

University of Sheffield

Department of Probability & Statistics

Stochastic Processes

PAS6052 & PAS401

Autumn 2008

# Chapter 1

## Probability and Measure

### 1.1 Basic Concepts

Probability theory is the mathematical study of chance phenomena. In the first half of the twentieth century it was realised that it could be given a firm mathematical foundation as a special case of a more general theory called *measure theory*.

Measure theory is an abstract theory which is derived from properties of physical phenomena which we measure, such as mass, weight, length, volume, electric charge or energy. Probability fits into this context in that when we assign probabilities, we measure how likely an event is to occur on a scale from 0 to 1.

Measure theory is a deep and extensive subject which really deserves a course to itself. It also serves as a foundation for the modern theory of integration. In this course, we will not go very deeply into the technical side of it. We will need to know the language, some of the conceptual apparatus and an appreciation of how probability fits into it.

We begin our survey of measure theory with a given set  $S$  which is the universe we are interested in. Abstract measure theory assigns a “weight” to certain subsets of  $S$ . There is already a problem here in that some sets in  $S$  may be too badly behaved to have a “weight”. So we have to distinguish those sets which we are able to deal with.

**Definition.** A  $\sigma$ -algebra  $\Sigma$  is a collection of subsets of  $S$  which has the following properties:

S(i)  $S \in \Sigma$ .

S(ii) If  $A \in \Sigma$  then  $A^c \in \Sigma$ .

S(iii) If  $(A_n, n \in \mathbb{N})$  is a sequence of sets where each  $A_n \in \Sigma$  then  $\bigcup_{n \in \mathbb{N}} A_n \in \Sigma$ .

Facts about  $\sigma$ -algebras:

- By S(i) and S(ii),  $\emptyset = S^c \in \Sigma$ .
- We have seen in S(iii) that infinite unions of sets in  $\Sigma$  are themselves in  $\Sigma$ . The same is true of finite unions - to see this just take  $(A_n, n \in \mathbb{N})$  to be such that  $A_m = \emptyset$  for all  $m \geq N$ , where  $N \in \mathbb{N}$  is fixed.
- $\Sigma$  is also closed under finite intersections. To see this use de Morgan's law to write

$$A_1 \cap A_2 \cap \cdots \cap A_n = (A_1^c \cup A_2^c \cup \cdots \cup A_n^c)^c.$$

- $\Sigma$  is closed under set theoretic differences  $A - B$ , since (by definition)  $A - B = A \cap B^c$ .

Examples of  $\Sigma$ -algebras:

1. The set of all subsets of  $S$  is a  $\sigma$ -algebra called the *power set* of  $S$  and denoted by  $\mathcal{P}(S)$ . If  $S$  is finite with  $n$  elements then  $\mathcal{P}(S)$  has  $2^n$  elements.
2. Consider the real line  $\mathbb{R}$ .  $\mathcal{P}(\mathbb{R})$  is a  $\sigma$ -algebra but it is too big for our purposes. Some sets in  $\mathcal{P}(\mathbb{R})$  are quite wild. We want to restrict ourselves to sets which can have a "length" in some reasonable sense. Certainly we can measure the length of an interval  $(a, b)$ . It is of course  $b - a$ .

We define  $\mathcal{B}(\mathbb{R})$  - the *Borel  $\sigma$ -algebra* of  $\mathbb{R}$  to be the smallest  $\sigma$ -algebra containing all open intervals  $(a, b)$ . It is a fact that  $\mathcal{B}(\mathbb{R})$  does not contain all subsets of  $\mathbb{R}$ .  $\mathcal{B}(\mathbb{R})$  certainly contains all finite and infinite unions of open and closed intervals as well as isolated points. It is hard to construct a set which is not in  $\mathcal{B}(\mathbb{R})$ . Sets in  $\mathcal{B}(\mathbb{R})$  are often called *Borel sets*.

3. By taking Cartesian products of open intervals we can similarly construct  $\mathcal{B}(\mathbb{R}^n)$  - the Borel  $\sigma$ -algebra of  $\mathbb{R}^n$ , for any  $n \geq 2$ .

Now suppose we have a set  $S$  and we've fixed a  $\sigma$ -algebra which we're going to work with. Now we are going to learn how to measure sets in  $\Sigma$ .

**Definition.** A measure on  $(S, \Sigma)$  is a mapping  $m$  from  $\Sigma$  to  $[0, \infty]$  which satisfies

M(i)  $m(\emptyset) = 0$ ,

M(ii) ( $\sigma$ -additivity) If  $(A_n, n \in \mathbb{N})$  is a sequence of sets where each  $A_n \in \Sigma$  and if these sets are mutually disjoint, i.e.  $A_n \cap A_m = \emptyset$  if  $m \neq n$ , then

$$m\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} m(A_n).$$

The *total mass* of a measure is  $m(S)$  so  $0 \leq m(S) \leq \infty$ .

### Examples

1. *Counting measure* on  $(S, \mathcal{P}(S))$  where  $S$  is finite is defined by

$$m(A) = \#(A) = \text{number of elements in the set } A.$$

It has total mass  $\#(S)$ .

2. *Lebesgue measure* on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  captures the notion of length. It is the unique measure for which

$$m((a, b)) = b - a,$$

for each  $-\infty < a < b < \infty$ . It has total mass  $\infty$ .

3. *Lebesgue measure* on  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  is the unique measure for which

$$m((a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n)) = (b_1 - a_1)(b_2 - a_2) \cdots (b_n - a_n).$$

In the cases where  $n = 2, 3$  we capture the notions of area and volume, respectively.

In general, if  $m$  is a measure on  $(S, \Sigma)$ , we call  $(S, \Sigma, m)$  a *measure space*.

**Definition** A *probability measure* on  $(S, \Sigma)$  is a measure which has total mass 1.

In the case of probability measures we usually use a different notation and write  $\Omega$  instead of  $S$ ,  $\mathcal{F}$  instead of  $\Sigma$  and  $P$  instead of  $m$ . The triple  $(\Omega, \mathcal{F}, P)$  is called a *probability space*.

$\Omega$  is called the *sample space* and elements of  $\Omega$  are called *outcomes*.

Sets in the  $\sigma$ -algebra  $\mathcal{F}$  are called *events*.

If  $A \in \mathcal{F}$ , then  $P(A)$  is called the *probability* of the event  $A$ . We always have  $0 \leq P(A) \leq 1$ .

**Example** If  $S$  is finite, we can define a probability measure on  $(S, \mathcal{P}(S))$  by

$$P(A) = \frac{\#(A)}{\#(S)} \text{ for each } A \in \mathcal{P}(S).$$

## 1.2 Measurable Functions and Random Variables

We begin by recalling some set theoretic ideas. Suppose that  $S_1$  and  $S_2$  are sets and  $f : S_1 \rightarrow S_2$  is a mapping (function). If  $A \subseteq S_1$  we define

$$f(A) = \{f(x), x \in A\},$$

so  $f(A) \subseteq S_2$ .

If  $B \subseteq S_2$ , we define

$$f^{-1}(B) = \{x \in S_1, f(x) \in B\},$$

so  $f^{-1}(B) \subseteq S_1$ . It is important to appreciate that  $f^{-1}(B)$  is the name of a set. It can always be found, at least in principle, whether or not the function  $f$  is invertible.

e.g. consider  $f : \mathbb{R} \rightarrow \mathbb{R}$  given by  $f(x) = x + 1$ .  $f$  is invertible and  $f^{-1}(x) = x - 1$ . We have  $f([0, 1]) = [1, 2]$  and  $f^{-1}([0, 1]) = [-1, 0]$ . On the other hand  $f(x) = \sin(x)$  is not invertible as a mapping from  $\mathbb{R}$  to  $\mathbb{R}$ . We have  $f(\{0\}) = \{0\}$  and  $f^{-1}(\{0\}) = \{n\pi, n \in \mathbb{Z}\}$ .

Just as the notion of  $\sigma$ -algebra allows us to focus only on well-behaved sets so the concept of “measurability” enables us to identify a nice class of functions to work with.

Suppose that  $\Sigma_1$  and  $\Sigma_2$  are given  $\sigma$ -algebras of  $S_1$  and  $S_2$ , respectively.

**Definition** A function  $f$  from  $S_1$  to  $S_2$  is said to be  $(\Sigma_1 - \Sigma_2)$  measurable if for all  $B \in \Sigma_2, f^{-1}(B) \in \Sigma_1$ . In future, we will usually just say that a function is measurable when the specific  $\sigma$ -algebras involved are clear.

All the functions you’ve ever met (or are likely to meet) are measurable. Its quite hard to construct one that isn’t.

In the case where  $S_1 = S_2 = \mathbb{R}$  and  $\Sigma_1 = \Sigma_2 = \mathcal{B}(\mathbb{R})$ , measurable functions are sometimes called *Borel functions*.  $f(x) = x, x^n, e^x, \log(x), \sin(x)$  are all measurable. In fact any continuous function is measurable. Sums, products, inverses (when they exist) and compositions of measurable functions remain measurable.

If we have a probability space  $(\Omega, \mathcal{F}, P)$  a *random variable* is defined to be a measurable function  $X$  from  $(\Omega, \mathcal{F})$  to  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . A *random vector* is a measurable function from  $(\Omega, \mathcal{F})$  to  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ .

If  $X$  is a random variable and  $g : \mathbb{R} \rightarrow \mathbb{R}$  is a Borel function then we can form the function  $g(X) = g \circ X$  from  $\Omega$  to  $\mathbb{R}$ . This is also a measurable

function and so is a random variable, so e.g.  $X^n, \sin(X), e^X$  are all random variables.

Why measurability ? Suppose  $B = (a, b)$ , so  $X^{-1}(B) = \{\omega \in \Omega; X(\omega) \in (a, b)\}$ . If  $X$  is e.g. the price of a stock next week, we would like to be able to assign a probability to the event that it lies between  $a$  and  $b$ . But we can only do this if  $X^{-1}(B)$  is a measurable set, i.e.  $X^{-1}(B) \in \mathcal{F}$  as these are the only sets that we can apply  $P$  to. This forces us to require random variables to be measurable functions.

Taking this idea further, we would like to be able to compute

$$\begin{aligned} P(X \in B) &= P(X^{-1}(B)) \\ &= P(\{\omega \in \Omega; X(\omega) \in B\}). \end{aligned}$$

Define a new measure on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  which is denoted  $p_X$  by

$$p_X(B) = P(X^{-1}(B)),$$

i.e.  $p_X = P \circ X^{-1}$ .  $p_X$  is called the *probability law* or *probability distribution* of the random variable  $X$ . In the case where the range of  $X$  is finite or countable,  $p_X$  is called a *probability mass function*.

e.g.  $X$  is a Bernoulli random variable with parameter  $p$  with  $0 < p < 1$ . Its range is  $\{0, 1\}$  and

$$p_X(\{0\}) = 1 - p, \quad p_X(\{1\}) = p.$$

$X$  has a Poisson distribution with parameter  $\lambda > 0$ . The range of  $X$  is  $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$  and for each  $n \in \mathbb{Z}_+$ ,

$$p_X(\{n\}) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

Returning to the general case, the *cumulative distribution function*  $F_X$  of  $X$  is a Borel function from  $\mathbb{R}$  to  $[0, 1]$ . It is defined by

$$F_X(x) = p_X((-\infty, x]) = P(X \leq x).$$

$F_X$  is a right continuous, increasing function for which  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  and  $\lim_{x \rightarrow \infty} F_X(x) = 1$ .

$X$  has a *probability density function* (or *pdf* for short) if there exists a non-negative Borel function  $f_X : \mathbb{R} \rightarrow [0, \infty]$  for which

$$F_X(x) = \int_{-\infty}^x f_X(y) dy.$$

e.g. Uniform distribution on  $(a, b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Normal (or Gaussian) distribution with parameters  $\mu$  and  $\sigma^2$ :

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

In this case we write  $X \sim N(\mu, \sigma^2)$ . The standard normal is  $X \sim N(0, 1)$ .

Two random variables  $X$  and  $Y$  are said to be *identically distributed* if they have the same probability law, i.e.  $p_X = p_Y$ . In this case we sometimes write  $X \stackrel{d}{=} Y$ .

### 1.2.1 Sub $\sigma$ -algebras and Independence

Let  $\Sigma$  be a  $\sigma$ -algebra of subsets of  $S$ .  $\Sigma'$  is a *sub- $\sigma$ -algebra* of  $\Sigma$  if

1.  $\Sigma'$  is a  $\sigma$ -algebra of subsets of  $S$ .
2.  $A \in \Sigma' \Rightarrow A \in \Sigma$ .

In this case we write  $\Sigma' \subseteq \Sigma$ . For example, if  $\mathcal{T}$  is the *trivial  $\sigma$ -algebra* defined by  $\mathcal{T} = \{\emptyset, S\}$ , then  $\mathcal{T} \subseteq \Sigma$  for any  $\sigma$ -algebra of subsets of  $S$ . Sub- $\sigma$ -algebras are important in probability theory. They describe “partial information”.

Now let  $(\Omega, \mathcal{F}, P)$  be a probability space and suppose that we are given two sub- $\sigma$ -algebras  $\mathcal{G}$  and  $\mathcal{H}$  of  $\mathcal{F}$ . We say that  $\mathcal{G}$  and  $\mathcal{H}$  are *independent* if

$$P(A \cap B) = P(A).P(B)$$

for all  $A \in \mathcal{G}$  and all  $B \in \mathcal{H}$ .

Let  $X$  be a random variable. It can be shown that the collection of all sets of the form  $X^{-1}(A)$ , where  $A \in \mathcal{B}(\mathbb{R})$  forms a sub- $\sigma$ -algebra of  $\mathcal{F}$  and we denote this by  $\mathcal{F}_X$  (or sometimes  $\sigma(X)$ .) It is called the *sub- $\sigma$ -algebra generated by  $X$* . Two random variables  $X$  and  $Y$  are said to be *independent* if  $\mathcal{F}_X$  and  $\mathcal{F}_Y$  are independent in the sense given above, i.e. for all  $A, B \in \mathcal{B}(\mathbb{R})$ :

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

## 1.3 Integration and Expectation

### 1.3.1 Lebesgue Integration

Many of you will have met a rigorous approach to integration through its nineteenth century version - the *Riemann integral*. If  $f$  is a bounded function on an interval  $[a, b]$ , we seek to define  $\int_a^b f(x)dx$ . Define a partition  $a = x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n = b$  and construct the *Riemann sums*  $\sum_{j=1}^n f(t_j)(x_j - x_{j-1})$ , where each  $x_{j-1} < t_j < x_j$ . By taking finer and finer partitions, we can investigate whether the Riemann sums converge to a suitable limit, which (if it exists) is the integral we seek.

For applications to probability and much of modern mathematics, this approach isn't sufficiently refined. We will need the *Lebesgue integral* which was developed in the first half of the twentieth century. Let  $(S, \Sigma, m)$  be a measure space and fix a measurable function  $f : S \rightarrow \mathbb{R}$ . We want to be able to define  $\int_S f(x)m(dx)$  - at least for suitably well-behaved  $f$ .

Lebesgue's approach to this problem is as follows. The construction is split up into four steps.

*Step 1.* First consider the case where  $A \in \Sigma$ . Define the *indicator function* of  $A$ , which is denoted  $1_A$  by the prescription

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

$1_A$  is easily seen to be measurable. We make the obvious definition

$$\int_S 1_A(x)m(dx) = m(A). \quad (1.3.1)$$

*Step 2.* We say that  $f$  is a *simple function* if for some  $n \in \mathbb{N}$ , there exist disjoint sets  $A_1, A_2, \dots, A_n$  and real numbers  $c_1, c_2, \dots, c_n$  such that

$$f = \sum_{j=1}^n c_j 1_{A_j}.$$

$f$  is again easily seen to be measurable. We define its integral by linear extension of the rule in step 1.

$$\int_S f(x)m(dx) = \sum_{j=1}^n c_j m(A_j).$$

*Step 3.* Now suppose that  $f$  is an arbitrary non-negative measurable function, i.e.  $f(x) \geq 0$  for all  $x \in S$ . In this case we approximate the integral from below using integrals of step functions as defined in step 2, i.e.

$$\int_S f(x)m(dx) = \sup \left\{ \int_S g(x)m(dx), g \text{ simple}, 0 \leq g \leq f \right\}.$$

With this definition,  $\int_S f(x)m(dx) \in [0, \infty]$ .

*Step 4.* For the final step we take  $f$  to be an arbitrary measurable function. We define the positive and negative parts of  $f$ , which we denote as  $f_+$  and  $f_-$  respectively by:

$$f_+(x) = \max\{f(x), 0\}, \quad f_-(x) = \max\{-f(x), 0\},$$

so both  $f_+$  and  $f_-$  are measurable and non-negative. We have

$$f = f_+ - f_-,$$

and using Step 3, we see that we can construct both  $\int_S f_+(x)m(dx)$  and  $\int_S f_-(x)m(dx)$ . Provided both of these are not infinite, we define

$$\int_S f(x)m(dx) = \int_S f_+(x)m(dx) - \int_S f_-(x)m(dx).$$

With this definition,  $\int_S f(x)m(dx) \in [-\infty, \infty]$ . We say that  $f$  is *integrable* if  $\int_S f(x)m(dx) \in (-\infty, \infty)$ . Clearly  $f$  is integrable if and only if each of  $f_+$  and  $f_-$  are. Since

$$|f| = f_+ + f_-,$$

it follows that  $f$  is integrable if and only if  $|f|$  is. Using this fact, the condition for integrability of  $f$  is often written

$$\int_S |f(x)|m(dx) < \infty.$$

We also have the useful inequality

$$\left| \int_S f(x)m(dx) \right| \leq \int_S |f(x)|m(dx). \quad (1.3.2)$$

We have defined the integral of  $f$  over the whole of  $S$ . If we want to restrict it to  $A \in \Sigma$ , we define

$$\int_A f(x)m(dx) = \int_S f(x)1_A(x)m(dx).$$

The integral has a number of natural properties. We list some of the most useful of these:

- (*Linearity*) If  $f$  and  $g$  are integrable then so is  $\alpha f + \beta g$  for any  $\alpha, \beta \in \mathbb{R}$  and

$$\int_S (\alpha f(x) + \beta g(x))m(dx) = \alpha \int_S f(x)m(dx) + \beta \int_S g(x)m(dx).$$

- (*Domination*) If  $g$  is measurable and  $f$  is integrable with  $0 \leq g \leq f$ , then  $g$  is integrable and

$$\int_S g(x)m(dx) \leq \int_S f(x)m(dx).$$

- (*The Triangle Inequality*) If  $f$  and  $g$  are integrable then

$$\int_S |f(x) + g(x)|m(dx) \leq \int_S |f(x)|m(dx) + \int_S |g(x)|m(dx).$$

- (*Set bounded integrability*) If  $A \in \Sigma$  with  $m(A) < \infty$  and  $f$  is bounded on  $A$ , i.e. there exists  $C > 0$  such that  $|f(x)| \leq C$  for all  $x \in A$ , then  $f1_A$  is integrable and

$$\int_A |f(x)|m(dx) \leq Cm(A).$$

One of the ways that Lebesgue integration improves on Riemann integration is that it gives us useful theorems on interchange of limits and integrals. One of the most useful of these is *Lebesgue's dominated convergence theorem*:

**Theorem 1.3.1 (Dominated Convergence)** *Let  $(f_n, n \in \mathbb{N})$  be a sequence of integrable functions which converges pointwise to  $f$ , i.e.  $\lim_{n \rightarrow \infty} f_n(x) = f(x)$  for all  $x \in S$ . If there exists an integrable function  $g \geq 0$  such that  $|f_n(x)| \leq g(x)$  for all  $n \in \mathbb{N}$  and all  $x \in S$  then  $f$  is integrable and*

$$\lim_{n \rightarrow \infty} \int_S f_n(x)m(dx) = \int_S f(x)m(dx).$$

If  $S = \mathbb{R}$ ,  $\Sigma = \mathcal{B}(\mathbb{R})$  and  $m$  is Lebesgue measure, we write  $\int_S f(x)m(dx)$  as  $\int_{\mathbb{R}} f(x)dx$ . Any function which is Riemann integrable will also be Lebesgue integrable (with the same value) but many more functions can be integrated in the Lebesgue sense.

### 1.3.2 Expectation

We now take  $(S, \Sigma, m)$  to be a probability space  $(\Omega, \mathcal{F}, P)$  and then our measurable function  $f$  becomes a random variable to be denoted  $X$ , as usual. We define the *expectation* of  $X$  (when it exists) by

$$\mathbb{E}(X) = \int_{\Omega} X(\omega)P(d\omega).$$

Of course this exists and is finite if and only if  $X$  is *integrable*, i.e.

$$\mathbb{E}(|X|) < \infty.$$

Note that we also have (by change of variable)

$$\mathbb{E}(X) = \int_{\mathbb{R}} xp_X(dx) = \int_{\mathbb{R}} xF_X(dx).$$

If  $X$  has a pdf  $f_X$ , we recover the familiar formula

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf_X(x)dx.$$

The formula (1.3.1) for integrating indicator functions in Step (1) gives us a nice direct link between expectation and probability:

$$\mathbb{E}(1_A) = P(A).$$

Let  $X$  and  $Y$  be integrable random variables, and  $\alpha, \beta \in \mathbb{R}$ . The following useful results are all direct translations of results about integrals given above. Make sure that you understand where these come from.

- If  $X(\omega) \geq 0$  for all  $\omega \in \Omega$ , then  $\mathbb{E}(X) \geq 0$ .
- $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$ .
- $\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$ .
- $\mathbb{E}(|X + Y|) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|)$ .

Another useful result is that if  $X$  and  $Y$  are integrable random variables that are independent, then they are *uncorrelated*, i.e.

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

In general, if  $X$  and  $Y$  are uncorrelated, we cannot assume that they are independent, except when  $X$  and  $Y$  are jointly normal.

If  $X^n$  is integrable, then  $\mathbb{E}(X^n)$  is called the  $n$ th moment of  $X$ . If it exists  $\mathbb{E}(X)$  is called the *mean* and it is sometimes denoted by  $\mu$  or  $\mu_X$ .  $\mathbb{E}((X - \mu)^2)$  is called the *variance* (again, when it exists) and it is often denoted by  $\sigma^2$  or  $\sigma_X^2$  or  $\text{Var}(X)$ .  $\sigma$  is called the *standard variation*. Note the useful (and easily derived) formula

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

If  $X$  and  $Y$  are independent and each has a finite variance, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \quad (1.3.3)$$

Beware that this formula is not true in general.

The Cauchy distribution is an example of a random variable which has no mean or variance. Its pdf is  $f(x) = \frac{1}{\pi(1+x^2)}$ .

Suppose that  $\mathbb{E}(e^{tX}) < \infty$  for all  $t \in \mathbb{R}$ . In this case the function  $\Phi_X(t) = \mathbb{E}(e^{tX})$  is called the *moment generating function* of  $X$ . Its name is justified by the fact that all moments exist in this case and the following calculation can be justified

$$\left. \frac{d^n}{dt^n} \mathbb{E}(e^{tX}) \right|_{t=0} = \mathbb{E} \left( \left. \frac{d^n}{dt^n} e^{tX} \right|_{t=0} \right) = \mathbb{E}(X^n).$$

In the general case, the *characteristic function*  $\Psi_X(t) = \mathbb{E}(e^{itX})$  always exists (Why ?) but of course the  $n$ th moment may not. When it does it can be found by differentiating  $\Psi_X$  in a similar manner to that just described.

### 1.3.3 Sets of Measure Zero

Let  $(S, \Sigma, m)$  be a measure space. A set  $A \in \Sigma$  has *measure zero* if  $m(A) = 0$ . Such sets play a subtle role in both measure theory and probability theory. Some quite complex sets can have measure zero, e.g. on the real line the natural numbers, integers and rational numbers all have Lebesgue measure zero (so only irrational numbers contribute to “length”). In fact so does any countable set. The famous Cantor set is an example of an uncountable set which has measure zero.

Lebesgue integration doesn't see sets of measure zero, i.e. if  $m(A) = 0$  then  $\int_A f(x)m(dx) = 0$ . It follows that two measurable functions  $f$  and  $g$  which have exactly the same values everywhere except on a set of measure zero will have the same integral.

e.g. Consider  $f : \mathbb{R} \rightarrow \mathbb{R}$  given by

$$f(x) = \begin{cases} 0 & \text{if } x \in \mathbb{Q} \\ 1 & \text{if } x \notin \mathbb{Q} \end{cases}$$

$$\int_a^b f(x)dx = \int_a^b 1dx = (b - a).$$

Note that this function (which is discontinuous at every point) does not even have an integral in the Riemann sense.

In measure theory, we identify functions  $f$  and  $g$  which agree except on set of measure zero and we write  $f = g$  (a.e.) where a.e. stands for “almost everywhere”.

In probability the same rules apply but we use a different language to express them. If  $f$  and  $g$  are random variables  $X$  and  $Y$ , we write  $X = Y$  (a.s.) when they only differ on a set of probability zero, where a.s. stands for “almost surely.”

Many theorems of measure theory and probability hold a.e. or a.s. Our strategy in this course will be to (in most cases) ignore these fine distinctions.

## 1.4 Auxiliary Topics in Probability Theory

### 1.4.1 Convergence of Random Variables

Let  $(X_n, n \in \mathbb{N})$  be a sequence of random variables. There are four different ways in which it can converge to a random variable  $X$ , i.e. four different ways of giving meaning to the idea of “ $\lim_{n \rightarrow \infty} X_n = X$ .”

(i) *almost sure convergence.* In this case the numbers  $X_n(\omega) \rightarrow X(\omega)$  as  $n \rightarrow \infty$  for all  $\omega \in \Omega$  except for a possible set of measure zero where convergence fails. If you have done a previous course on advanced analysis, then you can see that this is a slightly weaker notion to that of pointwise convergence of functions.

(ii) *convergence in mean square.* In this case we require that

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n - X|^2) = 0.$$

(iii) *convergence in probability.* Here we require that for all  $c > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - X| > c) = 0.$$

(iv) *convergence in distribution.* If  $F_n$  is the cdf of each  $X_n$  and  $F$  is the cdf of  $X$ , we require in this case that  $F_n(x) \rightarrow F(x)$  as  $n \rightarrow \infty$  for all  $x \in \mathbb{R}$ .

We have the following relationships between these modes of convergence:

almost sure convergence  $\Rightarrow$  convergence in probability  $\Rightarrow$  convergence in distribution.

convergence in mean square  $\Rightarrow$  convergence in probability  $\Rightarrow$  convergence in distribution.

There is no direct relationship between almost sure convergence and convergence in mean square.

## 1.4.2 Some Useful Inequalities

Inequalities play a very important role in advanced probability. Here are three particularly useful ones:

### Chebychev's Inequality

**Theorem 1.4.1 (Chebychev's Inequality)** *If  $X$  is a random variable with finite variance  $\sigma^2$  and mean  $\mu$  then for each  $c > 0$ :*

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

*Proof.* Let  $A = \{\omega \in \Omega; |X(\omega) - \mu| \geq c\}$ , then  $A \in \mathcal{F}$  and

$$\begin{aligned} \sigma^2 &= \mathbb{E}((X - \mu)^2) \\ &= \int_{\Omega} (X(\omega) - \mu)^2 P(d\omega) \\ &= \int_A (X(\omega) - \mu)^2 P(d\omega) + \int_{A^c} (X(\omega) - \mu)^2 P(d\omega) \\ &\geq \int_A (X(\omega) - \mu)^2 P(d\omega) \\ &\geq \int_A c^2 P(d\omega) = c^2 \int_A P(d\omega) = c^2 P(A). \end{aligned}$$

i.e.  $P(A) \leq \frac{\sigma^2}{c^2}$  and this is what we set out to prove.  $\square$

Note: In the next two inequalities we assume that all (functions of) random variables under consideration are integrable - so all expectations are finite numbers.

## The Cauchy-Schwarz Inequality

**Theorem 1.4.2 (Cauchy-Schwarz Inequality)** *If  $X$  and  $Y$  are random variables then*

$$|\mathbb{E}(XY)| \leq \mathbb{E}(X^2)^{\frac{1}{2}} \cdot \mathbb{E}(Y^2)^{\frac{1}{2}}.$$

*Proof.* Let  $t > 0$  and define the random variable  $X + tY$ . We must have  $\mathbb{E}((X + tY)^2) \geq 0$ . Expanding the brackets and using linearity yields

$$t^2\mathbb{E}(Y^2) + 2t\mathbb{E}(XY) + \mathbb{E}(X^2) \geq 0.$$

The left hand side of this inequality is a quadratic function in  $t$ . It follows that the inequality holds if and only if

$$4\mathbb{E}(XY)^2 \leq 4\mathbb{E}(X^2)\mathbb{E}(Y^2),$$

and the result then follows.  $\square$

## Convex Functions and Jensen's Inequality

Let  $f$  be a real-valued function defined on an interval in  $\mathbb{R}$ . We say that it is *convex* if for all  $0 < \lambda < 1$  and all  $x, y$

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

e.g. The following functions defined on  $(0, \infty)$  are convex:  $x^p$  (where  $p > 1$ ),  $-\log(x)$ ,  $e^x$ .

Fact: If  $f$  is twice differentiable and  $f''(x) \geq 0$  on  $[a, b]$ , then  $f$  is convex.

**Theorem 1.4.3 (Jensen's Inequality)** *If  $f$  is a convex function defined on  $[a, b]$  and  $X$  is a random variable taking values in  $[a, b]$ , then*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

*Proof.* We'll prove this only in the simple case where  $X$  is any random variable taking values  $x_1, x_2, \dots, x_n$  with probabilities  $p_1, p_2, \dots, p_n$ . So we want to prove that

$$f\left(\sum_{j=1}^n p_j x_j\right) \leq \sum_{j=1}^n p_j f(x_j).$$

We do this by using mathematical induction. The case  $n = 2$  is just the definition of convexity. Suppose it holds for some  $n$ , then

$$\begin{aligned}
f\left(\sum_{j=1}^{n+1} p_j x_j\right) &= f\left(p_{n+1} x_{n+1} + \sum_{j=1}^n p_j x_j\right) \\
&= f\left(p_{n+1} x_{n+1} + (1 - p_{n+1}) \sum_{j=1}^n \frac{p_j}{1 - p_{n+1}} x_j\right) \\
&\leq p_{n+1} f(x_{n+1}) + (1 - p_{n+1}) f\left(\sum_{j=1}^n \frac{p_j}{1 - p_{n+1}} x_j\right) \text{ by convexity} \\
&\leq p_{n+1} f(x_{n+1}) + \sum_{j=1}^n p_j f(x_j) \text{ by inductive hypothesis} \\
&= \sum_{j=1}^{n+1} p_j f(x_j).
\end{aligned}$$

We have shown that the result holds for  $n + 1$ , hence by induction it is true for all  $n \geq 2$  □

### 1.4.3 Limit Theorems

Limit theorems is a large area within probability. Here we'll just briefly mention three of the most important results.

Let  $(X_n, n \in \mathbb{N})$  be a sequence of *i.i.d.* random variables. *i.i.d.* is short for “independent and identically distributed” and identically distributed means that the laws  $p_{X_n}$  are the same for each  $n$ .

We further assume that each  $X_n$  has finite mean  $\mu$  and finite variance  $\sigma^2$ .

We consider the sequence of *empirical means*  $(\bar{X}_n, n \in \mathbb{N})$  where each

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

The *weak law of large numbers* states that  $\bar{X}_n$  converges to  $\mu$  in probability. This is easy to check using Chebychev's inequality. Indeed we have each

$$\mathbb{E}(\bar{X}_n) = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n},$$

and for all  $c > 0$

$$P(|\bar{X}_n - \mu| > c) \leq \frac{\sigma^2}{nc^2} \rightarrow 0,$$

as  $n \rightarrow \infty$ .

The *strong law of large numbers* states that  $\overline{X}_n$  converges to  $\mu$  almost surely. It is a stronger result than the weak law and considerably harder to prove.

Finally for each  $n \in \mathbb{N}$  define

$$Y_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}},$$

so that  $\mathbb{E}(Y_n) = 0$  and  $\text{Var}(Y_n) = 1$ . The celebrated *central limit theorem* states that  $Y_n$  converges in distribution to a standard normal  $Z$  as  $n \rightarrow \infty$ .

### 1.4.4 Change of Measure

Being able to change from one probability measure to another, hence altering the “rules of chance” is a vital technique in probability theory. It is also of great importance in option pricing - the famous Black-Scholes formula relies heavily on this.

We’ll look at this idea from the abstract measure theory point of view. Let  $(S, \Sigma, m)$  be a measure space and  $f : S \rightarrow \mathbb{R}$  be a non-negative<sup>1</sup> integrable function. We can use integration to define a new measure, i.e. for each  $A \in \Sigma$ , define

$$n(A) = \int_A f(x)m(dx).$$

$n$  is itself a measure. Indeed, we clearly have  $n(\emptyset) = 0$  and if  $A_1, A_2, \dots, A_p$  are disjoint sets in  $\Sigma$  with  $A = \bigcup_{r=1}^p A_r \in \Sigma$ , then we also have

$$n(A) = \sum_{r=1}^p \int_{A_r} f(x)m(dx) = \sum_{r=1}^p n(A_r).$$

This can be extended to give countable additivity using a limiting argument.

A key point to observe is that if  $A \in \Sigma$  is such that  $m(A) = 0$  then  $n(A) = 0$ . This property is important enough to deserve a name.

*Definition.* Let  $m$  and  $n$  be two arbitrary measures on  $(S, \Sigma)$ . We say that  $n$  is *absolutely continuous* with respect to  $m$  if for all  $A \in \Sigma$  which are such that  $m(A) = 0$  we also have  $n(A) = 0$ . If  $n$  is absolutely continuous with respect to  $m$  we write  $n \prec m$ .

If  $n \prec m$  and  $m \prec n$  we say that  $m$  and  $n$  are *equivalent*.

---

<sup>1</sup>i.e.  $f(s) \geq 0$  for all  $s \in S$ .

Note that if  $n \prec m$  then  $n(A) > 0 \Rightarrow m(A) > 0$ .

The reason why absolute continuity is so important is that it underlies the theory of “differentiation of measures”. The key result that we’ll need is the *Radon- Nikodým theorem*.

**Theorem 1.4.4 (Radon- Nikodým)** *Let  $m$  and  $n$  be measures defined on  $(S, \Sigma)$ .  $n$  is absolutely continuous with respect to  $m$  if and only if there exists a non-negative integrable function  $f$  (with respect to  $m$ ) such that*

$$n(A) = \int_A f(x)m(dx),$$

for each  $A \in \Sigma$ .

The theorem actually requires some further restrictions on  $m$  and  $n$  but we won’t worry about these here. It certainly holds in all of the situations where we’ll need it in this course, namely

- when  $m$  is Lebesgue measure and  $n$  is a probability measure.
- when  $m$  and  $n$  are both probability measures.

The function  $f$  which appears in theorem 1.4.4 is called the *Radon- Nikodým derivative* of  $n$  with respect to  $m$ . We sometimes write  $f = \frac{dn}{dm}$ .

**Example** Let  $(\Omega, \mathcal{F}, P)$  be a probability space and  $X$  be a random variable.  $X$  has a pdf  $f_X$  if and only if its law  $p_X$  is absolutely continuous with respect to Lebesgue measure. In this case  $f_X$  is the Radon- Nikodým derivative of  $p_X$  with respect to Lebesgue measure, i.e. for all  $A \in \mathcal{F}$ :

$$P(X \in A) = p_X(A) = \int_A f_X(x)dx.$$

## 1.4.5 Stochastic Processes

Let  $(\Omega, \mathcal{F}, P)$  be a fixed probability space. A *stochastic process*  $(X_i, i \in \mathcal{I})$  is a family of random variables defined on this space.  $\mathcal{I}$  is called the *index set* and usually designates “time” so the random variable  $X_i$  describes observations made at time  $i$ . We talk of *discrete time* stochastic processes when  $\mathcal{I}$  is a discrete set. Typically in this case,  $\mathcal{I} = \mathbb{N} = \{1, 2, 3, \dots\}$  or  $\mathcal{I} = \mathbb{Z}_+ = \mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \dots\}$ .

If  $\mathcal{I}$  is a continuous set we talk of *continuous time* stochastic processes. The usual choice here is  $\mathcal{I} = \mathbb{R}^+ = [0, \infty)$ .

We'll give one useful definition at this stage. A stochastic process  $(X_i, i \in \mathcal{I})$  is said to be *integrable* if each  $X_i$  is an integrable random variable, i.e. each  $\mathbb{E}(|X_i|) < \infty$ .

Typical examples of stochastic processes which you will probably have encountered before are random walks and Markov chains in discrete time and Poisson processes in continuous time. In this course we will study

Martingales - in both discrete and continuous time.

Brownian motion and diffusion processes in continuous time.

These are vital tools for both theoretical developments in modern probability and applications to e.g. physics, biology and finance, especially option pricing.