theorem 8** *(a) If* $X \backsim MVN_k(\mu, \Sigma)$ *then* $X_i \backsim N(\mu_i, \sigma_{ii})$ *(this is shown by integration of the joint density with respect to the other variables).*
*(b) The conditional distributions* $X = (X_1, X_2)$ *are Normal too*

$$f_{X_1|X_2}(x_1|x_2) \backsim N\left(\mu_{X_1|X_2}, \Sigma_{X_1|X_2}\right)$$

*where*

$$\mu_{X_1|X_2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \quad \Sigma_{X_1|X_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

*(c) Iff* $\Sigma$ *diagonal then* $X_1, X_2, .., X_k$ *are mutually independent. In this case*

$$\begin{aligned} \det(\Sigma) &= \sigma_{11}\sigma_{22}..\sigma_{kk} \\ -\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) &= -\frac{1}{2}\sum_{j=1}^{k}\frac{(x_j-\mu_j)^2}{\sigma_{jj}} \end{aligned}$$

*so that*

$$f_X(x|\mu, \Sigma) = \prod_{j=1}^{k}\frac{1}{\sqrt{2\pi\sigma_{jj}}}\exp\left(-\frac{1}{2}\frac{(x_j-\mu_j)^2}{\sigma_{jj}}\right)$$

*4.1.4 More on Conditional Distributions*

We now consider the relationship between two, or more, r.v. when they are not independent. In this case, conditional density $f_{Y|X}$ and c.d.f. $F_{Y|X}$ is in general varying with the conditioning point $x$. Likewise for conditional mean $E(Y|X)$, conditional median $M(Y|X)$, conditional variance $V(Y|X)$, conditional cf $E\left(e^{itY}|X\right)$, and other functionals, all of which characterize the relationship between $Y$ and $X$. Note that this is a directional concept, unlike covariance, and so for example $E(Y|X)$ can be very different from $E(X|Y)$.

**Regression Models:**

We start with random variable $(Y, X)$. We can write for any such random variable

$$Y = \overbrace{\underbrace{E(Y|X)}_{systematic\ \ part}}^{m(X)} + \overbrace{\underbrace{Y - E(Y|X)}_{\text{rand}\,om\ \ part}}^{\varepsilon}$$

By construction $\varepsilon$ satisfies $E(\varepsilon|X) = 0$, but $\varepsilon$ is not necessarily independent of $X$. For example, $Var(\varepsilon|X) = Var(Y - E(Y|X)|X) = Var(Y|X) = \sigma^2(X)$ can be expected to vary with $X$ as much as $m(X) = E(Y|X)$. A convenient and popular simplification is to assume that

$$\begin{aligned} E(Y|X) &= \alpha + \beta X \\ Var(Y|X) &= \sigma^2 \end{aligned}$$

For example, in the bivariate normal distribution $Y|X$ has

$$\begin{aligned} E(Y|X) &= \mu_Y + \rho_{YX}\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \\ Var(Y|X) &= \sigma_Y^2\left(1 - \rho_{YX}^2\right) \end{aligned}$$

and in fact $\varepsilon \perp\!\!\!\perp X$.

We have the following result about conditional expectations

**Theorem 9** *(1)* $E(Y) = E[E(Y|X)]$
*(2)* $E(Y|X)$ *minimizes* $E\left[(Y - g(X))^2\right]$ *over all measurable functions* $g(\cdot)$
*(3)* $Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$

**Proof.** (1) Write $f_{YX}(y,x) = f_{Y|X}(y|x) f_X(x)$ then we have $E(Y) = \int y f_Y(y)dy =$
$\int y\left(\int f_{YX}(y,x)dx\right)dy = \int y\left(\int f_{Y|X}(y|x) f_X(x)\,dx\right)dy =$

$= \int\left(\int y f_{Y|X}(y|x)\,dy\right) f_X(x)\,dx = \int [E(Y|X = x] f_X(x)\,dx = E(E(Y|X))$

(2) $E\left[(Y - g(X))^2\right] = E\left[[Y - E(Y|X) + E(Y|X) - g(X)]^2\right]$

$= E[Y - E(Y|X)]^2 + 2E[[Y - E(Y|X)][E(Y|X) - g(X)]] + E[E(Y|X) - g(X)]^2$

as now $E(YE(Y|X)) = E\left[(E(Y|X))^2\right]$, and $E(Yg(X)) = E(E(Y|X)g(X))$ we get that $E\left[(Y - g(X))^2\right] = E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 \geq E[Y - E(Y|X)]^2$.

(3) $Var(Y) = E[Y - E(Y)]^2 = E[Y - E(Y|X)]^2 + E[E(Y|X) - E(Y)]^2$

$+2E[[Y - E(Y|X)][E(Y|X) - E(Y)]]$

The first term is $E[Y - E(Y|X)]^2 = E\{E\left[[Y - E(Y|X)]^2|X\right]\} = E[Var(Y|X)]$

The second term is $E[E(Y|X) - E(Y)]^2 = Var[E(Y|X)]$

The third term is zero as $\varepsilon = Y - E(Y|X)$ is such that $E(\varepsilon|X) = 0$, and $E(Y|X) - E(Y)$ is measurable with respect to $X$. ■

**Covariance**

$$Cov\left(X,Y\right)=E\left[X-E\left(X\right)\right]E\left[Y-E\left(Y\right)\right]=E\left(XY\right)-E\left(X\right)E\left(Y\right)$$

Note that if $X$ or $Y$ is a constant then $Cov\left(X,Y\right)=0$. Also

$$Cov\left(aX+b,cY+d\right)=acCov\left(X,Y\right)$$

An alternative measure of association is given by the **correlation coefficient**

$$\rho_{XY}=\frac{Cov\left(X,Y\right)}{\sigma_X\sigma_Y}$$

Note that

$$\rho_{aX+b,cY+d}=sign\left(a\right)\times sign\left(c\right)\times\rho_{XY}$$

If $E\left(Y|X\right)=a=E\left(Y\right)$ almost surely, then $Cov\left(X,Y\right)=0$. Also if $X$ and $Y$ are independent r.v. then $Cov\left(X,Y\right)=0$.

Both the covariance and the correlation of random variables $X$ and $Y$ are measures of a linear relationship of $X$ and $Y$ in the following sense. $cov[X,Y]$ will be positive when $(X-\mu_X)$ and $(Y-\mu_Y)$ tend to have the same sign with high probability, and $cov[X,Y]$ will be negative when $(X-\mu_X)$ and $(Y-\mu_Y)$ tend to have opposite signs with high probability. The actual magnitude of the $cov[X,Y]$ does not much meaning of how strong the linear relationship between $X$ and $Y$ is. This is because the variability of $X$ and $Y$ is also important. The correlation coefficient does not have this problem, as we divide the covariance by the product of the standard deviations. Furthermore, the correlation is unitless and $-1\leq\rho\leq1$.

The properties are very useful for evaluating the **expected return** and **standard deviation** of a **portfolio**. Assume $r_a$ and $r_b$ are the returns on assets $A$ and $B$, and their variances are $\sigma_a^2$ and $\sigma_b^2$, respectively. Assume that we form a portfolio of the two assets with weights $w_a$ and $w_b$, respectively. If the correlation of the returns of these assets is $\rho$, find the expected return and standard deviation of the portfolio.

If $R_p$ is the return of the portfolio then $R_p = w_a r_a + w_b r_b$. The expected portfolio return is $E[R_p] = w_a E[r_a] + w_b E[r_b]$. The variance of the portfolio is $var[R_p] = var[w_a r_a + w_b r_b] = E[(w_a r_a + w_b r_b)^2] - (E[w_a r_a + w_b r_b])^2 =$

$= w_a^2 E[r_a^2] + w_b^2 E[r_b^2] + 2w_a w_b E[r_a r_b]$

$-w_a^2 (E[r_a])^2 - w_b^2 (E[r_b])^2 - 2w_a w_b E[r_a]E[r_b] =$

$= w_a^2 \{E[r_a^2] - (E[r_a])^2\} + w_b^2 \{E[r_b^2] - (E[r_b])^2\} + 2w_a w_b \{E[r_a r_b] - E[r_a]E[r_b]\}$

$= w_a^2 var[r_a] + w_b^2 var[r_b] + 2w_a w_b cov[r_a, r_b]$ or $= w_a^2 \sigma_a^2 + w_b^2 \sigma_b^2 + 2w_a w_b \rho \sigma_a \sigma_b$

In a vector format we have:

$$E[R_p] = \begin{pmatrix} w_a & w_b \end{pmatrix} \begin{pmatrix} E[r_a] \\ E[r_b] \end{pmatrix} \text{ and}$$

$$var[R_p] = \begin{pmatrix} w_a & w_b \end{pmatrix} \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \begin{pmatrix} w_a \\ w_b \end{pmatrix}$$

From the above example we can see that $var[aX+bY] = a^2 var[X] + b^2 var[Y] + 2abcov[X,Y]$ for random variables $X$ and $Y$ and constants $a$ and $b$. In fact we can generalize the formula above for several random variables $X_1, X_2, ..., X_n$ and constants $a_1, a_2, a_3, ..., a_n$ i.e. $var[a_1 X_1 + a_2 X_2 + ...a_n X_n] = \sum_{i=1}^{n} a_i^2 var[X_i] + 2\sum_{i<j}^{n} a_i a_j cov[X_i, X_j]$

## 4.2 Inequalities

This section gives some inequalities that are useful in establishing a variety of probabilistic results.

### *4.2.1 Markov*

Let $Y$ be a random variable and consider a function $g(.)$ such that $g(y) \geq 0$ for all $y \in \mathbb{R}$. Assume that $E[g(Y)]$ exists. Then

$$P[g(Y) \geq c] \leq c^{-1} E[g(Y)], \quad for \quad all \quad c > 0.$$

PROOF:

Assume that Y is continuous random variable (the discrete case follows anal-

ogously) with p.d.f. $f(.)$. Define $A_1 = \{y|g(y) \geq c\}$ and $A_2 = \{y|g(y) < c\}$. Then

$$\begin{aligned} E[g(Y)] &= \int_{A_1} g(y) f(y) dy + \int_{A_2} g(y) f(y) dy \\ &\geq \int_{A_1} g(y) f(y) dy \geq \int_{A_1} cf(y) dy = cP[g(Y) \geq c]. \end{aligned}$$

■

*4.2.2 Chebychev's Inequality*

$$P[|X - E(X)| \geq \eta] \leq \frac{Var(X)}{\eta^2}$$

or alternatively

$$P\left[|X - E(X)| \geq r\sqrt{Var(X)}\right] \leq \frac{1}{r^2}$$

PROOF:

To prove the above, assume that $E(X) = 0$ and compare $1(|X| \geq \eta)$ with $\frac{X^2}{\eta^2}$. Clearly $1(|X| \geq \eta) \leq \frac{X^2}{\eta^2}$ and it follows that $E[1(|X| \geq \eta)] \leq \frac{E(X^2)}{\eta^2} \Rightarrow P[|X| \geq \eta] \leq \frac{Var(X)}{\eta^2}$. Alternatively, apply Markov's inequality by setting $g(y) = [x - E(X)]^2$ and $c = r^2 Var(X)$.■

*4.2.3 Minkowski*

Let $Y$ and $Z$ be random variables such that $[E(|Y|^\alpha)] < \infty$ and $[E(|Z|^\alpha)] < \infty$ for some $1 \leq \alpha < \infty$. Then

$$[E(|Y + Z|^\alpha)]^{1/\alpha} \leq [E(|Y|^\alpha)]^{1/\alpha} + [E(|Z|^\alpha)]^{1/\alpha}$$

For $\alpha = 1$ we have the triangular inequality

*4.2.4 Triangle*

$$E|X + Y| \leq E|X| + E|Y|.$$

*4.2.5 Cauchy-Schwarz*

$$\begin{aligned} E^2(XY) &\leq E(X)^2 E(Y)^2 \\ (\sum a_j b_j)^2 &\leq \left(\sum a_j^2\right)\left(\sum b_j^2\right) \end{aligned}$$

PROOF:

Let $0 \leq h(t) = E\left[(tX - Y)^2\right] = t^2 E(X^2) + E(Y^2) - 2tE(XY)$. Then the function $h(t)$ is a quadratic function in $t$ which is increasing as $t \to \pm\infty$. It has a unique minimum at $h^{/}(t) = 0 \Rightarrow 2tE(X^2) - 2E(XY) = 0 \Rightarrow t = \frac{E(XY)}{E(X^2)}$. Hence $0 \leq h\left(\frac{E(XY)}{E(X^2)}\right) \Rightarrow E^2(XY) \leq E(X)^2 E(Y)^2$.■

*4.2.6 Hölder's Inequality*

For any $p$, $q$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$ we have

$$E|XY| \leq (E|X|^p)^{\frac{1}{p}} (E|Y|^q)^{\frac{1}{q}}$$

In fact the Cauchy-Schwarz inequality corresponds for $p = q = 2$.

*4.2.7 Jensen Inequality*

Let $X$ be a random variable with mean $E[X]$, and let $g(.)$ be a convex function. Then

$$E[g(X)] \geq g(E[X]).$$

Now a continuous function $g(.)$ with domain and counterdomain the real line is called **convex** if for any $x_0$ on the real line, there exist a line which goes through the point $(x_0, g(x_0))$ and lies on or under the graph of the function $g(.)$. Also if $g^{//}(x_0) \geq 0$ then $g(.)$ is convex.