STAT 901: PROBABILITY

Don McLeish

September 2005

Contents

Part I

Part One

Chapter 1

Mathematical Prerequisites

1.1 Sets and sequences of Real Numbers

The real numbers $\Re$ form a *field*. This is a set together with operations of addition and multiplication and their inverse operations (subtraction and inverse). They are totally ordered in the sense that any two real numbers can be compared; i.e. for any $a, b \in \Re$, either $a < b$, $a = b$, *or* $a > b$. The set of real numbers, unlike the set of rational numbers, is uncountable. A set is *countable* if it can be put in one-one correspondence with the positive integers. It is is *at most countable* if it can be put in one-one correspondence with *a subset* of the positive integers (i.e. finite or countable). The set of rational numbers is countable, for example, but it is easy to show that the set of all real numbers is not. We will usually require the concept of "at most countable" in this course and often not distinguish between these two terminologies, i.e. refer to the set as countable. If we wish to emphasize that a set is infinite we may describe it as *countably infinite*.

A brief diversion: why do we need the machinery of measure theory? Consider the simple problem of identifying a uniform distribution on all subsets of the unit iterval $[0, 1]$ so that this extends the notion of length. Specifically can we define a "measure" or distribution P so that

1. $P(([a, b)) = b - a$ for all $0 \leq a \leq b \leq 1$
2. $P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$ for any disjoint sequence of sets $A_n \subset [0, 1], n = 1, 2, ...$
3. $P(A \oplus r) = P(A)$ for any $r \in [0, 1]$ where for $A \subset [0, 1]$,we define the shift of a set

$$A \oplus r = \{x \in [0, 1]; x - r \in A \text{ or } x - r + 1 \in A\}.$$

Theorem 1 *There is no function P defined on all the subsets of the unit interval which satisfies properties 1-3 above.*

The consequence of this theorem is that in order to define even simple continuous distributions we are unable to deal with *all* subsets of the unit interval or the real numbers but must restrict attention to a subclass of sets or events in what we call a "sigma-algebra".

The set of all integers is not a field because the operation of subtraction (inverse of addition) preserves the set, but the operation of division (inverse of multiplication) does not. However, the set of rational numbers, numbers of the form p/q for integer p and q, forms a field with a countable number of elements. Consider $A \subset \Re$. Then A has an upper bound b if $b \geq a$ for all $a \in A$. If b_0 is the smallest number with this property, we define b_0 to be the *least upper bound*. Similarly lower bounds and greatest lower bounds.

The real numbers is endowed with a concept of distance. More generally, a set $\mathcal{X}$ with such a concept defined on it is called a *metric space* if there is a function $d(x, y)$ defined for all $x, y \in \mathcal{X}$(called the distance between points x and y) satisfying the properties

1. $d(x, y) > 0$ for all $x \neq y$ and $d(x, x) = 0$ for all x.

2. $d(x, y) = d(y, x)$

3. $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in \mathcal{X}$.

Obviously the real line is a metric space with distance $d(x, y) = |x - y|$ but so is any subset of the real line. Indeed any subset of Euclidean space $\Re^n$ is a metric space. A metric space allows us to define the notion of neighbourhoods and open sets. In particular, a neighbourhood of a point x is a set of the form $\{y; d(x, y) < r\}$ for some radius $r > 0$. A subset B of a metric space is *open* if every point x in B has a neighbourhood entirely contained in B.Formally B is open if, for every $x \in B$, there exists $r > 0$ such that $\{y; d(x, y) < r\} \subset B$. Note that the whole metric space X is open, and trivially the empty set φ is open.

We say that a set E in a metric space has an open cover consisting of (possibly infinitely many) open sets $\{G_s, s \in S\}$ if $E \subset \cup_{s \in S} G_s$,or in other words if every point in E is in at least one of the open sets G_s. The set E is *compact* if every open cover has a finite subcover– i.e. if for any open cover there are finitely many sets, say $G_{s_i}, i = 1, ..., n$ such that $E \subset \cup_i G_{s_i}$. Compact sets in Euclidean space are easily identified- they are closed and bounded. In a general metric space, a compact set is always closed.

Now consider a sequence of elements of a metric space $\{x_n, \quad n = 1, 2, \ldots\}$. We say this sequence *converges* to a point x if, for all $\epsilon > 0$ there exists an $N < \infty$ such that $d(x_n, x) < \epsilon$ for all $n > N$. The property that a sequence converges and the value of the limit is a property only of the *tail* of the sequence- i.e. the values for n arbitrarily large. If the sequence consists of real numbers and if we define $l_N = sup\{x_n; n \geq N\}$ to be the least upper bound of the set $\{x_n; n \geq N\}$, then we know the limit x, provided it exists, is less than or equal to each l_N. Indeed since the sequence l_N is a decreasing sequence, bounded below, it must converge to some limit l, and we know that any limit is less than or equal to l as well. The limit $l = lim_{N\to\infty} l_N$ we denote commonly by $l = lim\ sup_{n\to\infty} x_n$.

It is easy to identify $l = lim\ sup$ of a sequence of numbers x_n by comparing it to an arbitrary real number a. In general, $l > a$ if and only if $x_n > a$ infinitely many times or infinitely often (i.e. for infinitely many subscripts n). Similarly $l \le a$ if and only if $x_n > a + \epsilon$ at most finitely many times or finitely often for each $\epsilon > 0$.

We will deal throughout Stat 901 with subsets of the real numbers. For example, consider the set $\mathcal{O}$ of all *open intervals* $(a, b) = \{x; a < x < b\}$ and include $(a, a) = \phi$ the empty set. If we take the union of two (overlapping or non-overlapping) sets in $\mathcal{O}$ is the result in $\mathcal{O}$? What if we take the union of finitely many? Infinitely many? Repeat with intersections. These basic properties of open intervals are often used to describe more general *topologies* since they hold for more complicated spaces such as finite dimensional Euclidean spaces. Denote a closed interval $[a, b] = \{x; a \le x \le b\}$. Which of the above properties hold for closed intervals? Note that we can construct closed intervals from open ones provided we are permitted countably many operations of intersections for example:

$$[a, b] = \cap_{n=1}^{\infty} (a - 1/n,\ b + 1/n).$$

We shall normally use the following notation throughout this course. Ω is a fundamental measure (or probability, or sample) space. It is a set consisting of all *points* possible as the outcome to an experiment. For example what is the probability space if the experiment consists of choosing a random number from the interval $[0, 1]$? What if the experiment consists of tossing a coin repeatedly until we obtain exactly one head? We do not always assume that the space Ω has a topology (such as that induced by a metric) but in many cases it is convenient if the probability space does possess a metric topology. This is certainly the case if we are interested in the value of n random variables and so our space is $\Re^n$.

We denote by Ω a typical point in Ω. We wish to discuss events or classes of sets of possible outcomes.

Definition 2 *(Event) An* Event A *is a subset of* Ω*. The empty event* ϕ *and the whole space* Ω *are also considered events. However, the calculus of probability does not allow us in the most general case to accommodate the set of* all possible subsets of Ω *in general, and we need to restrict this class further.*

Definition 3 *(Topological Space) A topological Space* $(\Omega, \mathcal{O})$ *is a space* Ω *together with a class* $\mathcal{O}$ *of subsets of* Ω*. The members of the set* $\mathcal{O}$ *are called* open *sets.* $\mathcal{O}$ *has the property that unions of any number of the sets in* $\mathcal{O}$ *(finite or infinite, countable or uncountable) remain in* $\mathcal{O}$*, and intersections of finite numbers of sets in* $\mathcal{O}$ *also remain in* $\mathcal{O}$*. The* closed sets *are those whose complements are in* $\mathcal{O}$*.*

Definition 4 *(Some Notation)*

1. Union of sets $A \cup B$
2. Intersection of sets $A \cap B$
3. Complement : $A^c = \Omega \setminus A$
4. Set differences : $A \setminus B = A \cap B^c$.
5. Empty set : $\phi = \Omega^c$

Theorem 5 *(De Morgan's rules)* $(\cup_i A_i)^c = \cap_i A_i^c$ *and* $(\cap_i A_i)^c = \cup_i A_i^c$

Definition 6 *(Boolean Algebra) A Boolean Algebra (or algebra for short) is a family* $\mathcal{F}_\prime$ *of subsets of* Ω *such that*

1. $A, B \in \mathcal{F}_\prime$ *implies* $A \cup B \in \mathcal{F}_\prime$.
2. $A \in \mathcal{F}_\prime$ *implies* $A^c \in \mathcal{F}_\prime$.
3. $\phi \in \mathcal{F}_\prime$.

While Boolean algebras have satisfying mathematical properties, they are not sufficiently general to cover most probability spaces of interest. In particular, they may be used to model experiments with at most a finite number of possible outcomes. In the next chapter, we will deal with extending Boolean algebras to cover more general probability spaces.

1.2 Problems

1. Suppose we consider the space Ω of positive integers and define a measure by $P(A) = 0$ of the number of integers in A is finite, $P(A) = 1$ if the number is infinite. Does this measure satisfy the property of *countable additivity:*

   $$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$$

   for any disjoint sequence of sets $A_n \subset \Omega$, $n = 1, 2, ...$?

2. Prove that the equation $p^2 = 2$ is not satisfied by any rational number p.(Let $p = m/n$ where not both integers m, n are even).

3. The extended real number system consists of the usual real numbers $\{x; -\infty < x < \infty\}$ together with the symbols ∞ and $-\infty$. Which of the following have a meaning in the extended real number system and what is the meaning? Assume x is real $(-\infty < x < \infty)$.

   (a) $x + \infty$

(b) $x - \infty$

(c) $x(+\infty)$

(d) x/∞

(e) $\frac{x}{-\infty}$

(f) $\infty - \infty$

(g) ∞/∞

4. Prove: the set of rational numbers, numbers of the form p/q for integer p and q, has a countable number of elements.

5. Prove that the set of all real numbers is not countable.

6. Let the sets $E_n, n = 1, 2, ...$ each be countable. Prove that $\cup_{n=1}^{\infty} E_n$ is countable.

7. In a metric space, prove that for fixed x and $r > 0$, the set $\{y; d(x, y) < r\}$ is an open set.

8. In a metric space, prove that the union of any number of open sets is open, the intersection of a finite number of open sets is open, but the intersection of an infinite number of open sets might be closed.

9. Give an example of an open cover of the interval $(0, 1)$ which has no finite subcover.

10. Consider A to be the set of rational numbers $a \in Q$ such that $a^2 < 2$. Is there least upper bound, and a greatest lower bound, and are they in Q?

11. Show that any non-decreasing sequence of numbers that is bounded above converges.

12. Show that if $x \leq l_N$ for each $N < \infty$ and if l_N converges to some number l, then $x \leq l$.

13. Find an example of a double sequence $\{a_{ij}, \quad i = 1, 2, \ldots, \quad j = 1, 2, \ldots\}$ such that
$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} \neq \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}$$

14. Define the set $\mathcal{O}$ of *open intervals* $(a, b) = \{x; -a < x < b\}$, $\infty \geq a \geq 0$, $\infty \geq b \geq 0$.

    (a) Verify that the union or intersection of finitely many sets in $\mathcal{O}$ is in $\mathcal{O}$.

    (b) Verify that the union of a countably infinite number of sets in $\mathcal{O}$ is in $\mathcal{O}$

(c) Show that the intersection of a countably infinite number of sets in $\mathcal{O}$ may not be in $\mathcal{O}$.

15. Prove the triangle inequality:

$$|a+b| \leq |a| + |b|$$

whenever $a, b \in \Re^n$.

16. Define the metric $d(X,Y) = \sqrt{E(X-Y)^2}$ on a space of random variables with finite variance. Prove the triangle inequality

$$d(X,Z) \leq d(X,Y) + d(Y,Z)$$

for arbitrary choice of random variables X, Y, Z.(Hint: recall that $cov(W_1, W_2) \leq \sqrt{var(W_1)}\sqrt{var(W_2)}$)

17. Verify that

$$[a,b) = \cap_{n=1}^{\infty}(a-1/n,\ b).$$
$$(a,b) = \cup_{n=1}^{\infty}(a+1/n,\ b-1/n).$$
$$[a,b) = \cup_{n=1}^{\infty}[a,\ b-1/n).$$

18. Let a_n be a sequence or real numbers converging to a. Prove that $|a_n|$ converges to $|a|$. Prove that for any function $f(x)$ continuous at the point a then $f(a_n) \to f(a)$.

19. Give an example of a convergent series $\sum p_n = 1$ with all $p_n \geq 0$ such that the expectation of the distribution does not converge; i.e. $\sum_n np_n = \infty$.

20. Define Ω to be the interval (0,1] and $\mathcal{F}_0$ to be the class of all sets of the form $(a_0, a_1] \cup (a_2, a_3] \cup ... \cup (a_{n-1}, a_n]$ where $0 \leq a_0 \leq ... \leq a_n \leq 1$. Then is $\mathcal{F}_0$ a Boolean algebra? Verify.

21. Prove that any open subset of $\Re$ is the union of countable many intervals of the form (a,b) where $a < b$.

22. Suppose the probability space $\Omega = \{1,2,3\}$ and $P(\varphi) = 0, P(\Omega) = 1$. What conditions are necessary for the values $x = P(\{1,2\}), y = P(\{2,3\}), z = P(\{1,3\})$ for the measure P to be countably additive?

23. Suppose a measure satisfies the property of *countable additivity:*

$$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$$

for any disjoint sequence of sets $A_n \subset \Omega$, $n = 1, 2, ...$?

Prove that for an arbitrary sequence of sets B_j,

$$P(B_1 \cup B_2 \cup ...) \leq P(B_1) + P(B_2) + ...$$

24. Prove for any probability measure and for an arbitrary sets $B_j, j = 1, 2, ..., n$

$$P(B_1 \cup B_2 \cup ...B_n) = \sum_{j=1}^{n} P(B_j) - \sum_{i<j} P(B_i B_j) + \sum_{i<j<k} P(B_i B_j B_k)....$$

25. Find two Boolean Algebras $\mathcal{F}_0$ and $\mathcal{F}_1$ both defined on the space $\Omega = \{1, 2, 3\}$ such that the union $\mathcal{F}_0 \cup \mathcal{F}_1$ is NOT a Boolean Algebra.

26. For an arbitrary space Ω, is it true that

$$\mathcal{F}_0 = \{A \subset \Omega; A \text{ is a finite set}\}$$

is a Boolean algebra?

27. For two Boolean Algebras $\mathcal{F}_0$ and $\mathcal{F}_1$ both defined on the space Ω is it true that the intersection $\mathcal{F}_0 \cap \mathcal{F}_1$ is a Boolean Algebra?

28. The smallest non-empty events belonging to a Boolean algebra are called the *atoms.* Find the atoms of

$$\mathcal{F}_0 = \{\varphi, \Omega, \{1\}, \{2, 3\}, \{4\}, \{1, 2, 3\}, \{1, 4\}, \{2, 3, 4\}\}$$

where $\Omega = \{1, 2, 3, 4\}$.

29. The smallest non-empty events belonging to a Boolean algebra are called the *atoms.* Show that in general different atoms must be disjoint. If a Boolean algebra $\mathcal{F}_0$ has a total of n atoms how many elements are there in $\mathcal{F}_0$?

Chapter 2

Measure Spaces

2.1 Families of Sets

Definition 7 *($\pi-$ systems) A family of subsets $\mathcal{F}$ of Ω is a $\pi-$system if, $A_k \in \mathcal{F}$ for $k = 1, 2$ implies $A_1 \cap A_2 \in \mathcal{F}$.*

A $\pi-$system is closed under finitely many intersections but not necessarily under unions. The simplest example of a $\pi-$system is the family of rectangles in Euclidean space. Clearly a Boolean algebra is a π-system but there are $\pi-$systems that are not Boolean algebras (see the problems).

Definition 8 *(Sigma-Algebra) $\mathcal{F}$ is sigma algebra if,*

(i) $A_k \in \mathcal{F}$ *for all* k *implies* $\cup_{k=1}^{\infty} A_k \in \mathcal{F}$

(ii) $A \in \mathcal{F}$ *implies* $A^c \in \mathcal{F}$.

(iii) $\phi \in \mathcal{F}$.

Note that only the first property of a Boolean algebra has been changed-it is slightly strengthened. Any sigma algebra is automatically a Boolean algebra.

Theorem 9 *(Properties of a Sigma-Algebra) If $\mathcal{F}$ is a sigma algebra, then*

(iv) $\Omega \in \mathcal{F}$.

(v) $A_k \in \mathcal{F}$ *for all* k *implies* $\cap_{k=1}^{\infty} A_k \in \mathcal{F}$

Proof. Note that $\Omega = \varphi^c \in \mathcal{F}$ by properties (ii) and (iii). This verifies (iv). Also $\cap_{k=1}^{\infty} A_k = (\cup_{k=1}^{\infty} A_k^c)^c \in \mathcal{F}$ by properties (i) and (ii). ■

Theorem 10 *(Intersection of sigma algebras) Let $\mathcal{F}_\lambda$ be sigma algebras for each $\lambda \in \Lambda$. The index set Λ may be finite or infinite, countable or uncountable. Then $\cap_\lambda \mathcal{F}_\lambda$ is a sigma-algebra.*

Proof. Clearly if $\mathcal{F} = \cap_\lambda \mathcal{F}_\lambda$ then $\varphi \in \mathcal{F}$ since $\varphi \in \mathcal{F}_\lambda$ for every λ. Similarly if $A \in \mathcal{F}$ then $A \in \mathcal{F}_\lambda$ for every λ and so is A^c. Consequently $A^c \in \mathcal{F}$. Finally if $A_n \in \mathcal{F}$ for all $n = 1, 2, ...$then $A_n \in \mathcal{F}_\lambda$ for every n, λ and $\cup_{n=1}^{\infty} A_n \in \mathcal{F}_\lambda$ for every λ. This implies $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$. ■

Definition 11 *(sigma algebra generated by family of sets) If $\mathcal{C}$ is a family of sets, then the* sigma algebra generated by $\mathcal{C}$ *, denoted $\sigma(\mathcal{C})$, is the intersection of all sigma-algebras containing $\mathcal{C}$. It is the smallest sigma algebra which contains all of the sets in $\mathcal{C}$.*

Example 12 *Consider $\Omega = [0, 1]$ and $\mathcal{C} = \{[0, .3], [.5, 1]\} = \{A_1, A_2\}$, say. Then $\sigma(\mathcal{C}) = \{\varphi, A_1, A_2, A_3, A_1 \cup A_2, A_1 \cup A_3, A_2 \cup A_3, \Omega\}$ where we define $A_3 = (.3, .5)$. (There are 8 sets in $\sigma(\mathcal{C})$).*

Example 13 *Define Ω to be the interval (0,1] and $\mathcal{F}_\prime$ to be the class of all sets of the form $(a_0, a_1] \cup (a_2, a_3] \cup ... \cup (a_{n-1}, a_n]$ where $0 \leq a_0 \leq ... \leq a_n \leq 1$. Then $\mathcal{F}_\prime$ is a Boolean algebra but not a sigma algebra.*

Example 14 *(all subsets) Define $\mathcal{F}_\prime$ to be the class of* all *subsets of any given set Ω. Is this a Boolean algebra? Sigma Algebra? How many distinct sets are there in $\mathcal{F}_\prime$ if Ω has a finite number, N points?*

Example 15 *A and B play a game until one wins once (and is declared winner of the match). The probability that A wins each game is* 0.3*, the probability that B wins each game is* 0.2 *and the probability of a draw on each game is* 0.5*. What is a suitable probability space, sigma algebra and the probability that A wins the match?*

Example 16 *(Borel Sigma Algebra) The Borel Sigma Algebra is defined on a topological space $(\Omega, \mathcal{O})$ and is $\mathcal{B} = \sigma(\mathcal{O})$.*

Theorem 17 *The Borel sigma algebra on $\mathcal{R}$ is $\sigma(\mathcal{C})$, the sigma algebra generated by each of the classes of sets $\mathcal{C}$ described below;*

1. $\mathcal{C}_1 = \{(a, b);\ a \leq b\}$
2. $\mathcal{C}_2 = \{(a, b];\ a \leq b\}$
3. $\mathcal{C}_3 = \{[a, b);\ a \leq b\}$
4. $\mathcal{C}_4 = \{[a, b];\ a \leq b\}$
5. $\mathcal{C}_5$ =the set of all open subsets of $\mathcal{R}$
6. $\mathcal{C}_6$ =the set of all closed subsets of $\mathcal{R}$

To prove the equivalence of 1 and 5 above, we need the following theorem which indicates that any open set can be constructed from a countable number of open intervals.

Theorem 18 *Any open subset of $\mathcal{R}$ is a countable union of open intervals of the form (a,b).*

Proof. Let O be the open set and $x \in O$. Consider the interval $I_x = \cup\{(a,b); a < x < b, (a,b) \subset O\}$. This is the largest open interval around x that is entirely contained in O. Note that if $x \neq y$, then $I_x = I_y$ or $I_x \cap I_y = \varphi$. This is clear because if there is some point $z \in I_x \cap I_y$, then $I_x \cup I_y$ is an open interval containing both x and y and so since they are, by definition, the largest such open interval, $I_x \cup I_y = I_x = I_y$. Then we can clearly write

$$\begin{aligned} O &= \cup\{I_x; x \in O\} \\ &= \cup\{I_x; x \in O, x \text{ is rational}\} \end{aligned}$$

since every interval I_x contains at least one rational number. ∎

Definition 19 *(* Lim Sup, Lim Inf) *For an arbitrary sequence of events A_k*

$$\lim_{n\to\infty} \sup \mathrm{A}_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = [A_n \ i.o.]$$

$$\lim_{n\to\infty} \inf A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k = [A_n \ a.b.f.o.]$$

The notation A_n i.o. refers to A_n infinitely often and A_n $a.b.f.o.$ refers to A_n "all but finitely often".

A given point ω is in $\lim_{n\to\infty} \sup \mathrm{A}_n$ if and only if it lies in infinitely many of the individual sets A_n. The point is in $\lim_{n\to\infty} \inf A_n$ if and only if it is in *all but a finite number of the sets.* Which of these two sets is bigger? Compare them with $\cup_{k=n}^{\infty} A_k$ and $\cap_{k=n}^{\infty} A_k$ for any fixed n. Can you think of any circumstances under which $\limsup A_n = \liminf A_n$? You should be able to prove that

$$[\limsup A_n]^c = \liminf A_n^c.$$

Theorem 20 *Assume $\mathcal{F}$ is a sigma-algebra. If each of $A_n \in \mathcal{F}$, $n = 1, 2, \ldots$, then both $\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$ and $\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$ are in $\mathcal{F}$.*

Definition 21 *(measurable space) A pair $(\Omega, \mathcal{F})$ where the former is a set and the latter a sigma algebra of subsets of Ω is called a* measurable space.

Definition 22 *(additive set function) Consider a space Ω and a family of subsets $\mathcal{F}_0$ of Ω such that $\phi \in \mathcal{F}_0$. Suppose μ_0 is a non-negative set function; i.e. has the properties that*

- $\mu_0 : \mathcal{F}_0 \rightarrow [0, \infty]$
- When F, G and $F \cup G \in \mathcal{F}_0$ and $F \cap G = \phi$, then $\mu_0(F) + \mu_0(G) = \mu_0(F \cup G)$.

Then we call μ_0 an *additive set function* on $(\Omega, \mathcal{F}_0)$.

Note that it follows that $\mu_0(\phi) = 0$ (except in the trivial case that $\mu_0(A) = \infty$ for every subset including the empty set. We rule this out in our definition of a measure.)

Definition 23 *We call* μ_0 *a* countably additive set function *on* $(\Omega, \mathcal{F}_0)$ *if, whenever all* $A_n,\ n = 1, 2, \ldots$ *are members of* $\mathcal{F}_0$ *and* $\cup_{n=1}^{\infty} A_n \in \mathcal{F}_0$ *, and the sets are disjoint (* $A_i \cap A_j = \phi,\ \ i \neq j$*) then it follows that*

$$\mu_0(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_0(A_n)$$

We saw at the beginning of this chapter that the concept of a $\pi-$system provides one basic property of a Boolean algebra, but does not provide for unions. In order to insure that such a family is a $\sigma-$algebra we need the additional conditions provided by a $\lambda-$system (below).

Definition 24 *A family of events* $\mathcal{F}$ *is called a* λ*-system if the following conditions hold:*

1. $\Omega \in \mathcal{F}$
2. $A, B \in \mathcal{F}$ and $B \subset A$ implies $A \backslash B \in \mathcal{F}$
3. If $A_n \in \mathcal{F}$ for all $n = 1, 2, ...$ and $A_n \subset A_{n+1}$ then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$

A $\lambda-$system is closed under set differences if one set is included in the other and monotonically increasing countable unions. It turns out this this provides the axioms that are missing in the definition of a π-system to guarantee the conditions of a sigma-field are satisfied.

Proposition 25 *If* $\mathcal{F}$ *is both a* π*-system and a* λ*-system then it is a sigma-algebra.*

Proof. By the properties of a $\lambda-system$, we have that $\Omega \in \mathcal{F}$ and if $A \in \mathcal{F}$ then $A^c = \Omega \setminus A \in \mathcal{F}$. So we need only show that $\mathcal{F}$ is closed under countable unions. Note that since $\mathcal{F}$ is a $\pi-$system it is closed under finite intersections. Therefore if $A_n \in \mathcal{F}$ for each $n = 1, 2, ...$ then $B_n = \cup_{i=1}^{n} A_i = (\cap_{i=1}^{n} A_i^c)^c \in \mathcal{F}$ for each n and since $B_n \subset B_{n+1}$, $\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n \in \mathcal{F}$ by the third property of a $\lambda - system$. ■

Theorem 26 *(The* $\pi - \lambda$ *Theorem) Suppose a family of sets* $\mathcal{F}$ *is a* $\pi-$*system and* $\mathcal{F} \subset \mathcal{G}$ *where* $\mathcal{G}$ *is a* λ*-system. Then* $\sigma(\mathcal{F}) \subset \mathcal{G}$ *.*

This theorem is due to Dynkin and is proved by showing that the smallest $\lambda-$system containing $\mathcal{F}$ is a π-system and is therefore, by the theorem above, a sigma-algebra.

2.2 Measures

Definition 27 *(*measure*)* μ *is a (non-negative)* measure *on the measurable space* $(\Omega, \mathcal{F})$ *where* $\mathcal{F}$ *is a sigma-algebra of subsets of* Ω *if it is a countably additive (non-negative) set function* $\mu(); \mathcal{F} \to [0, \infty]$.

A measure μ satisfies the following conditions

(i) $\mu(A) \geq 0$ for all A.

(ii) If A_k disjoint, $\mu(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mu(A_k)$

(iii) $\mu(\phi) = 0$

(iv) (monotone) $A \subset B$ implies $\mu(A) \leq \mu(B)$.

(v) (subadditive) $\mu(\cup_k A_k) \leq \sum_k \mu(A_k)$

(vi) (inclusion-exclusion). For finitely many sets,

$$\mu(\cup_{k=1}^{n} A_k) = \sum_k \mu(A_k) - \sum_{i<j} \mu(A_i \cap A_j) + \ldots.$$

(vii) If A_k converges (i.e. is nested increasing or decreasing)

$$\mu(lim_n A_n) = lim_n \mu(A_n)$$
$$\text{where } \lim_n A_n = \begin{cases} \cup_n A_n \text{ if } A_n \text{ increasing} \\ \cap_n A_n \text{ if } A_n \text{ decreasing} \end{cases}$$

Definition 28 *(Measure space)The triple* $(\Omega, \mathcal{F}, \mu)$ *is called a* measure space.

Measures do exist which may take negative values as well but we leave discussion of these for later. Such measures we will call *signed measures*. For the present, however, we assume that every measure takes non-negative values only.

Definition 29 *(Probability measure) A* Probability measure *is a measure* P *satisfying* $P(\Omega) = 1$.

(Additional property) A probability measure also satisfies

(viii) $P(A^c) = 1 - P(A)$

Definition 30 *(Probability space) When the measure* P *is a probability measure, the triple* $(\Omega, \mathcal{F}, \mathcal{P})$ *is called a* probability space.

Theorem 31 *(Conditional Probability) For* $B \in \mathcal{F}$ *with* $P(B) > 0$, $Q(A) = P(A|B) = P(A \cap B)/P(B)$ *is a probability measure on the same space* $(\Omega, \mathcal{F})$.

2.3 Extending a measure from an algebra

Although measures generally need to be supported by sigma-algebras of sets, two probability measures are identical if they are identical on an algebra. The following Theorem is fundamental to this argument, and to the construction of Lebesgue measure on the real line.

Theorem 32 *(Caratheodory Extension) Suppose* $\mathcal{F}_0$ *is a (Boolean) algebra and* μ_0 *a countably additive set function from* $\mathcal{F}_0$ *into* $[0, \infty]$*. Then there is an extension of* μ_0 *to a measure* μ *defined on all of* $\sigma(\mathcal{F}_0)$*. Furthermore, if the total measure* $\mu_0(\Omega) < \infty$ *then the extension is unique.*

Proof. We do not provide a complete proof-details can be found in any measure theory text (e.g. Rosenthal, p.10-14.) Rather we give a short sketch of the proof. We begin by defining the outer measure of any set $E \subset \Omega$ (note it does not have to be in the algebra or sigma-algebra) by the smallest sum of the measures of sets in the algebra which cover the set E, i.e.

$$\mu^*(E) = \inf\{\sum_{n=1}^{\infty} \mu_0(A_n); E \subset \cup_{n=1}^{\infty} A_n, A_n \in \mathcal{F}_0\}.$$

Notice that the outer measure of a set in the algebra is the measure itself $\mu^*(E) = \mu_0(E)$ if $E \in \mathcal{F}_0$. Therefore, this outer measure is countably additive when restricted to the algebra $\mathcal{F}_0$. Generally, however, this outer measure is only subadditive; the measure of a countable union of disjoint events is less than or equal to the sum of the measures of the events. **If it were additive**, then it would satisfy the property;

$$\mu^*(E) = \mu^*(EQ) + \mu^*(EQ^c). \tag{2.1}$$

However, let us consider the class $\mathcal{F}$ of all sets Q for which the above equation (2.1) does hold. The rest of the work in the proof consists of showing that the class of sets $\mathcal{F}$ forms a sigma algebra and when restricted to this sigma algebra, the outer measure μ^* is countably additive, so is a measure. ■

The last condition in the extension theorem can be replaced by a weaker condition, that the measure is *sigma-finite.* In other words it suffices that we can write the whole space as a countable union of subsets A_i (i.e. $\Omega = \cup_{i=1}^{\infty} A_i$) each of which has finite measure $\mu_0(A_i) < \infty$. Lebesgue measure on the real line is sigma-finite but not finite.

Example 33 *Lebesgue measure Define* $\mathcal{F}_0$ *to be the set of all finite unions of intervals (open, closed or half and half) such as*

$$A = (a_0, a_1] \cup (a_2, a_3] \cup ... \cup (a_{n-1}, a_n]$$

where $-\infty \le a_0 \le ... \le a_n \le \infty$*. For* A *of the above form, define* $\mu(A) = \sum_i (a_{2i+1} - a_{2i})$*. Check that this is well-defined. Then there is a unique extension of this measure to all* $\mathcal{B}$ *, the Borel subsets of* $\mathcal{R}$*. This is called the **Lebesgue measure**.*

It should be noted that in the proof of Theorem 11, the sigma algebra $\mathcal{F}$ may in fact be a larger sigma algebra than $\sigma(\mathcal{F}_0)$ generated by the algebra. For example in the case of measures on the real line, we may take $\mathcal{F}_0$ to be all finite union of intervals. In this case $\sigma(\mathcal{F}_0)$ is the class of all Borel subsets of the real line but it is easy to check that $\mathcal{F}$ is a larger sigma algebra having the property of completeness, i.e. for any $A \in \mathcal{F}$ such that $\mu(A) = 0$, all subsets of A are also in $\mathcal{F}$ (and of course also have measure 0).

Example 34 *(the Cantor set) This example is useful for dispelling the notions that closed sets must either countain intervals or consist of a countable selection of points. Let $\Omega = [0,1]$ with P Lebesgue measure. Define $A_1 = \Omega \backslash \{(\frac{1}{3}, \frac{2}{3})\}$ and $A_2 = A \backslash \{(\frac{1}{9}, \frac{2}{9}) \cup (\frac{7}{9}, \frac{8}{9})\}$ etc. In each case, $A_n$is obtained from A_{n-1} by deleting the open interval in the middle third of each interval in A_{n-1}. Define $A = \cap_{n=1}^{\infty} A_n$. Then A is a closed, uncountable set such that $P(A) = 0$ and A contains no nondegenerate intervals.*

2.4 Independence

Definition 35 *(Independent Events) A family of events $\mathcal{C}$ is (mutually) independent if*

$$P(A_{\lambda_1} \cap A_{\lambda_2} .... A_{\lambda_n}) \;=\; P(A_{\lambda_1}) P(A_{\lambda_2}) .... P(A_{\lambda_n}) \qquad (*)$$

for all n, $A_{\lambda_i} \in \mathcal{C}$ and for distinct λ_i.

Properties: Independent Events

1. A, B independent implies A, B^c independent.
2. Any A_λ can be replaced by A_λ^c in equation (*).

Definition 36 *Families of sigma-algebras $\{\mathcal{F}_\lambda;\ \lambda \in \Lambda\}$ are independent if for any $A_\lambda \in \mathcal{F}_\lambda$, the family of events$\{A_\lambda; \lambda \in \Lambda\}$ are mutually independent.*

Example 37 *(Pairwise independence does not imply independence) Two fair coins are tossed. Let A = first coin is heads, B = second coin is heads, C = we obtain exactly one heads. Then A is independent of B and A is independent of C but A, B, C are **not mutually independent**.*

2.4.1 The Borel Cantelli Lemmas

Clearly if events are individually too small, then there little or no probability that their lim sup will occur, i.e. that they will occur infinitely often.

Lemma 38 *For an arbitrary sequence of events A_n, $\sum_n P(A_n) < \infty$ implies $P[A_n \ i.o.] = 0$.*

Proof. Notice that

$$P(\cap_{n=1}^{\infty}\cup_{m=n}^{\infty}A_m)\le P(\cup_{m=n}^{\infty}A_m)\le\sum_{m=n}^{\infty}P(A_m)\ \text{ for each } n=1,2,.....$$

For any $\epsilon>0$, since the series $\sum_{m=1}^{\infty}P(A_m)$ converges we can find a value of n sufficiently large that $\sum_{m=n}^{\infty}P(A_m)<\epsilon$. Therefore for every positive ϵ, $P(\cap_{n=1}^{\infty}\cup_{m=n}^{\infty}A_m)\le\epsilon$ and so it must equal 0. ■

The converse of this theorem is false without some additional conditions. For example suppose that Ω is the unit interval and the measure is Lebesgue. Define $A_n=[0,\frac{1}{n}], n=1,2,.....$ Now although $\sum P(A_n)=\infty$, it is still true that $P(A_n \text{ i.o.})=0$. However if we add the condition that the events are independent, we do have a converse as in the following.

Lemma 39 *For a sequence of* independent events A_n, $\sum_n P(A_n)=\infty$ *implies* $P[A_n\ i.o]=1$.

Proof. We need to show that $P(A_n^c \quad \text{a.b.f.o.})=0$. This is

$$\begin{aligned}P(\cup_{n=1}^{\infty}\cap_{m=n}^{\infty}A_m^c)&\le\sum_{n=1}^{\infty}P(\cap_{m=n}^{\infty}A_m^c)\\&\le\sum_{n=1}^{\infty}\prod_{m=n}^{N_n}(1-P(A_m))\text{ for any sequence }N_n\\&\le\sum_{n=1}^{\infty}\exp\{-\sum_{m=n}^{N_n}P(A_m)\}\end{aligned}$$

where we have used the inequality $(1-P(A_m))\le\exp(-P(A_m))$. Now if the series $\sum_{m=1}^{\infty}P(A_m)$ diverges to ∞ then we can choose the sequence N_n so that $\sum_{m=n}^{N_n}P(A_m)>nln2-ln\epsilon$ in which case the right hand side above is less than or equal to ϵ. Since this holds for arbitrary $\epsilon>0$, this verifies that $P(\cup_{n=1}^{\infty}\cap_{m=n}^{\infty}A_m^c)=0$. ■

Definition 40 *(Almost surely)A statement S about the points in Ω holds almost surely(a.s.) or with probability one if the set of ω such that the statement holds has probability one. Thus Lemma 13 above states that A_n occurs infinitely often almost surely (a.s.) and Lemma 12 that A_n^c occurs all but finitely often (a.s.).*

2.4.2 Kolmogorov's Zero-one law

For independent events A_n, put

$$\mathcal{F}=\cap_{n=1}^{\infty}\sigma(A_n,A_{n+1},...)$$

(call this the *tail sigma-algebra*). Events that are determined by the sequence $\{A_1,A_2,...\}$ but not by a finite number such as $\{A_1,...A_N\}$ are in the tail sigma-algebra. This includes events such as $[\limsup A_n],[\liminf A_n]$, $[\limsup A_{2^n}]$, etc.

Theorem 41 ***(zero-one law)*** *Any event in the tail sigma-algebra* $\mathcal{F}$ *has probability either 0 or 1.*

Proof. Define $\mathcal{F}_n = \sigma(A_1, A_2, ..., A_n)$ and suppose $B \in \mathcal{F}_n$ for fixed n. Then B is independent of $\mathcal{F}$ because it is independent of all sets in the larger sigma algebra $\sigma(A_{n+1}, A_{n+2}, ...)$. This means that every set $A \in \mathcal{F}$ is independent of every set in each $\mathcal{F}_n$ and therefore A is independent of each member of the Boolean Algebra of sets $\cup_{n=1}^{\infty}\mathcal{F}_n$. Therefore A is independent of $\sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$. But since

$$\cap_{n=1}^{\infty}\sigma(A_n, X_{n+1}, ...) \subset \sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$$

A is independent of itself, implying it has probability either 0 or 1 (see problem 18). ∎

2.5 Problems.

1. Give an example of a family of subsets of the set $\{1, 2, 3, 4\}$ that is a $\pi-$system but NOT a Boolean algebra of sets.

2. Consider the space $\Re^2$and define the family of all rectangles with sides parallel to the axes. Show that this family is a π-system.

3. Let Ω be the real line and let $\mathcal{F}_n$ be the sigma-algebra generated by the subsets
$$[0, 1), [1, 2), ..., [n-1, n)$$
Show that the sigma-algebras are *nested* in the sense that $\mathcal{F}_1 \subset \mathcal{F}_2$. How do you know if a given set is in $\mathcal{F}_n$? Show that $\cup_{n=1}^{100}\mathcal{F}_n$ is a sigma-algebra.

4. As above, let Ω be the real line and let $\mathcal{F}_n$ be the sigma-algebra generated by the subsets
$$[0, 1), [1, 2), ..., [n-1, n)$$
Show that $\cup_{n=1}^{\infty}\mathcal{F}_n$ is not a sigma-algebra.

5. How do we characterise the open subsets of the real line $\Re$? Show that the Borel sigma algebra is also generated by all sets of the form $(\infty, x], x \in \Re$.

6. For an arbitrary sequence of events A_k , give a formula for the event $B_k =$ [the first of the A_j's to occur is A_k].

7. Write in set-theoretic terms the event that exactly two of the events A_1, A_2, A_3, A_4, A_5 occur.

8. Prove that if A_k is a nested sequence of sets (increasing or decreasing), then $\limsup A_n = \liminf A_n$ and both have probability equal to $lim_n P(A_n)$.

9. Prove Bayes Rule:

   If $P(\cup_n B_n) = 1$ for a disjoint finite or countable sequence of events B_n all with positive probability, then

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_n P(A|B_n)P(B_n)}$$

10. Prove that if $A_1, ..., A_n$ are independent events, then the same is true with any number of A_i replaced by their complement A_i^c. This really implies therefore that any selection of one set from each of $\sigma(A_1), \sigma(A_2), \ldots \sigma(A_n)$ is a set of mutually independent events.

11. Find an example such that A, B are independent and B, C are independent but $P(A \cup B|C) \neq P(A \cup B)$.

12. Prove that for any sequence of events A_n ,

$$P(\liminf A_n) \leq \liminf P(A_n)$$

13. Prove the *multiplication rule*. That if $A_1 \ldots A_n$ are arbitrary events,

$$P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 A_1) \ldots P(A_n|A_1 A_2 \ldots A_{n-1})$$

14. Consider the unit interval with Lebesgue measure defined on the Borel subsets. For any point x in the interval, let $0.x_1 x_2 x_3 \ldots$ denote its decimal expansion (terminating wherever possible) and suppose A is the set of all points x such that $x_i \neq 5,\ i = 1, 2, \ldots$.

    (a) Prove that the set A is Borel measurable and find the measure of the set A.

    (b) Is the set A countable?

15. Give an example of a sequence of sets $A_n, n = 1, 2, ...$ such that $lim\ sup\ A_n = lim\ inf\ A_n$ but the sequence is not nested. Prove in this case that $P(lim\ sup\ A_n) = limP(A_n)$.

16. In a given probability space, every pair of distinct events are independent so if $B \neq A$, then

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

    What values for the probabilities $P(A)$ are possible? Under what circumstances is it possible that

$$P(A \cap B) \leq P(A)P(B)$$

    for all $A \neq B$?

17. Prove that a $\lambda-$system does not need to be closed under general unions or finite intersections. For example let $\mathcal{F}$ consist of all subsets of $\{1, 2, 3, 4\}$ which have either 0 or 2, or 4 elements.

18. Suppose $\mathcal{F}_0$ is a Boolean algebra of sets and $A \in \sigma(\mathcal{F}_0)$ has the property that A is independent of every set in $\mathcal{F}_0$. Prove that $P(A) = 0$ or 1.

19. Prove: If $\mathcal{F}$ is both a π-system and a λ-system then it is a sigma-field.

20. Is the family consissting of all countable subsets of a space Ω and their complements a sigma-algebra?

21. Find $\limsup A_n$ and $\liminf A_n$ where $A_n = (\frac{1}{n}, \frac{2}{3} - \frac{1}{n}), n = 1, 3, 5, ...$ and $A_n = (\frac{1}{3} - \frac{1}{n}, 1 + \frac{1}{n}), n = 2, 4, 6, ...$

22. Consider a measure μ_0 defined on a Boolean algebra of sets $\mathcal{F}_0$ satisfying the conditions of Theorem 11. For simplicity assume that $\mu_0(\Omega) = 1$. Consider the class of sets $\mathcal{F}$ defined by

$$\mathcal{F} = \{A \subset \Omega; \mu^*(AE) + \mu^*(A^c E) = \mu^*(E) \text{ for all } E \subset \Omega\}.$$

Prove that $\mathcal{F}$ is a Boolean algebra.

23. Consider $\mathcal{F}$ as in Problem 22. Prove that if $A_1, A_2, ...$ disjoint subsets of $\mathcal{F}$ then $\mu^*(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu^*(A_i)$ so that this outer measure is countably additive.

24. Consider $\mathcal{F}$ as in Problem 22. Prove that $\mathcal{F}$ is a sigma-algebra.

25. Consider $\mathcal{F}$ as in Problem 22. Prove that if $A \in \mathcal{F}_0$ then $\mu^*(A) = \mu(A)$.

26. Prove or disprove: the family consisting of all finite subsets of a space Ω and their complements is a sigma-algebra.

27. Prove or disprove: the family consisting of all countable subsets of a space Ω and their complements is a sigma-algebra.

28. Find two sigma-algebras such that their union is not a sigma algebra.

29. Suppose P and Q are two probability measures both defined on the same sample space Ω and sigma algebra $\mathcal{F}$. Suppose that $P(A) = Q(A)$ for all events $A \in \mathcal{F}$ such that $P(A) \leq \frac{1}{2}$. Prove that $P(A) = Q(A)$ for all events A. Show by counterexample that this statement is not true if we replace the condition $P(A) \leq \frac{1}{2}$ by $P(A) < \frac{1}{2}$.

Chapter 3

Random Variables and Measurable Functions.

3.1 Measurability

Definition 42 *(Measurable function) Let* f *be a function from a measurable space* $(\Omega, \mathcal{F})$ *into the real numbers. We say that the function is* measurable *if for each Borel set* $B \in \mathcal{B}$ *, the set* $\{\omega; f(\omega) \in B\} \in \mathcal{F}$*.*

Definition 43 *(random variable) A* random variable X *is a measurable function from a probability space* $(\Omega, \mathcal{F}, \mathcal{P})$ *into the real numbers* $\Re$*.*

Definition 44 *(Indicator random variables) For an arbitrary set* $A \in \mathcal{F}$ *define* $I_A(\omega) = 1$ *if* $\omega \in A$ *and 0 otherwise. This is called an* indicator random variable*.*

Definition 45 *(Simple Random variables) Consider events* $A_i \in \mathcal{F}, i = 1, 2, 3, ..., n$ *such that* $\cup_{i=1}^n A_i = \Omega$*. Define* $X(\omega) = \sum_{i=1}^n c_i I_{A_i}(\omega)$ *where* $c_i \in \Re$*. Then* X *is measurable and is consequently a random variable. We normally assume that the sets* A_i *are disjoint. Because this is a random variable which can take only finitely many different values, then it is called* simple *and any random variable taking only finitely many possible values can be written in this form.*

Example 46 *(binomial tree) A stock, presently worth $20, can increase each day by $1 or decrease by $1. We observe the process for a total of 5 days. Define* X *to be the value of the stock at the end of five days. Describe* $(\Omega, \mathcal{F})$ *and the function* $X(\omega)$ *. Define another random variable* Y *to be the value of the stock after 4 days.*

Define $X^{-1}(B) = \{\omega; X(\omega) \in B\}$. We will also sometimes denote this event $[X \in B]$. In the above example, define the events $X^{-1}(B)$ and $Y^{-1}(B)$ where $B = [20, \infty)$.

Then we have the following **properties.**
For any Borel sets $B_n \subset \Re$, and any random variable X,

1. $X^{-1}(\cup_n B_n) = \cup_n X^{-1}(B_n)$
2. $X^{-1}(\cap_n B_n) = \cap_n X^{-1}(B_n)$
3. $[X^{-1}(B)]^c = X^{-1}(B^c)$

These three properties together imply that for any class of sets $\mathcal{C}$, $X^{-1}(\sigma(\mathcal{C})) = \sigma(X^{-1}(\mathcal{C}))$. So X is measurable if, for all x, $\{\omega; X(\omega) \le x\} \in \mathcal{F}$ (see Theorem 16 and Problem 3.16).

BEWARE: The fact that we use notation X^{-1} does not imply that the function X has an inverse in the usual sense. For example, if $X(\omega) = sin(\omega)$ for $\omega \in \Re$, then what is $X^{-1}([.5, 1])$?

Theorem 47 *(combining random variables) Suppose* $X_i, i = 1, 2, \ldots$ *are all (measurable) random variables. Then so are*

1. $X_1 + X_2 + X_3 + \ldots X_n$
2. X_1^2
3. cX_1 for any $c \in \mathcal{R}$
4. X_1X_2
5. $inf\ \{X_n; n \ge 1\}$
6. $\liminf X_n$
7. $sup\{X_n; n \ge 1\}$
8. $\limsup_{n\to\infty} X_n$

Proof. For 1. notice that $[X_1 + X_2 > x]$ if and only if there is a rational number q in the interval $X_1 > q > x - X_2$ so that $[X_1 > q]$ and $[X_2 > x - q]$. In other words

$[X_1 + X_2 > x] = \cup_q [X_1 > q] \cap [X_2 > x - q]$ where the union is over all rational numbers q.

For 2, note that for $x \ge 0$,

$$[X_1^2 \le x] = [X_1 \ge 0] \cap [X_1 \le \sqrt{x}] \cup [X_1 < 0] \cap [X_1 \ge -\sqrt{x}].$$

For 3, in the case $c > 0$, notice that

$$[cX_1 \le x] = [X_1 \le \frac{x}{c}].$$

Finally 4 follows from properties 1, 2 and 3 since

$$X_1X_2 = \frac{1}{2}\{(X_1 + X_2)^2 - X_1^2 - X_2^2\}$$

For 5. note that $[\inf X_n \geq x] = \cap_{n=1}^{\infty}[X_n \geq x]$.

For 6. note that $[\liminf X_n \geq x] = [X_n > x - 1/m \quad \text{a.b.f.o.}]$ for all $m = 1, 2, ...$ so

$$[\liminf X_n \geq x] = \cap_{m=1}^{\infty} \liminf [X_n > x - 1/m].$$

The remaining two properties follow by replacing X_n by $-X_n$. ■

Definition 48 *(sigma-algebra generated by random variables) For X a random variable, define $\sigma(X) = \{X^{-1}(B); B \in \mathcal{B}\}$.*

$\sigma(X)$ is the smallest sigma algebra $\mathcal{F}$ such that X is a measurable function into $\Re$. The fact that it is a sigma-algebra follows from Theorem 16. Similarly, for a set of random variables $X_1, X_2, \ldots X_n$, the sigma algebra $\sigma(X_1, \ldots X_n)$ generated by these is the smallest sigma algebra such that all X_i are measurable.

Theorem 49 *$\sigma(X)$ is a sigma-algebra and is the same as $\sigma\{[X \leq x], \quad x \in \Re\}$.*

Definition 50 *A* Borel measurable function *f from $\Re \to \Re$ is a function such that $f^{-1}(B) \in \mathcal{B}$ for all $B \in \mathcal{B}$.*

For example if a function $f(x)$ is a continuous function from a subset of $\Re$ into a subset of $\Re$ then it is Borel measurable.

Theorem 51 *Suppose $f_i, i = 1, 2, \ldots$ are all Borel measurable functions. Then so are*

1. $f_1 + f_2 + f_3 + \ldots f_n$
2. f_1^2
3. cf_1 for any real number c.
4. $f_1 f_2$
5. $\inf\{f_n; n \geq 1\}$
6. $\liminf f_n$
7. $\sup\{f_n; n \geq 1\}$
8. $\lim_{n\to\infty} f_n$

Theorem 52 *If X and Y are both random variables, then Y can be written as a Borel measurable function of X, i.e. $Y = f(X)$ for some Borel measurable f if and only if*

$$\sigma(Y) \subset \sigma(X)$$

Proof. Suppose $Y = f(X)$. Then for an arbitrary Borel set B, $[Y \in B] = [f(X) \in B] = [X \in f^{-1}(B)] = [X \in B_2]$ for Borel set $B_2 \in \mathcal{B}$. This shows that $\sigma(Y) \subset \sigma(X)$.

For the converse, we assume that $\sigma(Y) \subset \sigma(X)$ and we wish to find a Borel measurable function f such that $Y = f(X)$. For fixed n consider the set $A_{m,n} = \{\omega; m2^{-n} \leq Y(\omega) < (m+1)2^{-n}\}$ for $m = 0, \pm 1, \ldots.$ Since this set is in $\sigma(Y)$ it is also in $\sigma(X)$ and therefore can be written as $\{\omega; X(\omega) \in B_{m,n}\}$ for some Borel subset $B_{m,n}$ of the real line. Consider the function $f_n(x) = \sum_m m2^{-n} I(x \in B_{m,n})$. Clearly this function is defined so that $f_n(X)$ is close to Y, and indeed is within $\frac{1}{2^n}$ units of Y. The function we seek is obtained by taking the limit

$$f(x) = \lim_{n\to\infty} f_n(x).$$

We require two results, first that the limit exists and second that the limit satisfies the property $f(X) = Y$. Convergence of the sequence follows from the fact that for each x, the sequence $f_n(x)$ is monotonically increasing (this is Problem 22). The fact that $Y = f(X)$ follows easily since for each n, $f_n(X) \leq Y \leq f_n(X) + \frac{1}{2^n}$. Taking limits as $n \to \infty$ gives $f(X) \leq Y \leq f(X)$. ■

Example 53 *Consider* $\Omega = [0,1]$ *with Lebesgue measure and define a random variable* $X(\omega) = a_1, a_2, a_3$ *(any three distinct real numbers) for* $\omega \in [0,1/4], (1/4,1/2], (1/2,1]$ *respectively. Find* $\sigma(X)$. *Now consider a random variable* Y *such that* $Y(\omega) = 0$ *or* 1 *as* $\omega \in [0,1/2], (1/2,1]$ *respectively. Verify that* $\sigma(Y) \subset \sigma(X)$ *and that we can write* $Y = f(X)$ *for some Borel measurable function* $f(.)$.

3.2 Cumulative Distribution Functions

Definition 54 *The* cumulative distribution function *(c.d.f.) of a* random variable X *is defined to be the function* $F(x) = P[X \leq x]$, *for* $x \in \Re$. *Similarly, if* μ *is a measure on* $\Re$, *then the cumulative distribution function is defined to be* $F(x) = \mu(-\infty, x]$. *Note in the latter case, the function may take the value* ∞.

Theorem 55 *(Properties of the Cumulative Distribution Function)*

1. A c.d.f. $F(x)$ is non-decreasing. i.e. $F(y) \geq F(x)$ whenever $y \geq x$.

2. $F(x) \to 0$, as $x \to -\infty$.

3. When $F(x)$ is the c.d.f. of a random variable, $F(x) \to 1$, as $x \to \infty$.

4. $F(x)$ is right continuous. i.e. $F(x) = \lim F(x+h)$ as h decreases to 0.

Proof.

1. If $x \leq y$ then $X \leq x$ implies $X \leq y$ or in set theoretic terms $[X \leq x] \subset [X \leq y]$. Therefore $P(X \leq x) \leq P(X \leq y)$.

2. If X is a real-valued random variable then $[X = -\infty] = \varphi$ the empty set. Therefore for any sequence x_n **decreasing** to $-\infty$,

$$\begin{aligned}\lim F(x_n) &= \lim P(X \leq x_n) \\ &= P(\cap_{n=1}^{\infty}[X \leq x_n]) \quad \text{(since the sequence is nested)} \\ &= P(\varphi) = 0\end{aligned}$$

3. Again if X is a real-valued random variable then $[X < \infty] = \Omega$ and for any sequence x_n **increasing** to ∞,

$$\begin{aligned}\lim F(x_n) &= \lim P(X \leq x_n) \\ &= P(\cup_{n=1}^{\infty}[X \leq x_n]) \quad \text{(since the sequence is nested)} \\ &= P(\Omega) = 1.\end{aligned}$$

4. For any sequence h_n decreasing to 0,

$$\begin{aligned}\lim F(x+h_n) &= \lim P(X \leq x+h_n) \\ &= P(\cap_{n=1}^{\infty}[X \leq x+h_n]) \quad \text{(since the sequence is nested)} \\ &= P(X \leq x) = F(x)\end{aligned}$$

■

Theorem 56 *(existence of limits) Any bounded non-decreasing function has at most countably many discontinuities and possesses limits from both the right and the left. In particular this holds for cumulative distribution functions.*

Suppose we denote the limit of $F(x)$ from the left by $F(x-) = lim_h F(x-h)$ as h decreases to 0. Then $P[X < x] = F(x-)$ and $P[X = x] = F(x) - F(x-)$, the jump in the c.d.f. at the point x.

Definition 57 *Let x_i be any sequence of real numbers and p_i a sequence of non-negative numbers such that $\sum_i p_i = 1$. Define*

$$F(x) = \sum_{\{i; x_i \leq x\}} p_i. \tag{3.1}$$

This is the c.d.f. of a distribution which takes each value x_i with probability p_i. A discrete *distribution is one with for which there is a countable set S with $P[X \in S] = 1$. Any discrete distribution has cumulative distribution function of the form (3.1).*

Theorem 58 *If $F(x)$ satisfies properties 1-4 of Theorem 19, then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X defined on this probability space such that F is the c.d.f. of X.*

Proof. We define the probability space to be $\Omega = (0,1)$ with $\mathcal{F}$ the Borel sigma algebra of subsets of the unit interval and P the Borel measure. Define $X(\omega) = sup\{z; F(z) < \omega\}$. Notice that for any c , $X(\omega) > c$ implies $\omega > F(c)$. On the other hand if $\omega > F(c)$ then since F is right continuous, for some $\epsilon > 0$, $\omega > F(c+\epsilon)$ and this in turn implies that $X(\omega) \geq c+\epsilon > c$. It follows that $X(\omega) > c$ if and only if $\omega > F(c)$. Therefore $P[X(\omega) > c] = P[\omega > F(c)] = 1 - F(c)$ and so F is the cumulative distribution function of X. ■

3.3 Problems

1. If $\Omega = [0,1]$ and P is Lebesgue measure, find $X^{-1}(C)$ where $C = [0, \frac{1}{2})$ and $X(\omega) = \omega^2$.

2. Define $\Omega = \{1,2,3,4\}$ and the sigma algebra $\mathcal{F} = \{\phi, \Omega, \{1\}, \{2,3,4\}\}$. Describe all random variables that are measurable on the probability space $(\Omega, \mathcal{F})$.

3. Let $\Omega = \{-2,-1,0,1,2\}$ and consider a random variable defined by $X(\omega) = \omega^2$. Find $\sigma(X)$,the sigma algebra generated by X.Repeat if $X(\omega) = |\omega|$ or if $X(\omega) = \omega + 1$.

4. Find two different random variables defined on the space $\Omega = [0,1]$ with Lebesgue measure which have exactly the same distribution. Can you arrange that these two random variables are independent of one another?

5. If X_i; $i = 1,2,...$ are random variables, prove that $max_{i \leq n} X_i$ is a random variables and that $limsup \frac{1}{n} \sum_i X_i$ is a random variable.

6. If X_i; $i = 1,2,...$ are random variables, prove that $X_1 X_2 ... X_n$ is a random variable.

7. Let Ω denote the set of all outcomes when tossing an unbiased coin three times. Describe the probability space and the random variable $X =$ *the number of heads observed*. Find the cumulative distribution function $P[X \leq x]$.

8. A number x is called a point of increase of a distribution function F if $F(x+\epsilon) - F(x-\epsilon) > 0$ for all $\epsilon > 0$. Construct a discrete distribution function such that every real number is a point of increase. (Hint: Can you define a discrete distribution supported on the set of all rational numbers?).

9. Consider a stock price process which goes up or down by a constant factor (e.g. $S_{t+1} = S_t u$ or $S_t d$ (where $u > 1$ and $d < 1$) with probabilities p and $1-p$ respectively (based on the outcome of the toss of a biased coin). Suppose we are interested in the path of the stock price from time $t = 0$ to time $t = 5$. What is a suitable probability space? What is $\sigma(S_3)$? What are the advantages of requiring that $d = 1/u$?

10. Using a Uniform random variable on the interval $[0,1]$, find a random variable X with distribution $F(x) = 1 - p^{\lfloor x \rfloor}, x > 0$, where $\lfloor x \rfloor$ denotes the floor or integer part of. Repeat with $F(x) = 1 - e^{-\lambda x}, x > 0, \lambda > 0$.

11. Suppose a coin with probability p of heads is tossed repeatedly. Let A_k be the event that a sequence of k or more consecutive heads occurs amongst tosses numbered $2^k, 2^k+1, \ldots, 2^{k+1}-1$. Prove that $P[A_k \ \ i.o.] = 1$ if $p \geq 1/2$ and otherwise it is 0.

    (*Hint:* Let E_i be the event that there are k consecutive heads beginning at toss numbered $2^k + (i-1)k$ and use the inclusion-exclusion formula.)

12. *The Hypergeometric Distribution* Suppose we have a collection (the *population*) of N objects which can be classified into two groups S or F where there are r of the former and $N-r$ of the latter. Suppose we take a random sample of n items without replacement from the population. Show the probability that we obtain exactly x S's is

$$f(x) = P[X = x] = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \ x = 0, 1, \ldots$$

    Show in addition that as $N \to \infty$ in such a way that $r/N \to p$ for some parameter $0 < p < 1$, this probability function approaches that of the *Binomial Distribution*

$$f(x) = P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,\ 1, \ldots n$$

13. *The Negative Binomial distribution*

    The binomial distribution is generated by assuming that we repeated trials a fixed number n of times and then counted the total number of successes X in those n trials. Suppose we decide in advance that we wish a fixed number (k) of successes instead, and sample repeatedly until we obtain exactly this number. Then the number of trials X is random. Show that the probability function is:

$$f(x) = P[X = x] = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots$$

14. Let $g(u)$ be a cumulative distribution function on $[0,1]$ and $F(x)$ be the cumulative distribution function of a random variable X. Show that we can define a *deformed* cumulative distribution function such that $G(x) = g(F(x))$ at at all continuity points of $g(F(x))$. Describe the effect of this transformation when

$$g(u) = \Phi\left(\Phi^{-1}(u) - \alpha\right)$$

    for Φ the standard normal cumulative distribution function. Take a special case in which F corresponds to the $N(2,1)$ cumulative distribution function.

15. Show that if X has a continuous c.d.f. $F(x)$ then the random variable $F(X)$ has a uniform distribution on the interval $[0, 1]$.

16. Show that if $\mathcal{C}$ is a class of sets which generates the Borel sigma algebra in $\mathcal{R}$ and X is a random variable then $\sigma(X)$ is generated by the class of sets
$$\{X^{-1}(A); A \in \mathcal{C}\}.$$

17. Suppose that $X_1, X_2, ...$are independent Normal(0,1) random variables and $S_n = X_1 + X_2 + ... + X_n$. Use the Borel Cantelli Lemma to prove the strong law of large numbers for normal random variables. i.e. prove that for and $\varepsilon > 0$,
$$P[S_n > n\varepsilon \quad i.o.] = 0.$$
Note: you may use the fact that if $\Phi(x)$ and $\phi(x)$ denote the standard normal cumulative distribution function and probability density function respectively, $1 - \Phi(x) \leq Cx\phi(x)$ for some constant C. Is it true that $P[S_n > \sqrt{n}\varepsilon \quad i.o.] = 0$?

18. Show that the following are equivalent:

    (a) $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ for all x, y

    (b) $P(X \in A, Y \in B) = P(X \in A)P(X \in B)$ for all Borel subsets of the real numbers A, B.

19. Let X and Y be independent random variables. Show that for any Borel measurable functions f, g on $\mathcal{R}$, the random variables $f(X)$ and $g(Y)$ are independent.

20. Show that if A is an uncountable set of non-negative real numbers, then there is a sequence of elements of A, $a_1, a_2, ...$ such that $\sum_{i=1}^{\infty} a_i = \infty$.

21. Mrs Jones made a rhubarb crumble pie. While she is away doing heart bypass surgery on the King of Tonga, her son William (graduate student in Stat-Finance) comes home and eats a random fraction X of the pie. Subsequently her daughter Wilhelmina (PhD student in Stat-Bio) returns and eats a random fraction Y of the remainder. When she comes home, she notices that more than half of the pie is gone. If one person eats more than a half of a rhubard-crumble pie, the results are a digestive catastrophe. What is the probability of such a catastrophe if X and Y are independent uniform on $[0, 1]$?

22. Suppose for random variables Y and X, $\sigma(Y) \subset \sigma(X)$. Define sets by
$$A_{m,n} = \{\omega; m2^{-n} \leq Y(\omega) < (m+1)2^{-n} \text{ for } m = 0, \pm1, ...$$
and define a function f_n by
$$f_n(x) = \sum_m m2^{-2} I(x \in B_{m,n})$$

where

$$[X \in B_{m,n}] = A_{m,n}.$$

Prove that the sequence of functions f_n is non-decreasing in n.

23. Let (Ω, F, P) be the unit interval $[0,1]$ together with the Borel subsets and Borel measure. Give an example of a function from $[0,1]$ into $\mathcal{R}$ which is NOT a random variable.

24. Let (Ω, F, P) be the unit interval $[0,1]$ together with the Borel subsets and Borel measure. Let $0 \le a < c < c < d \le 1$ be arbitrary real numbers. Give and example of a sequence of events $A_n, n = 1, 2, ...$ such that the following all hold:

$$P(\liminf A_n) = a$$
$$\liminf P(A_n) = b$$
$$\limsup P(A_n) = c$$
$$P(\limsup A_n) = d$$

25. Let $A_n, n = 1, 2, ...$ be a sequence of events such that A_i and A_j are independent whenever

$$|i - j| \ge 2$$

and $\sum_n P(A_n) = \infty$. Prove that

$$P(\limsup A_n) = 1$$

26. For each of the functions below find the smallest sigma-algebra for which the function is a random variable. $\Omega = \{-2, -2, 0, 1, 2\}$ and

    (a) $X(\omega) = \omega^2$
    (b) $X(\omega) = \omega + 1$
    (c) $X(\omega) = |\omega|$

27. Let $\Omega = [0,1]$ with the sigma-algebra $\mathcal{F}$ of Borel subsets B contained in this unit interval which have the property that $B = 1 - B$.

    (a) Is $X(\omega) = \omega$ a random variable with respect to this sigma-algebra?
    (b) Is $X(\omega) = |\omega - \frac{1}{2}|$ a random variable with respect to this sigma-algebra?

28. Suppose Ω is the unit square in two dimensions together with Lebesgue measure and for each $\omega \in \Omega$, we define a random variable $X(\omega)$ = minimum distance to an edge of the square. Find the cumulative distribution function of X and its derivative.

29. Suppose that X and Y are two random variables on the same probability space with joint distribution

$$P(X=m,Y=n)=\begin{cases}\frac{1}{2^{m+1}} & \text{if } m\geq n\\ 0 & \text{if } m<n\end{cases}.$$

Find the marginal cumulative distribution functions of X and Y.

Chapter 4

Integration

4.1 Great Expections

An indicator random variable I_A takes two values, the value 1 with probability $P(A)$ and the value 0 otherwise. Its expected value, or average over many trials would therefore be $0(1-P(A))+1P(A)=P(A)$. This is the simplest case of an integral or expectation. It is also the basic building block from which expected value in general (or the Lebesgue integral) is constructed. We begin, however, with an example illustrating the problems associated with the Rieman integral, usually defined by approximating the integral with inner and outer sums of rectangles.

Example 59 So what's so wrong with the Riemann integral anyway? *Let $f(x)=1$ for x irrational and in the interval $[0,1]$, otherwise $f(x)=0$. What is the Riemann integral $\int_0^1 f(x)dx$? What should this integral be?*

Recall that a simple random variable takes only finitely many possible values, say $c_1,....c_n$ on sets $A_1,...,A_n$ in a partition of the probability space. The definition of the integral or expected value for indicator random variables together with the additive properties expected of integrals leads to only one possible definition of integral for simple random variables:

Definition 60 *(Expectation of simple random Variables) A simple random variable can be written in the form $X=\sum_{i=1}^n c_i I_{A_i}$. In this case, we define $E(X)=\sum_{i=1}^n c_i P(A_i)$.* Note: we must show that this is well-defined; i.e. that if there are two such representations of the same random variable X then both lead to the same value of $E(X)$.

4.1.1 Properties of the Expected Value for Simple Random Variables

Theorem 61 *For simple random variables X, Y,*

1. $X(\omega) \le Y(\omega)$ for all ω implies $E(X) \le E(Y)$.
2. For real numbers α, β, $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

Proof. Suppose $X = \sum_i c_i I_{A_i} \le \sum_j d_j I_{B_j}$ where A_i forms a disjoint partition of the space Ω (i.e. are disjoint sets with $\cup_i A_i = \Omega$) and B_j also forms a disjoint partition of the space. Then $c_i \le d_j$ whenever $A_i B_j \ne \phi$. Therefore

$$\begin{aligned} E(X) &= \sum_i c_i P(A_i) = \sum_i c_i \sum_j P(A_i B_j) \\ &\le \sum_i \sum_j d_j P(A_i B_j) = \sum_j d_j P(B_j) = E(Y) \end{aligned}$$

For the second part, note that $\alpha X + \beta Y$ is also a simple random variable that can be written in the form $\sum_i \sum_j (\alpha c_i + \beta d_j) I_{A_i B_j}$ where the sets $A_i B_j$ form a disjoint partition of the sample space Ω. Now take expectation to verify that this equals $\alpha \sum_i c_i P(A_i) + \beta \sum_j d_j P(B_j)$.

4.1.2 Expectation of non-negative measurable random variables.

Suppose X is a non-negative random variable so that $X(\omega) \ge 0$ for all $\omega \in \Omega$. Then we define

$$E(X) = sup\{E(Y);\ Y \text{ is simple and } Y \le X\}.$$

The supremum is well-defined, although it might be infinite. There should be some concern, of course, as to whether this definition will differ **for simple random variables** from the one listed previously, but this is resolved in property 1 below.

4.1.3 Some Properties of Expectation.

Assume X, Y are non-negative random variables. Then ;

1. If $X = \sum_i c_i I_{A_i}$ simple, then $E(X) = \sum_i c_i P(A_i)$.
2. If $X(\omega) \le Y(\omega)$ for all ω, then $E(X) \le E(Y)$.
3. If X_n is increasing to X pointwise, then $E(X_n)$ increases to $E(X)$ (this is usually called the *Monotone Convergence Theorem*).
4. For non-negative numbers α, and β,

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$

Proof. *Proof of Properties.*

1. If $Z \le X$ and Z is a simple function, then $E(Z) \le E(X)$. It follows that since X is a simple function and we take the supremum over all simple functions Z , that this supremum is $E(X)$.

2. Suppose Z is a simple function and $Z \le X$. Then $Z \le Y$. It follows that the set of Z satisfying $Z \le X$ is a subset of the set satisfying $Z \le Y$ and therefore $sup\{E(Z); Z \text{ is simple}, Z \le X\} \le sup\{E(Z); Z \text{ is simple}, Z \le Y\}$.

3. Since $X_n \le X$ if follows from property (2) that $E(X_n) \le E(X)$. Similarly $E(X_n)$ is monotonically non-decreasing and it therefore converges. Thus it converges to a limit satisfying

$$limE(X_n) \le E(X).$$

We will now show that $lim\ E(X_n) \ge E(X)$ and then conclude equality holds above. Suppose $\epsilon > 0$ is arbitrary and $Y = \sum_i c_i I_{A_i}$ where $Y \le X$ is a simple random variable. Define $B_n = \{\omega; X_n(\omega) \ge (1-\epsilon)Y(\omega)\}$ Note that as $n \to \infty$, this sequence of sets increases to a set containing $\{\omega; X(\omega) \ge (1-\epsilon/2)Y(\omega)\}$ and since $X \ge Y$ the latter is the whole space Ω. Therefore,

$$E(X_n) \ge E(X_n I_{B_n}) \ge (1-\epsilon)E(Y I_{B_n}).$$

But

$$E(Y I_{B_n}) = \sum_i c_i P(A_i B_n) \to \sum_i c_i P(A_i)$$

as $n \to \infty$. Therefore

$$\lim E(X_n) \ge (1-\epsilon)E(Y)$$

whenever Y is a simple function satisfying $Y \le X$. Note that the supremum of the right hand side over all such Y is $(1-\epsilon)E(X)$. We have now shown that for any $\epsilon > 0$, $\lim E(X_n) \ge (1-\epsilon)E(X)$ and it follows that this is true also as $\epsilon \to 0$.

4. Take two sequences of simple random variables X_n increasing to X and Y_n increasing to Y. Assume α and β are non-negative. Then by Property 2. of 4.1.1,

$$E(\alpha X_n + \beta Y_n) = \alpha E(X_n) + \beta E(Y_n)$$

By monotone convergence, the left side increases to the limit $E(\alpha X + \beta Y)$ while the right side increases to the limit $\alpha E(X) + \beta E(Y)$. We leave the more general case of a proof to later.

■

Definition 62 *(General Definition of Expected Value) For an arbitrary random variable X, define $X^+ = \max(X, 0)$, and $X^- = \max(0, -X)$. Note that $X = X^+ - X^-$. Then we define $E(X) = E(X^+) - E(X^-)$. This is well defined even if one of $E(X^+)$ or $E(X^-)$ are equal to ∞ as long as both or not infinite since the form $\infty - \infty$ is meaningless.*

Definition 63 *(integrable) If both $E(X^+) < \infty$ and $E(X^-) < \infty$ then we say X is* integrable*.*

Notation;

$E(X) = \int X(\omega)dP$

$\int_A X(\omega)dP = E(XI_A)$ for $A \in \mathcal{F}$.

4.1.4 Further Properties of Expectation.

In the general case, expectation satisfies 1-4 of 4.1.3. above plus the the additional property:

5. If $P(A) = 0$, $\int_A X(\omega)dP = 0$.

Proof. (property 5)

Suppose the non-negative random variable $Z = \sum_{i=1}^n c_i I_{B_i}$ is simple and $Z \leq XI_A$. Then for any i, $c_i I_{B_i} \leq XI_A$ which implies either $c_i = 0$ or $B_i \subset A$. In the latter case, $P(B_i) \leq P(A) = 0$. Therefore $E(Z) = \sum_{i=1}^n c_i P(B_i) = 0$. Since this holds for every simple random variable $Z \leq XI_A$ it holds for the supremum

$$E(XI_A) = \sup\{E(Z); Z \text{ is simple, } Z \leq XI_A\} = 0.$$

■

Theorem 64 *(An integral is a measure) If X is non-negative r.v. and we define $\mu(A) = \int_A X(\omega)dP$, then μ is a (countably additive) measure on $\mathcal{F}$.*

Proof. Note that by property 5 above, $\mu(\varphi) = 0$ and since $XI_A \geq 0$, $E(XI_A) \geq 0$ by property 2 of the integal. Note also that the set function μ is finitely additive. In particular if A_1 and A_2 are **disjoint** events,

$$\mu(A_1 \cup A_2) = E(XI_{A_1 \cup A_2}) = E(X(I_{A_1} + I_{A_2})) = \mu(A_1) + \mu(A_2).$$

This shows that the set function is additive. By induction we can easily prove that it is finitely additive; that for disjoint sets $A_i, i = 1, 2, ...$

$$\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i).$$

To show that the set function is countably additive, define $B_n = \cup_{i=1}^n A_i$. Notice that the random variables XI_{B_n} form a non-decreasing sequence converging to XI_B where $B = \lim_{n\to\infty} B_n$ (recall that the limit of a nested sequence of sets is well-defined and in this case equals the union). Therefore by the monotone convergence theorem (property 3 above),

$$\sum_{i=1}^{n} \mu(A_i) = E(XI_{B_n}) \to E(XI_B) = \mu(\cup_{i=1}^\infty A_i).$$

Therefore, the set function is countably additive, i.e.

$$\mu(\cup_{i=1}^\infty A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

Consequently the set function satisfies the conditions of a measure. If $E(X) < \infty$ then this measure is finite. Otherwise, if we define events $C_n = [X \le n]$, then notice that $\mu(C_n) \le n$. Moreover, $\Omega = \cup_n C_n$. This shows that the measure is sigma-finite (i.e. it is the countable union of sets C_n each having finite measure). ■

Lemma 65 *(Fatou's lemma: limits of integrals) If X_n is a sequence of non-negative r.v.,*

$$\int [\liminf X_n] dP \le \liminf \int X_n dP$$

Proof. Define $Y_n(\omega) = inf_{\{m;m\ge n\}} X_m(\omega)$. Note that Y_n is a non-decreasing sequence of random variables and $lim Y_n = lim\ inf\ X_n = X$, say. Therefore by monotone convergence, $E(Y_n) \to E(X)$. Since $Y_n \le X_n$ for all n ,

$$E(X) = lim E(Y_n) \le lim\ inf\ E(X_n).$$

■

Example 66 *(convergence a.s. implies convergence in expectation?) It is possible for $X_n(\omega) \to X(\omega)$ for all ω but $E(X_n)$ does not converge to $E(X)$. Let $\Omega = (0,1)$ and the probability measure be Lebesgue measure on the interval. Define $X(\omega) = n$ if $0 < \omega < 1/n$ and otherwise $X(\omega) = 0$. Then $X_n(\omega) \to 0$ for all ω but $E(X_n) = 1$ does not converge to the expected value of the limit.*

Theorem 67 *(Lebesgue dominated convergence Theorem) If $X_n(\omega) \to X(\omega)$ for each ω, and there exists integrable Y with $|X_n(\omega)| \le Y(\omega)$ for all n, ω, then X is integrable and $E(X_n) \to E(X)$.*

(*Note for future reference: the Lebesgue Dominated Convergence Theorem can be proven under the more general condition that X_n converges in distribution* to X)

Proof. Since $Y \geq |X_n|$ the random variables $Y + X_n$ are non-negative. Therefore by Fatou's lemma,

$$E[\liminf(Y + X_n)] \leq \liminf E(Y + X_n)$$

or $E(Y) + E(X) \leq E(Y) + \liminf E(X_n)$ or $E(X) \leq \liminf E(X_n)$. Similarly, applying the same argument to the random variables $Y - X_n$ results in

$$E[\liminf(Y - X_n)] \leq \liminf E(Y - X_n)$$

or $E(Y) - E(X) \leq E(Y) - \limsup E(X_n)$ or

$$E(X) \geq \limsup E(X_n).$$

It follows that $E(X) = \lim E(X_n)$. ■

4.2 The Lebesgue-Stieltjes Integral

Suppose $g(x)$ is a Borel measurable function $\Re \to \Re$. By this we mean that $\{x; g(x) \in B\}$ is a Borel set for each Borel set $B \subset \Re$. Suppose $F(x)$ is a Borel measurable function satisfying two of the conditions of 3.2.2, namely

1. $F(x)$ is non-decreasing. i.e. $F(x) \geq F(y)$ whenever $x \geq y$.
2. $F(x)$ is right continuous. i.e. $F(x) = \lim F(x+h)$ as h decreases to 0.

Notice that we can use F to define a measure μ on the real line; for example the measure of the interval $(a, b]$ we can take to be $\mu((a, b]) = F(b) - F(a)$. The measure is extended from these intervals to all Borel sets in the usual way, by first defining the measure on the algebra of finite unions of intervals, and then extending this measure to the Borel sigma algebra generated by this algebra. We will define $\int g(x)dF(x)$ or $\int g(x)d\mu$ exactly as we did expected values in section 4.1 but with the probability measure P replaced by μ and $X(\omega)$ replaced by $g(x)$. In particular, for a simple function $g(x) = \sum_i c_i I_{A_i}(x)$, we define $\int g(x)dF = \sum_i c_i \mu(A_i)$.

4.2.1 Integration of Borel measurable functions.

Definition 68 *Suppose $g(x)$ is a non-negative Borel measurable function so that $g(x) \geq 0$ for all $x \in \Re$. Then we define*

$$\int g(x)d\mu = \sup\{\int h(x)d\mu;\ h \text{ simple},\ h \leq g\}.$$

Definition 69 *(General Definition: integral) As in Definitions 62 and 63, for a general function $f(x)$ we write $f(x) = f^+(x) - f^-(x)$ where both f^+ and f^- are non-negative functions. We then define $\int f d\mu = \int f^+ d\mu - \int f^- d\mu$ provided that this makes sense (i.e. is not of the form $\infty - \infty$). Finally we say that f is integrable if both f^+ and f^- have* finite integrals*, or equivalently, if $\int |f(x)| d\mu < \infty$.*

4.2.2 Properties of integral

For arbitrary Borel measurable functions $f(x)$, $g(x)$,

1. $f(x) \le g(x)$ for all x implies $\int f(x)d\mu \le \int g(x)d\mu$.
2. For real numbers α, β, $\int(\alpha f + \beta g)d\mu = \alpha \int f d\mu + \beta \int g d\mu$.
3. If f_n increasing to f, then $\int f_n d\mu$ increases to $\int f d\mu$ (called the *Monotone Convergence Theorem*).

The monotone convergence theorem holds even if the limiting function f is not integrable, i.e. if $\int f d\mu = \infty$. In this case it says that $\int f_n d\mu \to \infty$ as $n \to \infty$.

Example 70 *Consider a discrete function defined for non-negative constants* $p_j, j = 1, 2, ...$ *and real numbers* $x_j, j = 1, 2, ...$ *by*

$$F(x) = \sum_{\{j; x_j \le x\}} p_j$$

Then

$$\int_{-\infty}^{\infty} g(x)dF = \sum_j g(x_j)p_j.$$

If the constants p_j *are probabilities, i.e. if* $\sum p_j = 1$, *then this equals* $E[g(X)]$ *where* X *is a random variable having c.d.f.* F.

Example 71 *(completion of Borel sigma algebra) The* Lebesgue measure λ *is generated by the function* $F(x) = x$. *Thus we define* $\lambda((a, b]) = b - a$ *for all* a, b, *and then extend this measure to a measure on all of the Borel sets. A sigma-algebra* $\mathcal{L}$ *is* complete *with respect to Lebesgue measure* λ *if whenever* $A \in \mathcal{L}$ *and* $\lambda(A) = 0$ *then every subset of* A *is also in* $\mathcal{L}$. *The* completion *of the Borel sigma algebra with respect to Lebesgue measure is called the* Lebesgue sigma algebra. *The extension of the measure* λ *above to all of the sets in* $\mathcal{L}$ *is called* Lebesgue measure.

Definition 72 *(absolutely continuous) A measure* μ *on* $\Re$ *is* absolutely continuous *with respect to Lebesgue measure* λ *(denoted* $\mu << \lambda$*) if there is an integrable function* $f(x)$ *such that* $\mu(B) = \int_B f(x)d\lambda$ *for all Borel sets* B. *The function* f *is called the* density *of the measure* μ *with respect to* λ.

Intuitively, two measures μ, λ on the same measurable space $(\Omega, \mathcal{F})$ (not necessarily the real line) satisfy $\mu << \lambda$ if the support of the measure λ includes the support of the measure μ. For a discrete space, the measure μ simply reweights those points with non-zero probabilities under λ. For example if λ represents the discrete uniform distribution on the set $\Omega = \{1, 2, 3, ..., N\}$ (so that $\lambda(B)$ is $N^{-1}\times$the number of integers in $B \cap \{1, 2, 3, ..., N\}$) and $f(x) = x$, then if $\mu(B) = \int_B f(x)d\lambda$,we have $\mu(B) = \sum_{x \in B \cap \{1,2,3,...,N\}} x$. Note that the measure μ assigns weights $\frac{1}{N}, \frac{2}{N}, ...1$ to the points $\{1, 2, 3, ..., N\}$ respectively.

4.2.3 Notes on absolute continuity

The so-called *continuous distributions* such as the normal, gamma, exponential, beta, chi-squared, student's t, etc. studied in elementary statistics should have been called *absolutely continuous with respect to Lebesgue measure.*

Theorem 73 (The Radon-Nykodym Theorem); *For arbitrary measures μ and λ defined on the same measure space, the two conditions below are equivalent:*

1. μ is absolutely continuous with respect to λ so that there exists a function $f(x)$ such that
$$\mu(B) = \int_B f(x) d\lambda$$
2. For all B, $\lambda(B) = 0$ implies $\mu(B) = 0$.

The first condition above asserts the existence of a "density function" as it is usually called in statistics but it is the second condition above that is usually referred to as absolute continuity. The function $f(x)$ is called the *Radon Nikodym* derivative of μ w.r.t. λ. We sometimes write $f = \frac{d\mu}{d\lambda}$ but f is not in general unique. Indeed there are many $f(x)$ corresponding to a single μ, i.e. many functions f satisfying $\mu(B) = \int_B f(x)d\lambda$ for all Borel B. However, for any two such functions f_1, f_2, $\lambda\{x; f_1(x) \neq f_2(x)\} = 0$. This means that f_1 and f_2 are *equal almost everywhere* (λ).

The so-called discrete distributions in statistics such as the binomial distribution, the negative binomial, the geometric, the hypergeometric, the Poisson or indeed any distribution concentrated on the integers is absolutely continuous with respect to the counting measure $\lambda(A)$ =number of integers in A.

If the measure induced by a c.d.f. $F(x)$ is absolutely continuous with respect to Lebsegue measure, then $F(x)$ is a continuous function. However it is possible that $F(x)$ be a continuous function without the corresponding measure being absolutely continuous with respect to Lebesgue measure.

Example 74 *Consider $F(x)$ to be the cumulative distribution of a random variable uniformly distributed on the Cantor set. In other words, if X_i are independent Bernoulli (1/2) random variables, define*

$$X = \sum_{i=1}^{\infty} \frac{2X_i}{3^i}$$

and $F(x) = P[X \leq x]$. Then it is not hard to see that the measure corresponding to this cumulative distribution function is continuous but not absolutely continuous with respect to Lebesgue measure. In fact if C is the Cantor set, $\mu(C) = P(X \in C) = 1$ but $\lambda(C) = 0$ so condition 2 of the Theorem above fails. On the other hand the cumulative distribution function is a continuous function because for any real number $x \in [0, 1]$ we have

$$P[X = x] = 0.$$

The measure $\mu(B) = P(X \in B)$ is an example of one that is singular with respect to Lebesgue measure. This means in effect that the support of the two measures μ and λ is non-overlapping.

Definition 75 *Measures μ and λ defined on the same measurable space are mutually singular if they have disjoint supports; i.e. if there are disjoint sets A and A^c such that $\mu(A) = 0$ and $\lambda(A^c) = 0$.*

Proof. (Radon-Nykodym Theorem.) The fact that condition 1. implies condition 2. is the result of 4.1.4 property 5. so we need only prove the reverse. Assume both measures are defined on the measure space $(\Omega, \mathcal{F})$ and that for all $B \in F$, $\lambda(B) = 0$ imples $\mu(B) = 0$. Also assume for simplicity that both measures are finite and so $\lambda(\Omega) < \infty, \mu(\Omega) < \infty$. Define a class of measurable functions $\mathcal{C}$ by

$$\mathcal{C} = \{g; g(x) \geq 0, \int_E g d\lambda \leq \mu(E) \text{ for all } E \in \mathcal{F}\}.$$

We wish to show that there is a function $f \in \mathcal{C}$ that is maximal in the sense that

$$\int_\Omega f d\lambda = \sup\{\int_\Omega g d\lambda; g \in \mathcal{C}\} = \alpha, \text{ say.}$$

and that this function has the properties we need. First, note that if two functions $g_1, g_2 \in \mathcal{C}$, then $\max(g_1, g_2) \in \mathcal{C}$. This is because we can write

$$\begin{aligned} \int_E \min(g_1, g_2) d\lambda &= \int_{EA} g_1 d\lambda + \int_{EA^c} g_2 d\lambda \text{ where } A = \{\omega; g_1(\omega) > g_2(\omega)\} \\ &\leq \mu(EA) + \mu(EA^c) \\ &\leq \mu(E) \end{aligned}$$

Similarly the maximum of a finite number of elements of $\mathcal{C}$ is also in $\mathcal{C}$.Suppose, for each n, we choose g_n such that $\int_\Omega g_n d\lambda \geq \alpha - \frac{1}{n}$. Then the sequence

$$f_n = \max(g_1, ..., g_n)$$

is an increasing sequence and by monotone convergence it converges to a function $f \in \mathcal{C}$ for which $\int_\Omega f d\lambda = \alpha$. If we can show that $\alpha = \mu(\Omega)$ then the rest of the proof is easy. Define a new measure by $\mu_s(E) = \mu(E) - \int_E f d\lambda$. Suppose that there is a set A such that $\lambda(A) > 0$ and assume for the moment that the measures μ_s, λ are **not** mutually singular. Then by problem 25 there exists $\varepsilon > 0$ and a set A with $\lambda(A) > 0$ such that

$$\varepsilon \lambda(E) \leq \mu_s(E)$$

for all measurable sets $E \subset A$.Consequently for all E,

$$\begin{aligned}\int_E (f+\varepsilon I_A)d\lambda &= \int_E fd\lambda + \varepsilon\lambda(A\cap E)\\ &\le \int_E fd\lambda + \mu_s(A\cap E)\\ &\le \int_E fd\lambda + \mu(AE) - \int_{AE} fd\lambda\\ &\le \int_{E\backslash A} fd\lambda + \mu(AE)\\ &\le \mu(E\backslash A)+\mu(AE)=\mu(E).\end{aligned}$$

In other words, $f+\varepsilon I_A \in \mathcal{C}$. This contradicts the fact that f is maximal, since $\int_\Omega (f+\varepsilon I_A)d\lambda = \alpha + \varepsilon\lambda(A) > \alpha$. Therefore, by contradiction, the measures $\mu_s$and λ must be mutually singular. This implies that there is a set B such that $\mu_s(B)=0$ and $\lambda(B^c)=0$. But since $\mu << \lambda$, $\mu(B^c)=0$ and $\mu_s(B^c) \le \mu(B^c)=0$ which shows that the measure μ_s is identically 0. This now shows that

$$\mu(E) = \int_E fd\lambda \text{ for all } E, \text{ as was required.}$$

■

Definition 76 *Two measures μ and λ defined on the same measure space are said to be* equivalent *if both $\mu << \lambda$ and $\lambda << \mu$.*

Two measures μ, λ on the same measurable space are equivalent if $\mu(A)=0$ if and only if $\lambda(A)=0$ for all A.Intuitively this means that the two measures share exactly the same support or that the measures are either both positive on a given event or they are both zero an that event.

4.2.4 Distribution Types.

There are three different types of probability distributions, when expressed in terms of the cumulative distribution function.

1. Discrete: For countable x_n, p_n, $F(x) = \sum_{\{n;x_n\le x\}} p_n$. The corresponding measure has countably many atoms.

2. Continuous singular. $F(x)$ is a continuous function but for some Borel set B having Lebesgue measure zero, $\lambda(B)=0$, we have $P(X\in B)=1$. (For example, the uniform distribution on the Cantor set is singular since it is supported entirely by a set of Lebesgue measure 0 . We will later denote $P(X\in B)$ as obtained from its cumulative distribution function F by $\int_B F(dx)$).

3. Absolutely continuous (with respect to Lebesgue measure).

$$F(x) = \int_{-\infty}^{x} f(x) d\lambda$$

for some function f called the *probability density function*.

There is a general result called the Lebesgue decomposition which asserts that any any cumulative distribution function can be expressed as a mixture of those of the above three types. In terms of measures, any sigma-finite measure μ on the real line can be written

$$\mu = \mu_d + \mu_{ac} + \mu_s,$$

the sum of a discrete measure μ_d, a measure μ_{ac} absolutely continuous with respect to Lebesgue measure and a measure μ_s that is continuous singular. For a variety of reasons of dubious validity, statisticians concentrate on absolutely continuous and discrete distributions, excluding, as a general rule, those that are singular.

4.3 Moments and the Moment Generating Function

Many of the properties of a random variable X are determined from its moments. The $k'th$ moment of X is $E(X^k)$. If the first moment $\mu = E(X)$, the $k'th$ central moment is $E[(X-\mu)^k]$. For example the variance is the second central moment $var(X) = \sigma^2 = E[(X-\mu)^2]$. We also define the skewness

$$\frac{E[(X-\mu)^3]}{\sigma^3}$$

and the Kurtosis

$$\frac{E[(X-\mu)^4]}{\sigma^4}.$$

The normal distribution is often taken as the standard against which skewness and kurtosis is measured and for the normal distribution (or any distribution symmetric about its mean with third moments), *skewness = 0* . Similarly for the normal distribution *kurtosis = 3* . Moments are often most easily obtained from the moment generating function of a distribution. Thus if X has a given c.d.f. $F(x)$, the moment generating function is defined as

$$m_X(t) = E[exp\{Xt\}] = \int_{-\infty}^{\infty} e^{xt} dF, \quad t \in \Re.$$

Since this is the expected value of a non-negative quantity it is well-defined but might, for some t, take the value ∞. The domain of the moment generating function, the set of t for which this integral is finite, is often a proper subset of the real numbers. For example consider the moment generating function of an exponential random variable with probability density function

$$f(x) = \frac{1}{4}\exp(-\frac{x}{4}), \text{ for } x > 0.$$

The moments are easily extracted from the moment generating function since

$$m_X(t) = \sum_{j=0}^{\infty} \frac{t^j E(X^j)}{j!}$$

provided that this series converges absolutely in an open neighbourhood of $t = 0$. Differentating n times and then setting $t = 0$ recovers the moment, viz.

$$E(X^n) = m_X^{(n)}(0).$$

The moment generating function of the normal (μ, σ) distribution is $m(t) = \exp\{\mu t + \frac{\sigma^2 t^2}{2}\}$.

Definition 77 *(convex function) A function* $g(x)$ *on an interval of the real line is said to be* convex *if for every pair of points* x, y *in the interval, and every point* $0 < p < 1$,

$$g(px + (1-p)y) \leq pg(x) + (1-p)g(y).$$

This can be restated as "the graph of the function always lies below any chord" *or alternatively* "the function of a weighted average is less than the weighted average of the function". *In view of the last statement, since expected value is a form of weighted average, the following theorem is a natural one.*

Theorem 78 *(Jensen's Inequality) If* $g(x)$ *is a convex function and both* X *and* $g(X)$ *are integrable, then*

$$g(EX) \leq E[g(X)]$$

Proof. Let us denote the point $(EX, g(EX))$ by $p_0 = (x_0, g_0)$. Since g is convex, it is not difficult to show that there exists a line $l(x)$ through the point p_0 such that the graph of g lies on or above this line. In particular, with

$$l(x) = g_0 + k(x - x_0)$$

we have $g(x) \geq l(x)$ for all x. Therefore

$$E(g(X)) \geq E(l(X)) = g_0 + k(EX - EX) = g(E(X)),$$

thus proving Jensen's inequality. ■

For example the functions $g_1(x) = x^2$ and $g_2(x) = e^{tX}, t > 0$ are both convex functions and so $[E(X)]^2 \leq E[X^2]$ and $e^{tEX} \leq E[e^{tX}]$.

4.4 Problems

1. Prove that a c.d.f $F(x)$ can have at most a countable number of discontinuities (i.e. points x such that $F(x) > F(x-)$).

2. A stock either increases or decreases by 5 % each day with probability p, each day's movement independent of the preceding days. Find p so that the expected rate of return matches that of a risk free bond whose return is a constant r units per day. Give an expression for the probability that the stock will more than double in price in 50 days. Use the normal approximation to the Binomial distribution to estimate this probability when $r = .01\%$.

3. One of the fundamental principals of finance is the *no-arbitrage* principle, which roughly states that all financial products should be priced in such a way that it is impossible to earn a positive return with probability one. To take a simple example, suppose a market allows you to purchase or borrow any amount of a stock and an interest free bond, both initially worth \$1. It is known that at the end of the next time interval the stock will either double or halve its value to either \$2.00 or \$0.50. Suppose you own an option which pays you exactly \$1.00 if the stock goes up, zero otherwise. Construct a portfolio of stocks and bonds which is identical to this option and thereby determine the value of the option. Note that its value was determined without knowing the probabilities with which the stock increased or decreased. Repeat this calculation if the bond pays interest r per unit time. Note that the no-arbitrage principle generates probabilities for the branches. Although these may not be the true probabilities with which movements up or down occur, they should nevertheless be used in valuing a derivative.

4. Suppose a stock moves in increments of ± 1 and S_n is the stock price on day n so that $S_{n+1} = S_n \pm 1$. If we graph the possible values of S_n as $n = 0, 1, 2, \ldots N$ we obtain a *binomial tree.* Assume on day n the interest rate is r_n so that 1 dollar invested on day n returns $(1 + r_n)$ on day $n + 1$. Use the above no-arbitrage principle to determine the probabilities of up and down movements throughout the binomial tree. Use these probabilities in the case $N = 6$ to determine the initial value of derivative that will pay $S_N - 14$ if this is positive, and otherwise pay 0 asssuming $S_n = 10$. Assume constant interest rate $r_n = .01$.

5. (*A constructive definition of the integral)* For a given non-negative random variable X, define a simple random variable $X_n = \sum_{i=1}^{n2^n} c_i I_{A_i}$ where

$$c_i = (i-1)/2^n, \quad A_i = [(i-1)/2^n \le X < i/2^n], \quad i < n2^n,$$

and

$$A_{n2^n} = [(n2^n - 1)/2^n \le X].$$

Prove that X_n is an increasing function and that $E(X) = lim\ E(X_n)$. This is sometimes used as the definition of the integral.

6. Show that if X is integrable, then $|E(X)| \leq E(|X|)$. Similarly, show $|E(X)| \leq \sqrt{E(|X|^2)}$.

7. Suppose X_n is a sequence of random variables such that for some event A with $P(A) = 1$ and for all $\omega \in A$, $X_n(\omega)$ increases to $X(\omega)$. Prove that $E(X_n)$ increases to $E(X)$.

8. Show that if X, Y are two integrable random variables for which $P[X \neq Y] = 0$, then $\int_A X dP = \int_A Y dP$ for all $A \in \mathcal{F}$.

9. Show that if $X \geq 0$ is integrable and $X \geq |Y|$ then Y is integrable.

10. Prove property 5, page 37: if $P(A) = 0$, $\int_A X(\omega)dP = 0$.

11. If X is non-negative r.v., $\mu(A) = \int_A X(\omega)dP$ defines a (countably additive) measure on $\mathcal{F}$. (proved as Theorem 22)

12. Restate the theorems in section 4.1 for the Lebesgue-Stieltjes integral of functions. Give simple conditions on the functions g_n under which

$$lim \int g_n(x)d\lambda = \int lim\ g_n(x)d\lambda$$

13. Suppose X is a random variable with c.d.f. $F(x)$. Show that $E(X)$ as defined in section 4.1 is the same as $\int xdF$ as defined in section 4.2.

14. Suppose X is a non-negative random variable. Show that $E(X) = \int_0^\infty (1 - F(x))dx$. Why not use this as the definition of the (Lebesgue) integral, since $1 - F(x)$ is Riemann integrable?

15. *Chebyshev's inequality.* Suppose that X^p is integrable for $p \geq 1$. Then show that for any constant a ,

$$P[|X - a| \geq \epsilon] \leq \frac{E|X - a|^p}{\epsilon^p}$$

16. Is Chebyschev's inequality sharp? That is can we find a random variable X so that we have equality above, i.e. so that

$$P[|X - a| \geq \epsilon] = \frac{E|X - a|^p}{\epsilon^p}$$

17. Show that if $\mathcal{C}$ is the class of all random variables defined on some probability space (say the unit interval with the Borel sigma algebra),

    (a) if $\epsilon > 0$, $\inf\{P(|X| > \epsilon); X \in \mathcal{C}, E(X) = 0, var(X) = 1\} = 0$ and

    (b) if $y \geq 1$, $\inf\{P(|X| > y); X \in \mathcal{C}, E(X) = 1, var(X) = 1\} = 0$

18. A random variable Z has the *Standard normal distribution* if its density with respect to Lebesgue measure is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Then the price of a very simple non-dividend paying stock at time T is taken to be a random variable of the form

$$S_T = S_0 exp\{\mu T + \sqrt{T}\sigma Z\}$$

where $\mu = r - \frac{1}{2}\sigma^2$, r is the risk-free interest rate, σ the volatility or standard deviation per unit time, and Z is a random variable having the standard normal distribution.

(a) Find $E(S_T)$. Explain your answer.

(b) Find $E((S_T - K)^+)$ for a constant K. This is the price of a European call option having strike price K. (*Hint: Check that for for any choice of numbers* a, b, σ,

$$E(e^{\sigma Z} - e^{\sigma a})^+ = e^{\sigma^2/2} H(a - \sigma) - e^{\sigma a} H(a)$$

where $H(x)$ *is* $P[Z > x]$.)

19. Show that for any value of $t > 0$ and a random variable X with moment generating function m_X,

$$P[X > c] \leq e^{-tc} m_X(t)$$

20. A coin is tossed 5 times. Describe an appropriate probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Define random variables X =*number of heads in first 3 tosses* and Y = min(5, *number of tails before first head)*. Describe $\sigma(X)$ and $\sigma(X, Y)$ and show that $\sigma(X) \subset \sigma(X, Y)$. Determine the expected value and variance of $Y - X$.

21. Suppose you hold 1 option on a stock whose price at time T (the expiry date) is S_T with distribution given by

$$S_T = S_0 exp\{\mu T + \sqrt{T}\sigma Z\}$$

as in Question 18. We assume that the value of this option $E(S_T - K)^+ = V(S_0, T)$ is a function of the time to expiry and the current value of the stock. You wish also to hold $-\Delta$ units of the stock (Δ may be positive or negative). Find the value of Δ which minimizes the variance of the change in the portfolio; i.e. minimizing

$$var[\delta V - \Delta \delta S].$$

where δV is the change in the value of the option $V(S_T, 0) - V(S_0, T)$ and δS is the change in the value of the stock $S_T - S_0$.

Approximate δV by two terms of a Taylor series expansion $\delta V = \frac{\partial}{\partial S_0} V(S_0, T)\delta S - \frac{\partial}{\partial T} V(S_0, T)T$ and find an approximate value for the optimal choice of Δ. Suppose the linear approximation to δV is inadequate and we wish to use a quadratic approximation of the form

$$\delta V \approx a_T + b_T(S_T - ES_T) + c_T(S_T^2 - ES_T^2)$$

Then show that the optimal value of Δ is

$$\Delta = b_T + c_T\sqrt{Var(S_T)}\ \text{Skewness}(S_T).$$

22. *Bernstein polynomials.* If $g(p)$ is a continuous function on $[0, 1]$, then we may define $B_n(p) = E[g(X_{np}/n)]$ where $X_{np} \sim Bin(n, p)$. Show that $B_n(p) \to g(p)$ uniformly as $p \to \infty$. Note that the function $B_n(p)$ is a polynomial of degree n in p. This shows that any continuous function on a finite interval can be approximated uniformly by a polynomial. (*Hint: a continuous function on a compact interval* $[0, 1]$ *is uniformly continuous*).

23. In 1948 in a fundamental paper, C.E. Shannon defines the notion of entropy of a distribution as follows: Let X be a random variable with probability function or continuous probability density function $f(x)$. Suppose that the expectation $H(f) = E\{-\log(f(X))\}$ exists and is finite.

    (a) Prove that if g is the probability function of some function $h(X)$ of a discrete random variable X, then $H(g) \le H(f)$.

    (b) Prove that $H(f) \ge 0$.

24. Let μ be the measure on $\Re$ induced by the Poisson distribution with parameter 2. In other words if $p_n = P[X = n]$ where X has this Poisson distribution, define $\mu(A) = \sum\{p_n; n \in A\}$ for every Borel set $A \subset \Re$. Let λ be a similarly defined measure but with Poisson parameter 1. Show that $\mu << \lambda$ and find a function $f(x)$ such that

$$\mu(B) = \int_B f(x)d\lambda \tag{4.1}$$

for all Borel sets B. Is this function unique as a function on $\mathcal{R}$? How may it be modified while leaving property (4.1) unchanged?

25. Suppose two finite measures μ, λ defined on the same measurable space are not mutually singular. Prove that there exists $\varepsilon > 0$ and a set A with $\lambda(A) > 0$ such that

$$\varepsilon\lambda(E) \le \mu(E)$$

for all measurable sets $E \subset A. Hint :$ Solve this in the following steps:

(a) Consider the signed measure $\mu - n^{-1}\lambda$ for each value of $n = 1, 2, ...$ You may assume that you can decompose the probability space into disjoint sets A_n^- and A_n^+ such that $\mu(B) - n^{-1}\lambda(B) \le 0$ or ≥ 0 as $B \subset A_n^-$ or $B \subset A_n^+$ respectively (this is called the *Hahn decomposition*). Define

$$M = \cup A_n^+$$
$$M^c = \cap A_n^-.$$

Show that $\mu(M^c) = 0$.

(b) Show $\lambda(M) > 0$ and this implies $\lambda(A_n^+) > 0$ for some n.

(c) Finally conclude that $\frac{1}{n}\lambda(E) \le \mu(E)$ for all $E \subset A_n^+$.

26. (a) Find the moment generating function of a Binomial distribution.

    (b) Show that if the moment generating function has sufficiently many derivatives in a neighbourhood of the origin, we can use it to obtain the moments of X as follows:

$$E(X^p) = m_X^{(p)}(0), \quad p = 1, 2, \ldots$$

Show that the moments of the standard normal distribution are given by

$$E(Z) = 0, \quad E(Z^2) = 1, E(Z^3) = 0, \; E(Z^4) = 3, E(Z^{2n}) = \frac{(2n)!}{n!2^n}.$$

What is $E(Z^{2k})$?

27. Prove using only the definition of the expected value for simple random variables that if

$$\sum c_i I_{A_i} = \sum d_j I_{B_j}$$

then

$$\sum c_i P(A_i) = \sum d_j P(B_j)$$

28. Find an example of a random variable such that the $k'th$ moment exists i.e.

$$E(|X|^k) < \infty$$

but any higher moment does not, i.e.

$$E(|X|^{k+\epsilon}) = \infty \text{ for all } \epsilon > 0.$$

29. A city was designed entirely by probabilists so that traffic lights stay green for random periods of time (say $X_n, n = 1, 2, ...$) and then red for random periods (say $Y_n, n = 1, 2, ...$). There is no amber. Both X and Y have an exponential distribution with mean 1 minute and are independent. What is your expected delay if you arrive at the light at a random point of time?

30. Suppose that a random variable X has a moment generating function $m_X(t)$ which is finite on an interval $t \in [-\varepsilon, \varepsilon]$ for $\varepsilon > 0$. Prove rigorously that

$$E(X) = m'_X(0)$$

by interchanging a limit and an expected value.

31. A fair coin is tossed repeatedly. For each occurence of heads (say on the k'th toss) you win $\frac{2}{3^k}$, whereas for each occurrence of tails, you win nothing. Let

$$X = \text{total gain after infinitely many tosses.}$$

(a) What is the distribution of X. Is it discrete, absolutely continuous, or a mixture of the two?

(b) Find $E(X)$.

Chapter 5

Joint Distributions and Convergence

5.1 Product measures and Independence

In this section we duscuss the problem of constructing measures on a Cartesian product space and the properties that these measures possess. Such a discussion is essential if we wish to determine probabilities that depend on two or more random variables; for example calculating $P[|X-Y|>1]$ for random variables X, Y. First consider the analogous problem in $\Re^2$. Given Lebesgue measure λ on $\Re$ how would we construct a similar measure, compatible with the notion of area in two-dimensional Euclidean space? Clearly we can begin with the measure of rectangles or indeed any *product* set of the form $A\times B=\{(x,y);x\in A,\ y\in B\}$ for arbitrary Borel sets $A\subset\Re, B\subset\Re$. Clearly the measure of a product set $\mu(A\times B)=\lambda(A)\lambda(B)$. This defines a measure for any product set and by the extension theorem, since the product sets form a Boolean algebra, we can extend this measure to the sigma algebra generated by the product sets.

More formally, suppose we are given two measure spaces $(M,\mathcal{M},\mu)$ and $(N,\mathcal{N},\nu)$. Define the *product space* to be the space consisting of pairs of objects, one from each of M and N,

$$\Omega=M\times N=\{(x,y);\ x\in M,\ y\in N\}.$$

The Cartesian product of two sets $A\subset M,\quad B\subset N$ is denoted $A\times B=\{(a,b);a\in A,b\in B\}$. This is the analogue of a rectangle, a subset of $M\times N$, and it is easy to define a measure for such sets as follows. Define the *product measure* of product sets of the above form by $\pi(A\times B)=\mu(A)\nu(B)$. The following theorem is a simple consequence of the Caratheodory Extension Theorem.

Theorem 79 *The product measure π defined on the product sets of the form $\{A\times B;A\in\mathcal{N},\ \mathcal{B}\in\mathcal{M}\}$ can be extended to a measure on the sigma algebra*

$\sigma\{A \times B; A \in \mathcal{N}, \mathcal{B} \in \mathcal{M}\}$ *of subsets of* $M \times N$.

There are two cases of product measure of importance. Consider the sigma algebra on $\Re^2$ generated by the product of the Borel sigma algebras on $\Re$. This is called the Borel sigma algebra in $\Re^2$. We can similarly define the Borel sigma algebra on $\Re^n$.

In an analogous manner, if we are given two probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ we can construct a *product measure* Q on the Cartesian product space $\Omega_1 \times \Omega_2$ such that $Q(A \times B) = P_1(A)P_2(B)$ for all $A \in \mathcal{F}_1,\ B \in \mathcal{F}_2$.This guarantees the existence of a product probability space in which events of the form $A \times \Omega_2$ are independent of events of the form $\Omega_1 \times B$ for $A \in \mathcal{F}_1,\ B \in \mathcal{F}_2$.

Definition 80 *(Independence, identically distributed) A sequence of random variables* $X_1, X_2, \ldots$ *is* independent *if the family of sigma-algebras* $\sigma(X_1), \sigma(X_2), \ldots$ *are independent. This is equivalent to the requirement that for every finite set* $B_n, n = 1, \ldots N$ *of Borel subsets of* $\Re$*, the events* $[X_n \in B_n], n = 1, ..., N$ *form a mutually independent sequence of events. The sequence is said to be* identically distributed *every random variable* X_n *has the same c.d.f.*

Lemma 81 *If* X, Y *are independent integrable random variables on the same probability space, then* XY *is also integrable and*

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

Proof. Suppose first that X and Y are both simple functions, $X = \sum c_i I_{A_i}, Y = \sum d_j I_{B_j}$. Then X and Y are independent if and only if $P(A_i B_j) = P(A_i)P(B_j)$ for all i, j and so

$$\begin{aligned} E(XY) &= E[(\sum c_i I_{A_i})(\sum d_j I_{B_j})] \\ &= \sum \sum c_i d_j E(I_{A_i} I_{B_j}) \\ &= \sum \sum c_i d_j P(A_i)P(B_j) \\ &= E(X)E(Y). \end{aligned}$$

More generally suppose X, Y are non-negative random variables and consider independent simple functions X_n increasing to X and Y_n increasing to Y. Then $X_n Y_n$ is a non-decreasing sequence with limit XY. Therefore, by monotone convergence

$$E(X_n Y_n) \to E(XY).$$

On the other hand,

$$E(X_n Y_n) = E(X_n)E(Y_n) \to E(X)E(Y).$$

Therefore $E(XY) = E(X)E(Y)$. The case of general (positive and negative random variables X, Y we leave as a problem. ∎

5.1.1 Joint Distributions of more than 2 random variables.

Suppose $X_1, ..., X_n$ are random variables defined on the same probability space $(\Omega, \mathcal{F}, P)$. The joint distribution can be characterised by the *joint cumulative distribution function*, a function on $\Re^n$ defined by

$$F(x_1, ..., x_n) = P[X_1 \leq x_1, ..., X_n \leq x_n] = P([X_1 \leq x_1] \cap .... \cap [X_n \leq x_n]).$$

Example 82 *Suppose $n = 2$. Express the probability*

$$P[a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2]$$

using the joint cumulative distribution function.

Notice that the joint cumulative distribution function allows us to find $P[a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n]$. Using inclusion-exclusion,

$$\begin{aligned} & P[a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n] \\ & \quad = F(b_1, b_2, \ldots b_n) - \sum_j F(b_1, ..., a_j, b_{j+1}, ...b_n) \\ & \quad + \sum_{i<j} F(b_1, ..., a_i, b_{i+1}...a_j, b_{j+1}, ...b_n) - ... \end{aligned} \tag{5.1}$$

As in the case $n = 1$, we may then build a probability measure on an algebra of subsets of $\Re^n$. This measure is then extended to the Borel sigma-algebra on $\Re^n$.

Theorem 83 *The joint cumulative distribution function has the following properties:*

(a) $F(x_1, ..., x_n)$ is right-continuous and non-decreasing in each argument x_i when the other arguments x_j, $j \neq i$ are fixed.

(b) $F(x_1, ..., x_n) \to 1$ as $min(x_1, ..., x_n) \to \infty$ and $F(x_1, ..., x_n) \to 0$ as $\min(x_1, ..., x_n) \to -\infty$.

(c) The expression on the right hand side of (5.1) is greater than or equal to zero for all $a_1, \ldots a_n$, $b_1, \ldots, b_n$.

The joint probability distribution of the variables $X_1, ..., X_n$ is a measure on $\mathcal{R}^n$. It can be determined from the cumulative distribution function since (5.1) gives the measure of rectangles, these form a pi-system in $\mathcal{R}^n$ and this permits extension first to an algebra and then the sigma algebra generated by these intervals. This sigma algebra is the Borel sigma algebra in $\mathcal{R}^n$. Therefore, in order to verify that the random variables are mutually independent, it is sufficient to verify that the joint cumulative distribution function factors;

$$F(x_1, ..., x_n) = F_1(x_1)F_2(x_2)...F_n(x_n) = P[X_1 \leq x_1] \ldots P[X_n \leq x_n]$$

for all $x_1, ..., x_n \in \Re$.

The next theorem is an immediate consequence of Lemma 28 and the fact that $X_1, ..., X_n$ independent implies that $g_1(X_1), g_2(X_2), ..., g_n(X_n)$ are independent for arbitrary measurable functions $g_i, i = 1, ..., n$.

Theorem 84 *If the random variables $X_1, ..., X_n$ are mutually independent, then*

$$E[\prod g_j(X_j)] = \prod E[g_j(X_j)]$$

for any Borel measurable functions $g_1, \ldots, g_n$.

We say an infinite sequence of random variables $X_1, X_2, \ldots$ is mutually independent if every finite subset is mutually independent.

5.2 Strong (almost sure) Convergence

Definition 85 *Let X and $X_n, n = 1, 2, \ldots$ be random variables all defined on the same probability space $(\Omega, \mathcal{F})$. We say that the sequence X_n converges* almost surely *(or* with probability one*) to X (denoted $X_n \to X$ a.s.) if the event*

$$\{\omega; X_n(\omega) \to X(\omega)\} = \cap_{m=1}^{\infty}[|X_n - X| \le \frac{1}{m} \quad a.b.f.o.]$$

has probability one.

In order to show a sequence X_n converges almost surely, we need that X_n are (measurable) random variables for all n, and to show that there is some measurable random variable X for which the set $\{\omega; X_n(\omega) \to X(\omega)\}$ is measurable and hence an event, and that the probability of this event $P[X_n \to X]$ is 1.Alternatively we can show that for each value of $\epsilon > 0$, $P[|X_n - X| > \epsilon$ i.o.$] = 0$. It is sufficient, of course, to consider values of ϵ of the form $\epsilon = 1/m$, m=1,2,... above.

The law of large numbers (sometimes called the law of averages) is the single most immportant and well-known result in probability. There are many versions of it but the following is sufficient, for example, to show that the average of independent Bernoulli random variables, or Poisson, or negative binomial, or Gamma random variables, to name a few, converge to their expected value **with probability one**.

Theorem 86 *(Strong Law of Large Numbers) If X_n, $n = 1, 2, \ldots$ is a sequence of independent identically distributed random variables with $E|X_n| < \infty$, (i.e. they are integrable) and $E(X_n) = \mu$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad a.s. \quad as \quad n \to \infty$$

Proof. We shall prove this result in the special case $E(X^4) < \infty$. The more general proof will be left for later. First note that, by replacing X_i by $X_i - \mu$ we may assume that $\mu = 0$ without any loss of generality. Now note that with $S_n = \sum_{i=1}^n X_i$, and letting $var(X_i) = \sigma^2$ and $E(X_i^4) = K$, we have

$$E(S_n^4) = nK + 3n(n-1)\sigma^4 = d_n, \quad say.$$

Therefore for each $\epsilon > 0$, we have by Chebyschev's inequality

$$P\{|\frac{S_n}{n}| > \epsilon\} \leq \frac{E(S_n^4)}{\epsilon^4 n^4} = \frac{d_n}{\epsilon^4 n^4}$$

Note that since $\sum_n \frac{d_n}{n^4} < \infty$ we have by the first Borel Cantelli Lemma,

$$P\{|\frac{S_n}{n}| > \epsilon \ i.o.\} = 0.$$

Since this holds for all $\epsilon > 0$ it follows that the probability that $\frac{S_n}{n}$ does not converge to 0 is 0 and so the probability that it does converge is 1. ■

5.3 Weak Convergence (Convergence in Distribution)

Consider random variables that are constants; $X_n = 1 + \frac{1}{n}$. By any sensible definition of convergence, X_n converges to $X = 1$ as $n \to \infty$. Does the cumulative distribution function of X_n, F_n,say, converge to the cumulative distribution function of X pointwise? In this case it is true that $F_n(x) \to F(x)$ at all values of x except the value $x = 1$ where the function $F(x)$ has a discontinuity. Convergence in distribution (weak convergence, convergence in Law) is defined as pointwise convergence of the c.d.f. at all values of x except those at which $F(x)$ is discontinuous. Of course if the limiting distribution is absolutely continuous (for example the normal distribution as in the Central Limit Theorem), then $F_n(x) \to F(x)$ does hold for all values of x.

Definition 87 *(Weak Convergence) If $F_n(x)$, $n = 1, \ldots$ is a sequence of cumulative distribution functions and if F is a cumulative distribution function, we say that F_n converges to F* weakly *or* in distribution *if $F_n(x) \to F(x)$ for all x at which $F(x)$ is continuous. We will sometimes denote weak convergence of a sequence of random variables X_n whose c.d.f. converges in the above sense by $X_n \Rightarrow X$.*

Example 88 *(Maximum of independent exponential(α)) Suppose $(X_1, ..., X_n)$ are independent exponentially distributed random variables all with the exponential cumulative distribution function*

$$F(x) = 1 - e^{-\alpha x}.$$

Define $M_n = max(X_1, ..., X_n)$ *. Then the c.d.f. of* $M_n - (log\ n)/\alpha$ *is*

$$F_{M_n}(x) = (1 - e^{-(\alpha x + log\ n)})^n \rightarrow F(x) = e^{-e^{-\alpha x}}$$

Proof. Note that for arbitrary $x \in \mathcal{R}$

$$\begin{aligned} P[M_n - \frac{\ln n}{\alpha} \leq x] = P[M_n \leq x + \frac{\ln n}{\alpha}] &= [F(x + \frac{\ln n}{\alpha})]^n \\ &= (1 - e^{-\alpha x - \ln n})^n = (1 - \frac{1}{n}e^{-\alpha x})^n \\ &\rightarrow \exp(-e^{-\alpha x}) \text{ as } n \rightarrow \infty. \end{aligned}$$

■

For any independent identically distributed random variables such that the cumulative distribution function satisfies $1 - F(x) \sim e^{-\alpha x}$, the same result holds. The limiting distribution whose cumulative distribution function is of the form $F(x) = \exp(-e^{-\alpha x})$ is called an *extreme value distribution* and is commonly used in environmental, biostatistical and engineering applications of statistics. The corresponding probability density function is

$$\frac{d}{dx}e^{-e^{-\alpha x}} = \alpha \exp\left(-\alpha x - e^{-\alpha x}\right), -\infty < x < \infty$$

and is shaped like a slightly skewed version of the normal density function (see Figure 1 for the case $\alpha = 2$).

This example also shows approximately how large a maximum will be since $M_n - (\ln\ n)/\alpha$ converges to a proper distribution. Theoretically, if there were no improvement in training techniques over time, for example, we would expect that the world record in the high jump or the shot put at time t (assuming the number of competitors and events occurred at a constant rate) to increase like $ln(t)$. However, records in general have increased at a much higher rate, indicating higher levels of performance, rather than just the effect of the larger number of events over time. Similarly, record high temperatures since records in North America were begun increase at a higher rate than this, providing evidence of global warming.

Example 89 *Suppose* $1 - F(x) \sim x^{-\alpha}$ *for* $\alpha > 0$. *Then the cumulative distribution function of* $n^{-1/\alpha} M_n$ *converges weakly to* $F(x) = e^{-x^{-\alpha}}$, $x > 0$ *(The distribution with the cumulative distribution function* $F(x) = e^{-x^{-\alpha}}$ *is called the* Weibull *distribution).*

Proof. The proof of the convergence to a Weibull is similar to that for the extreme value distribution above.

$$\begin{aligned} P[n^{-1/\alpha} M_n \le x] &= [F(n^{1/\alpha} x)]^n \\ &= [1 - (n^{1/\alpha} x)^{-\alpha} + o(n^{-1})]^n \\ &= [1 - \frac{1}{n} x^{-\alpha} + o(n^{-1})]^n \\ &\to \exp(-x^{-\alpha}) \text{ as } n \to \infty \end{aligned}$$

■

We have used a slight extension of the well-known result that $(1+c/n)^n \to e^c$ as $n \to \infty$. This result continues to hold even if we include in the bracket and additional term $o(n^{-1})$ which satisfies $no(n^{-1}) \to 0$. The extension that has been used (and is easily proven) is $(1 + c/n + o(n^{-1}))^n \to e^c$ as $n \to \infty$.

Example 90 *Find a sequence of cumulative distribution functions* $F_n(x) \to F(x)$ *for some limiting function* $F(x)$ *where this limit is not a proper c.d.f.*

There are many simple examples of cumulative distribution functions that converge pointwise but not to a genuine c.d.f. All involve some of the mass of the distribution "excaping" to infinity. For example consider F_n the $N(0, n)$ cumulative distribution function. Of more simply, use F_n the cumulative distribution function of a point mass at the point n. However there is an additional condition that is often applied which insures that the limiting distribution is a "proper" probability distribution (i.e. has total measure 1). This condition is called tightness.

Definition 91 *A sequence of probability measures* P_n *on a measurable metric space is tight if for all* $\epsilon > 0$, *there exists a compact set* K *such that* $P_n(K^c) \le \epsilon$ *for all* n.

A sequence of cumulative distribution functions F_n is tight if it corresponds to a sequence of tight probability measures on $\mathcal{R}$. This is equivalent to the requirement that for every $\epsilon > 0$, there is a value of $M < \infty$ such that the probabilities outside the interval $[-M, M]$ are less than ϵ. In other words if

$$F_n(-M-) + (1 - F_n(M)) \le \epsilon \quad \text{for all } n = 1, 2, ...$$

If a sequence F_n converges to some limiting right-continuous function F at continuity points of F and if the sequence is tight, then F is a c.d.f. of a probability distribution and the convergence is in distribution or weak (see Problem 6).

Lemma 92 *If* $X_n \Rightarrow X$, *then there is a sequence of random variables* Y, Y_n *on some other probability space (for example the unit interval) such that* Y_n *has the same distribution as* X_n *and* Y *has the same distribution as* X *but* $Y_n \to Y$ *almost surely.*

Proof. Suppose we take a single uniform[0,1] random variable U. Recall the definition of pseudo inverse used in Theorem 20, $F^{-1}(y) = \sup\{z; F(z) < y\}$. Define $Y_n = F_n^{-1}(U)$ and $Y = F^{-1}(U)$ where F_n and F are the cumulative distribution functions of X_n and X respectively. We need to show that if $F_n(x) \to F(x)$ at all x which are continuity points of the function F, then $F_n^{-1}(U) \to F^{-1}(U)$ almost surely. First note that the set of $y \in [0,1]$ such that $F^{-1}(y)$ is NOT a

point of increase of the function has Lebesgue measure 0. For any point x which is a continuity point and a point of increase of F, we can find a closed neighbourhood around x, say of the form $[x-b, x+b]$ such that convergence hold uniformly in this neighbourhood. This means that for any $\varepsilon > 0$, there is a value of N so that $|F_n(z) - F(z)| \leq \epsilon$ whenever $n > N$. This in turn implies that, with $y = F(x)$, that $|F_n^{-1}(y) - F^{-1}(y)| \leq \delta$ where $\delta = \max(x - F^{-1}(y-\epsilon), F^{-1}(y+\epsilon) - x)$. It follows that $F_n^{-1}(U)$ converges almost surely to $F^{-1}(U)$ ■

Theorem 93 *Suppose* $X_n \Rightarrow X$ *and* g *is a Borel measurable function. Define* $D_g = \{x; g \text{ discontinuous at } x\}$. *If* $P[X \in D_g] = 0$, *then* $g(X_n) \Rightarrow g(X)$.

Proof. We prove this result assuming the last lemma which states that we can find a sequence of random variables Y_n and a random varaible Y which have the same distribution as X_n, X respectively but such that Y_n converges almost surely (i.e. with probability one) to Y. Note that in this case

$$g(Y_n(\omega)) \to g(Y(\omega))$$

provided that the function g is continuous at the point $Y(\omega)$ or in other words, provided that $Y(\omega) \notin D_g$. Since $P[Y(\omega) \notin D_g] = 1$, we have that

$$g(Y_n) \to g(Y) \text{ a.s.}$$

and therefore convergence holds also in distribution (you may either use Theorems 35 and 36 or prove this fact seperately). But since Y_n and X_n have the same distribution, so too do $g(Y_n)$ and $g(X_n)$ implying that $g(X_n)$ converges in distribution to $g(X)$. ■

In many applications of probability, we wish to consider stochastic processes $X_n(t)$ and their convergence to a possible limit. For example, suppose $X_n(t)$ is defined to be a random walk on discrete time, with time steps $1/n$ and we wish to consider a limiting distribution of this process as $n \to \infty$. Since X_n is a stochastic process, not a random variable, it does not have a cumulative distribution function, and any notion of weak convergence must not rely on the c.d.f. In this case, the following theorem is used as a basis for defining weak convergence. In general, we say that X_n *converges weakly to* X *if* $E[f(X_n)] \to E[f(X)]$ *for all bounded continuous functions* f. This is a more general definition of weak convergence.

Definition 94 *(general definition of weak convergence) A sequence of random elements of a metric space* X_n *converges weakly to* X *i.e.* $X_n \Rightarrow X$ *if and only if* $E[f(X_n)] \to E[f(X)]$ *for all bounded continuous functions* f.

Theorem 95 *If* X_n *and* X *are random variables,* X_n *converges weakly to* X *if and only if* $F_n(x) \to F(x)$ *for all* $x \notin D_F$.

Proof. The proof is based on lemma 32. Consider a sequence Y_n such that Y_n and X_n have the same distribution but $Y_n \to Y$ almost surely. Since $f(Y_n)$ is bounded above by a constant (and the expected value of a constant is finite), we have by the Dominated Convergence Theorem $Ef(Y_n) \to Ef(Y)$. (We have used here a slightly more general version of the dominated convergence theorem in which convergence holds almost surely rather than pointwise at all ω.) For the converse direction, assume $E[f(X_n)] \to E[f(X)]$ for all bounded continuous functions f. Suppose we take the function

$$f_\epsilon(t) = \begin{cases} 1, & t \le x \\ 0, & t > x + \epsilon \\ \frac{x+\epsilon-t}{\epsilon}, & x < t < x + \epsilon \end{cases}$$

Assume that x is a continuity point of the c.d.f. of X. Then $E[f_\epsilon(X_n)] \to E[f_\epsilon(X)]$. We may now take $\epsilon \to 0$ to get the convergence of the c.d.f. ■

5.4 Convergence in Probability

Definition 96 *We say a sequence of random variables* $X_n \to X$ in probability *if for all* $\epsilon > 0$, $P[|X_n - X| > \epsilon] \to 0$ *as* $n \to \infty$.

Convergence in probability is in general a somewhat more demanding concept than weak convergence, but less demanding than almost sure convergence. In other words, convergence almost surely implies convergence in probability and convergence in probability implies weak convergence.

Theorem 97 *If* $X_n \to X$ almost surely *then* $X_n \to X$ in probability.

Proof. Because we can replace X_n by $X_n - X$, we may assume without any loss of generality that $X = 0$. Then the set on which X_n converges almost surely to zero is

$$\{\omega; X_n(\omega) \to 0\} = \cap_{m=1}^{\infty} ([|X_n| \le 1/m] a.b.f.o.)$$

and so for each $\epsilon = 1/m > 0$, we have, since X_n converges almost surely,

$$P([|X_n| \le \epsilon] a.b.f.o.) = 1.$$

or

$$1 = P(\cup_{j=1}^{\infty} \cap_{n=j}^{\infty} [|X_n| \le \epsilon]) = lim_{j\to\infty} P(\cap_{n=j}^{\infty}[|X_n| \le \epsilon]).$$

Since $P(\cap_{n=j}^{\infty}[|X_n| \le \epsilon]) \le P[|X_j| \le \epsilon]$ it must follow that $P[|X_j| \le \epsilon] \to 1$ as $j \to \infty$. ■

Convergence in probability does not imply convergence almost surely. For example let $\Omega = [0,1]$ and for each n write it uniquely in the form $n = 2^m + j$ for $0 \le j < 2^m$. Define $X_n(\omega) = 1$ if $j/2^m \le \omega \le (j+1)/2^m$ so that X_n is the indicator of the interval $[j/2^m, (j+1)/2^m]$. Then X_n converges in probability to 0 but $P[X_n \to 0] = 0$.

Theorem 98 *If* $X_n \to X$ in probability, *then* $X_n \Rightarrow X$.

Proof. Assuming convergence in probability, we need to show that $P[X_n \le x] \to P[X \le x]$ whenever x is a continuity point of the function on the right hand side. Note that

$$P[X_n \le x] \le P[X \le x + \epsilon] + P[|X_n - X| > \epsilon]$$

for any $\epsilon > 0$. Taking limits on both sides as $n \to \infty$, we obtain

$$limsup_{n\to\infty} P[X_n \le x] \le P[X \le x + \epsilon].$$

By a similar argument

$$liminf_{n\to\infty} P[X_n \le x] \ge P[X \le x - \epsilon].$$

Now since $\epsilon > 0$ was arbitrary, we may take it as close as we wish to 0. and since the function $F(x) = P[X \le x]$ is continuous at the point x , the limit as $\epsilon \to 0$ of both $P[X \le x + \epsilon]$ and $P[X \le x - \epsilon]$ is $F(x)$. It follows that

$$F(x) \le liminf P[X_n \le x] \le limsup P[X_n \le x] \le F(x)$$

and therefore $P[X_n \le x] \to F(x)$ as $n \to \infty$. ■

Theorem 99 *If* $X_n \Rightarrow c$ i.e. in distribution *for some constant* c *then* $X_n \to c$ *in probability.*

Proof. Since the c.d.f. of the constant c is $F(x) = 0,\ \ x < c,\quad F(x) = 1,\ \ x \ge c$, and is continuous at all points except the point $x = c$, we have, by the convergence in distribution,

$$P[X_n \le c + \epsilon] \to 1 \quad \textit{and} \quad P[X_n \le c - \epsilon] \to 0$$

for all $\epsilon > 0$. Therefore,

$$P[|X_n - c| > \epsilon] \le (1 - P[X_n \le c + \epsilon]) + P[X_n \le c - \epsilon] \to 0.$$

■

Theorem 100 *If* $X_n, \quad n = 1, \ldots$ *and* $Y_n, \ n = 1, \ldots$ *are two sequences of random variables such that* $X_n \Rightarrow X$ *and* $Y_n - X_n \Rightarrow 0$ *, then* $Y_n \Rightarrow X$*.*

Proof. Assume that $F(x)$, the c.d.f of X is continuous at a given point x. Then for $\epsilon > 0$,

$$P[Y_n \leq x - \epsilon] \leq P[X_n \leq x] + P[|X_n - Y_n| > \epsilon].$$

Now take limit supremum as $n \to \infty$ to obtain

$$limsup P[Y_n \leq x - \epsilon] \leq F(x).$$

A similar argument gives

$$liminf P[Y_n \leq x + \epsilon] \geq F(x).$$

Since this is true for ϵ arbitrarily close to 0, $P[Y_n \leq x] \to F(x)$ as $n \to \infty$. ■

Theorem 101 *(A Weak Law of Large Numbers) If* $X_n, \quad n = 1, 2, \ldots$ *is a sequence of independent random variables all with the same expected value* $E(X_n) = \mu$*, and if their variances satisfy* $\frac{1}{n^2}\sum_{i=1}^n var(X_i) \to 0$ *, then* $\frac{1}{n}\sum_{i=1}^n X_i \to \mu$ *in probability.*

Proof. By Chebyschev's inequality,

$$P[|\frac{\sum_{i=1}^n X_i}{n} - \mu| \geq \epsilon] \leq \frac{\sum_{i=1}^n var(X_i)}{\epsilon^2 n^2}$$

and this converges to 0 by the assumptions. ■

5.5 Fubini's Theorem and Convolutions.

Theorem 102 *(Fubini's Theorem) Suppose* $g(x, y)$ *is integrable with respect to a product measure* $\pi = \mu \times \nu$ *on* $M \times N$ *. Then*

$$\int_{M\times N} g(x, y) d\pi = \int_M [\int_N g(x, y) d\nu] d\mu = \int_N [\int_M g(x, y) d\mu] d\nu.$$

We can dispense with the assumption that the function $g(x, y)$is integrable in Fubini's theorem (permiting infinite integrals) if we assume instead that $g(x, y) \geq 0$.

Proof. First we need to identify some measurability requirements. Suppose E is a set measurable with respect to the product sigma-algebra on $M \times N$. We need to first show that the set $E_y = \{x \in M; (x, y) \in E\}$ is a measurable set in M. Consider the class of sets

$$\mathcal{C} = \{E; \{x \in M; (x, y) \in E\} \text{ is measurable in the product sigma algebra}\}$$

It is easy to see that $\mathcal{C}$ contains all product sets of the form $A \times B$ and that it satisfies the properties of a sigma-algebra. Therefore, since the product sigma algebra is generated by $\{A \times B; A \in \mathcal{N}, B \in \mathcal{M}\}$, it is contained in $\mathcal{C}$. This shows that sets of the form E_y are measurable. Now define the measure of these sets $h(y) = \mu(E_y)$. The function $h(y)$ is a measurable function defined on N (see Problem 23).

Now consider a function $g(x,y) = I_E$ where $E \in \mathcal{F}$. The above argument is needed to show that the function $h(y) = \int_M g(x,y)d\mu$ is measurable so that the integral $\int_N h(y)d\nu$ potentially makes sense. Finally note that for a set E of the form $A \times B$,

$$\int I_E d\pi = \pi(A \times B) = \mu(A)\nu(B) = \int_N (\int_M I_E(x,y)d\mu)d\nu$$

and so the condition $\int I_E d\pi = \int_N(\int_M I_E(x,y)d\mu)d\nu$ holds for sets E that are product sets. It follows that this equality holds for all sets $E \in \mathcal{F}$(see problem 24). Therefore this holds also when I_E is replaced by a simple function. Finally we can extend this result to an arbitrary non-negative function g by using the fact that it holds for simple functions and defining a sequence of simple functions $g_n \uparrow g$ and using monotone convergence. ■

Example 103 *The formula for integration by parts is*

$$\int_{(a,b]} G(x)dF(x) \;=\; G(b)F(b) - G(a)F(a) - \int_{(a,b]} F(x)dG(x)$$

Does this formula apply if $F(x)$ is the cumulative distribution function of a of a constant z in the interval $(a,b]$ and the function G has a discontinuity at the point z?

Lemma 104 *(Integration by Parts) If F, G are two monotone right continuous functions on the real line having* no common discontinuities*, then*

$$\int_{(a,b]} G(x)dF(x) \;=\; G(b)F(b) - G(a)F(a) - \int_{(a,b]} F(x)dG(x)$$

5.5.1 Convolutions

Consider two independent random variables X, Y, both having a discrete distribution. Suppose we wish to find the probability function of the sum $Z = X + Y$. Then

$$P[Z = z] = \sum_x P[X = x]P[Y = z - x] = \sum_x f_X(x)f_Y(z-x).$$

Similarly, if X, Y are independent absolutely continuous distributions with probability density functions f_X, f_Y respectively, then we find the probability density function of the sum $Z = X + Y$ by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$$

In both the discrete and continuous case, we can rewrite the above in terms of the cumulative distribution function F_Z of Z. In either case,

$$F_Z(z) = E[F_Y(z - X)] = \int_{\Re} F_Y(z - x)F_X(dx)$$

We use the last form as a more general definition of a *convolution* between two cumulative distribution functions F, G . We define the *convolution* of F and G to be $F * G(x) = \int_{-\infty}^{\infty} F(x - y)dG(y)$.

5.5.2 Properties.

(a) If F, G are cumulative distributions functions, then so is $F*G$ (Problem 5.25)

(b) $F * G = G * F$ (Problem 6.3)

(c) If either F or G are absolutely continuous with respect to Lebesgue measure, then $F * G$ is absolutely continuous with respect to Lebesgue measure.

The convolution of two cumulative distribution functions $F * G$ represents the c.d.f of the sum of two independent random variables, one with c.d.f. F and the other with c.d.f. G. The next theorem says that if we have two independent sequences X_n independent of Y_n and $X_n \Longrightarrow X$, $Y_n \Longrightarrow Y$, then the pair (X_n, Y_n) converge weakly to the joint distribution of two random variables (X, Y) where X and Y are independent. There is an easier proof available using the characteristic functions in Chapter 6.

Theorem 105 *If* $F_n \Rightarrow F$ *and* $G_n \Rightarrow G$, *then* $F_n * G_n \Rightarrow F * G$.

Proof.
First suppose that X, X_n, Y, Y_n have cumulative distribution functions given by F, F_n, G, G_n respectively and denote the set of points at which a function such as F is discontinuous by D_F. Recall that by Lemma 32, we may redefine the random variables Y_n and Y so that $Y_n \to Y$ almost surely. Now choose a point $z \notin D_{F*G}$. We wish to show that

$$F_n * G_n(z) \to F * G(z)$$

for all such z. Note that since $F * G$ is the cumulative distribution function of $X + Y$, $z \notin D_{F*G}$ implies that

$$0 = P[X + Y = z] \geq \sum_{x \epsilon D_F} P[Y = z - x]P[X = x].$$

so $P[Y = z - x] = 0$ whenever $P[X = x] > 0$, implying $P[Y \in z - D_F] = 0$. Therefore the set $[Y \notin z - D_F]$ has probability one, and on this set, since $z - Y_n \to z - Y$ almost surely, we also have $F_n(z - Y_n) \to F(z - Y)$ almost surely. It follows from the dominated convergence theorem (since $F_n(z - Y_n)$ is bounded above by 1) that

$$F_n * G_n(z) = E(F_n(z - Y_n)) \to E(F(z - Y)) = F * G(z)$$

■

5.6 Problems

1. Prove that if X, Y are independent random variables, $E(XY) = E(X)E(Y)$(Lemma 28). Are there are random variables X, Y such that $E(XY) = E(X)E(Y)$ but X, Y are not independent? What if X and Y only take two possible values?

2. Find two absolutely continuous random variables such that the joint distribution (X, Y) is not absolutely continuous.

3. If two random variables X, Y has joint probability density function $f(x, y)$ show that the joint density function of $U = X + Y$ and $V = X - Y$ is

$$f_{U,V}(u, v) = \frac{1}{2} f_{X,Y}(\frac{u+v}{2}, \frac{u-v}{2}).$$

4. If X_n is a sequence of non-negative random variables, show that the set of

$$\{\omega; X_n(\omega) \ \text{converges}\} = \cap_{m=1}^{\infty} \cup_{N=1}^{\infty} \cap_{j=N}^{\infty} \cap_{n=N}^{\infty} [|X_n - X_j| \leq \frac{1}{m}]$$

5. Give an example of a sequence of random variables X_n defined on $\Omega = [0, 1]$ which converges in probability but does not converge almost surely. Is there an example of the reverse (i.e. the sequence converges almost surely but not in probability)? If X_n is a Binomial (n, p) random variable for each n, in what sense does $n^{-1} X_n$ converge to p as $n \to \infty$?

6. Suppose that F_n is a sequence of c.d.f.'s converging to a right continuous function F at all continuity points of F. Prove that if the sequence has the property that for every $\epsilon > 0$ there exists $M < \infty$ such that $F_n(-M) + 1 - F_n(M) < \epsilon$ for all n, then F must be a proper cumulative distribution function.

7. Prove directly (using only the definitions of almost sure and weak convergence) that if X_n is a sequence of random variables such that $X_n \to X$ almost surely, then $X_n \Longrightarrow X$ (convergence holds weakly).

8. Prove that if X_n converges in distribution (weakly) to a constant $c > 0$ and $Y_n \Rightarrow Y$ for a random variable Y, then $Y_n / X_n \Rightarrow Y/c$. Show also that if $g(x, y)$ is a continuous function of (x, y), then $g(X_n, Y_n) \Rightarrow g(c, Y)$.

9. Prove that if $X_n \Rightarrow X$ then there exist random variables Y_n, Y with the same distribution as X_n, X respectively such that $Y_n \to Y$ a.s. (Lemma 32).

10. Prove that if X_n converges with probability 1 to a random variable X then it converges in distribution to X (Theorem 36).

11. Suppose $X_i, i = 1, 2, ...$ are independent identically distributed random variables with finite mean and variance $var(X_i) = \sigma^2$.Let $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Prove that

$$\frac{1}{n-1}\sum (X_i - \overline{X_n})^2 \to \sigma^2 \text{ almost surely as } n \to \infty.$$

12. A multivariate c.d.f. $F(\mathbf{x})$ of a random vector $\mathbf{X} = (\mathbf{X_1}, ..., \mathbf{X_n})$ is *discrete* if there are countably many points $\mathbf{y_j}$ such that

$$\sum_j P[\mathbf{X} = \mathbf{y_j}] = \mathbf{1}.$$

Prove that a multivariate distribution function is discrete if and only if its marginal distribution functions are all discrete.

13. Let X_n, $n = 1, 2, \ldots$ be independent positive random variables all having a distribution with probability density function $f(x)$, $x > 0$. Suppose $f(x) \to c > 0$ as $x \to 0$. Define the random variable

$$Y_n = min(X_1, X_2, \ldots X_n).$$

   (a) Show that $Y_n \to 0$ in probability.

   (b) Show that nY_n converges in distribution to an exponential distribution with mean $1/c$.

14. *Continuity* Suppose X_t is, for each $t \in [a, b]$, a random variable defined on Ω. Suppose for each $\omega \in \Omega$, $X_t(\omega)$ is continuous as a function of t for $t \in [a, b]$.

    If for all $t \in [a, b]$, $|X_t(\omega)| \le Y(\omega)$ for all $\omega \in \Omega$, where Y is some integrable random variable, prove that $g(t) = E(X_t)$ is a continuous function of t in the interval $[a, b]$.

15. *Differentiation under Integral.* Suppose for each $\omega \in \Omega$ that the derivative $\frac{d}{dt}X_t(\omega)$ exists and $|\frac{d}{dt}X_t(\omega)| \le Y(\omega)$ for all $t \in [a, b]$, where Y is an integrable random variable. Then show that

$$\frac{d}{dt}E(X_t) = E[\frac{d}{dt}X_t]$$

16. Find the moment generating function of the *Gamma distribution* having probability density function

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}, \; x > 0$$

and show that the sum of n independent identically distributed Gamma (α, β) random variables is Gamma $(n\alpha, \beta)$. Use this fact to show that the moment generating function of the random variable

$$Z^* = \frac{\sum_{i=1}^n X_i - n\alpha\beta}{\sqrt{n\alpha\beta^2}}$$

approaches the moment generating function of the standard normal distribution as $n \to \infty$ and thus that $Z^* \Rightarrow Z \sim N(0, 1)$.

17. Let $X_1, \ldots X_n$ be independent identically distributed random variables with the uniform distribution on the interval $[0, b]$. Show convergence in distribution of the random variable

$$Y_n = n \; \min(X_1, X_2, \ldots, X_n)$$

and identify the limiting distribution.

18. Assume that the value of a stock at time n is given by

$$S_n = c(n) exp\{2X_n\}$$

where X_n has a binomial distribution with parameters (n, p) and $c(n)$ is a sequence of constants. Find $c(n)$ so that the expected value of the stock at time n is the risk-free rate of return e^{rn} . Consider the present value of a call option on this stock which has exercise price K.

$$V = e^{-rn}E\{max(S_n - K, 0)\}.$$

Show, using the weak convergence of the binomial distribution to the normal, that this expectation approaches a similar quantity for a normal random variable.

19. The usual student t-statistic is given by a form

$$t_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n}$$

where $\bar{X}_n$, s_n are the sample mean and standard deviation respectively. It is known that

$$z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

converges in distribution to a standard normal (N(0,1)) and that $s_n \to \sigma$ in probability. Show that t_n converges in distribution to the standard normal.

20. Let $X_1, X_2, \ldots, X_{2n+1}$ be independent identically distributed $U[0,1]$ random variables . Define $M_n = \textit{median}\ (X_1, X_2, \ldots, X_{2n+1})$. Show that $M_n \to \frac{1}{2}$ in probability and almost surely as $n \to \infty$.

21. We say that $X_n \to X$ in L_p for some $p \geq 1$ if

$$E(|X_n - X|^p) \to 0$$

as $n \to \infty$. Show that if $X_n \to X$ in L_p then $X_n \to X$ in probability. Is the converse true?

22. If $X_n \to 0$ in probability, show that there exists a subsequence n_k such that $X_{n_k} \to 0$ almost surely as $k \to \infty$.

23. Consider the product space $\{M \times N, \mathcal{F}, \pi)$ of two measure spaces $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$. Consider a set $E \in \mathcal{F}$ and define $E_y = \{x \in M; (x,y) \in E\}$.This is a measurable set in M. Now define the measure of these sets $g(y) = \mu(E_y)$. Show that the function $g(y)$ is a measurable function defined on N.

24. Consider the product space $\{M \times N, \mathcal{F}, \pi)$ of two measure spaces $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$. Suppose we verify that for all $E = A \times B$,

$$\pi(E) = \int_N \left(\int_M I_E(x,y) d\mu \right) d\nu. \tag{5.2}$$

Prove that (5.2) holds for all $E \in \mathcal{F}$.

25. Prove that if F, G are cumulative distributions functions, then so is $F * G$.

26. Prove: If either F or G are absolutely continuous with respect to Lebesgue measure, then $F * G$ is absolutely continuous with respect to Lebesgue measure.

Chapter 6

Characteristic Functions and the Central Limit Theorem

6.1 Characteristic Functions

6.1.1 Transforms and Characteristic Functions.

There are several transforms or generating functions used in mathematics, probability and statistics. In general, they are all integrals of an exponential function, which has the advantage that it converts sums to products. They are all functions defined on $t \in \Re$. In this section we use the notation $i = \sqrt{-1}$. For example;

1. *(Probability) Generating function.* $g(s) = E(s^X)$.
2. *Moment Generating Function.* $m(t) = E[e^{tX}] = \int e^{tx} dF$
3. *Laplace Transform.* $\mathcal{L}(t) = E[e^{-tX}] = \int e^{-tx} dF$
4. *Fourier Transform.* $E[e^{-itX}] = \int e^{-itx} dF$
5. *Characteristic function.* $\varphi_X(t) = E[e^{itX}] = \int e^{itx} dF$

Definition 106 *(*Characteristic Function) *Define the characteristic function of a random variable X or or its cumulative distribution function F_X to be the complex-valued function on $t \in \Re$*

$$\varphi_X(t) = E[e^{itX}] = \int e^{itx} dF = E(cos(tX)) + iE(sin(tX))$$

The main advantage of the characteristic function over transforms such as the Laplace transform, probability generating function or the moment generating function is property (a) below. Because we are integrating a bounded function; $|e^{itx}| = 1$ for all $x, t \in \Re$, the integral exists for any probability distribution.

6.1.2 Properties of Characteristic Function.

(a) φ exists for any distribution for X.

(b) $\varphi(0) = 1$.

(c) $|\varphi(t)| \leq 1$ for all t .

(d) φ is *uniformly continuous* . That is for all $\epsilon > 0$, there exists $\delta > 0$ such that $|\varphi(t) - \varphi(s)| \leq \epsilon$ whenever $|t - s| \leq \delta$.

(e) The characteristic function of $a + bX$ is $e^{iat}\varphi(bt)$.

(f) The characteristic function of $-X$ is the complex conjugate $\bar{\varphi}(t)$.

(g) A characteristic function φ is real valued if and only if the distribution of the corresponding random variable X has a distribution that is symmetric about zero, that is if and only if $P[X > z] = P[X < -z]$ for all $z \geq 0$.

(h) The characteristic function of a convolution $F * G$ is $\varphi_F(t)\varphi_G(t)$.

Proofs.

(a) Note that for each x and t, $|e^{itx}|^2 = sin^2(tx) + cos^2(tx) = 1$ and the constant 1 is integrable. Therefore

$$E|e^{itX}|^2 = 1.$$

It follows that

$$E|e^{itX}| \leq \sqrt{E|e^{itX}|^2} = 1$$

and so the function e^{itx} is integrable.

(b) $e^{itX} = 1$ when $t = 0$. Therefore $\varphi(0) = Ee^0 = 1$.

(c) This is included in the proof (a).

(d) Let $h = s - t$. Assume without loss of generality that $s > t$. Then

$$\begin{aligned} |\varphi(t) - \varphi(s)| &= |Ee^{itX}(e^{ihX} - 1)\ | \\ &\leq E[|e^{itX}(e^{ihX} - 1)|] \\ &\leq E[|e^{itX}||e^{ihX} - 1|] \\ &\leq E[|e^{ihX} - 1|]. \end{aligned}$$

But as $h \to 0$ the function $e^{ihX} - 1$ converges to 0 for each $\omega \in \Omega$ and it is dominated by the constant 2. Therefore, by the Lebesgue Dominated Convergence theorem, $E[|e^{ihX} - 1|] \to 0$ as $h \to 0$. So for a given $\epsilon > 0$, we chan choose h sufficiently small, for example $h = |s - t| \le \delta$ such that $|\varphi(t) - \varphi(s)| \le \epsilon$.

(e) By definition, $Ee^{it(a+bX)} = e^{ita}E[e^{itbX}] = e^{iat}\varphi(bt)$.

(f) Recall that the complex conjugate of $a + bi$ is $a - bi$ and of e^{iz} is e^{-iz} when $a, b,$ and z are real numbers. Then

$$E[e^{it(-X)}] = E[e^{-itX}] = E[cos(tX)+isin(-tX)] = E[cos(tX)-isin(tX)] = \bar{\varphi}(t).$$

(g) The distribution of the corresponding random variable X is symmetric if and only if X has the same distribution as does $-X$. This is true if and only if they have the same characteristic function. By properties (f) and the corollary below, this is true if and only if $\varphi(t) = \bar{\varphi}(t)$ which holds if and only if the function $\varphi(t)$ takes on only real values.

(h) Put $H = F * G$. Then

$$\int e^{itx}H(dx) = \int e^{itx} \int F(dx - y)G(dy)$$

$$= \int \int e^{it(z+y)}F(dz)G(dy), \text{with } z = x - y,$$

and this is $\varphi_F(t)\varphi_G(t)$.

The major reason for our interest in characteristic functions is that they uniquely describe the distribution. Probabilities of intervals can be recovered from the characteristic function using the following inversion theorem.

Theorem 107 *(*Inversion Formula*). If X has characteristic function $\varphi_X(t)$, then for any interval (a, b),*

$$P[a < X < b] + \frac{P[X = a] + P[X = b]}{2} = lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t)dt$$

Proof. Consider the integral

$$\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)dt = \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \int_{\Re} e^{itx}F(dx)dt$$

$$= \int_{-T}^{T} \int_{\Re} \frac{e^{it(x-a)} - e^{it(x-b)}}{2\pi it} F(dx)dt = \int_{\Re} \int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{2\pi it} dtF(dx).$$

Note that for real c we have

$$\int_{-T}^{T} \frac{e^{itc}}{2it} dt = \int_{0}^{T} \frac{sin(tc)}{t} dt$$

and so we obtain from above

$$\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt = \int_{\Re} \frac{1}{\pi} \{ \int_{0}^{T} \frac{sin(t(x-a))}{t} dt - \int_{0}^{T} \frac{sin(t(x-b))}{t} dt \} F(dx).$$

But as $T \to \infty$, it is possible to show that the integral (this is known as the sine integral function)

$$\frac{1}{\pi} \int_{0}^{T} \frac{sin(t(x-a))}{t} dt \to \begin{cases} -\frac{1}{2}, & x < a \\ \frac{1}{2}, & x > a \\ 0, & x = a \end{cases}$$

Substituting this above and taking limits through the integral using the Lebesgue Dominated Convergence Theorem, the limit is the integral with respect to $F(dx)$ of the function

$$g(x) = \begin{cases} \frac{1}{2}, & x = a \\ \frac{1}{2}, & x = b \\ 1, & a < x < b \\ 0, & elswhere \end{cases}$$

and this integral equals

$$P[a < X < b] + \frac{P[X = a] + P[X = b]}{2}.$$

■

Corollary 108 *If the characteristic function of two random variables X and Y agree, then X and Y have the same distribution.*

Proof. This follows immediately from the inversion formula above. ■

We have seen that if a sequence of cumulative distribution functions $F_n(x)$ converges pointwise to a limit, the limiting function $F(x)$ is not necessarily a cumulative distribution function. To ensure that it is, we require that the distributions are "tight". Similarly if a sequence of characteristic functions converge for each t, the limit is not necessarily the characteristic function of a probability distribution. However, in this case the tightness of the sequence translates to a very simple condition on the limiting characteristic function.

Theorem 109 *(Continuity Theorem) If X_n has characteristic function φ_n, then X_n converges weakly if and only if there exists a function φ which is continuous at 0 such that $\varphi_n(t) \to \varphi(t)$ for each t.* Note: In this case φ is the characteristic function of the limiting random variable X.

Proof. Suppose $X_n \Rightarrow X$. Then since the function e^{itx} is a continuous bounded function of x, then

$$E(e^{itX_n}) \to E(e^{itX}).$$

Conversely, suppose that $\varphi_n(t) \to \varphi(t)$ for each t and φ is a continuous function at $t = 0$. First prove that for all $\epsilon > 0$ there exists a $c < \infty$ such that $P[|X_n| > c] \leq \epsilon$ for all n. This is Problem 11 below. This shows that the sequence of random variables X_n is "tight" in the sense that any subsequence of it contains a further subsequence which converges in distribution to a proper cumulative distribution function. By the first half of the proof, $\varphi(t)$ is the characteristic function of the limit. Thus, since every subsequence has the same limit, $X_n \Rightarrow X$. ■

Example 110 *Suppose* $X_n \sim U[-n,n]$ *. Then the characteristic function of* X_n *is* $\varphi_n(t) = (sin\ tn)/tn$*. Does this converge as* $n \to \infty$*? Is the limit continuous at 0?*

Example 111 *Suppose* $X_1, \ldots X_n \ldots$ *are independent Cauchy distributed random variables with probability density function*

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \Re.$$

Then the sample mean $\bar{X}$ *has the same distribution as* X_1*.*

Note: We may use the integral formula

$$\int_0^\infty \frac{\cos(tx)}{b^2+x^2} dx = \frac{\pi}{2b} e^{-tb}, t \geq 0$$

to obtain the characteristic function of the above Cauchy distribution

$$\varphi(t) = e^{-|t|}.$$

6.1.3 Characteristic function of $N(\mu, \sigma^2)$.

The characteristic function of a random variable with the distribution $N(\mu, \sigma^2)$ is

$$\varphi(t) \ = \ exp\{i\mu t - \frac{\sigma^2 t^2}{2}\}.$$

(Note: Recall that for any real constant c,

$$\int_{-\infty}^{\infty} e^{-(x-c)^2/2} dx = \sqrt{2\pi}.$$

Use the fact that this remains true even if $c = it$).

6.2 The Central Limit Theorem

Our objective is to show that the sum of independent random variables, when standardized, converges in distribution to the standard normal distribution. The proof usually used in undergraduate statistics requires the moment generating function. However, the moment generating function exists only if moments of all orders exist, and so a more general result, requiring only that the random variables have finite mean and variance, needs to use characteristic functions. Two preliminary lemmas are used in the proof.

Lemma 112 *For real* x *,*

$$e^{ix} - (1 + ix - \frac{x^2}{2}) = r(x)$$

where $|r(x)| \le min[x^2, \frac{|x|^3}{6}]$. *Consequently,*

$$\varphi(t) \;=\; 1 + itE(X) - \frac{t^2}{2}E(X^2) + o(t^2)$$

where $\frac{o(t^2)}{t^2} \to 0$ *as* $t \to 0$.

Proof. By expanding e^{ix} in a Taylor series with remainder we obtain

$$\frac{e^{ix}-1}{i} = x + i\frac{x^2}{2} + i^2\frac{b_2}{2}$$

where $b_n(x) = \int_0^x (x-s)^n e^{is} ds$, and a crude approximation provides $|b_2| \le \int_0^x s^2 ds = x^3/3$. Integration by parts shows that $b_2 = \frac{2b_1 - x^2}{i}$ and substituting this provides the remaining bound on the error term. ∎

Lemma 113 *For any complex numbers* w_i, z_i, *if* $|z_i| \le 1$, $|w_i| \le 1$ *, then* $|\prod_i z_i - \prod_i w_i| \le \sum_i |z_i - w_i|$.

Proof. This is proved by induction using the fact that

$$\prod_{i=1}^n z_i - \prod_{i=1}^n w_i = (z_n - w_n)(\prod_{i=1}^{n-1} z_i) + w_n(\prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i) \le |z_n - w_n| + |(\prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i)|.$$

∎

This shows the often used result that

$$(1 - \frac{c}{n} + o(\frac{1}{n}))^n - (1 - \frac{c}{n})^n \to 0$$

and hence that

$$(1 - \frac{c}{n} + o(\frac{1}{n}))^n \to e^{-c} \quad \text{as } n \to \infty.$$

Theorem 114 *(Central Limit Theorem) If X_i are independent identically distributed random variables with $E(X_i) = \mu$, $var(X_i) = \sigma^2$, then*

$$S_n^* = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu)$$

converges weakly to $N(0,1)$.

Proof. Suppose we denote the characteristic function of $\frac{X_i-\mu}{\sigma}$ by $\varphi(t)$. By Lemma 112, $\varphi(t) = 1 - \frac{t^2}{2} + r(t)$ where $\frac{r(t)}{t^2} \to 0$ as $t \to 0$. Then the characteristic function of S_n^* is

$$\varphi^n(t/\sqrt{n}) = \{1 - \frac{t^2}{2n} + o(t^2/n)\}^n.$$

Note that by Lemma 113,

$$|\{1 - \frac{t^2}{2n} + o(t^2/n)\}^n - (1 - \frac{t^2}{2n})^n| \le n\ o(t^2/n) \to 0$$

and the second term $(1-\frac{t^2}{2n})^n \to e^{-t^2/2}$. Since this is the characteristic function of the standard normal distribution, it follows that S_n^* converges weakly to the standard normal distribution. ■

6.3 Problems

1. Find the characteristic function of the normal(0,1) distribution. Prove using characteristic functions that if F is the $N(\mu, \sigma^2)$ c.d.f. then $G(x) = F(\mu + \sigma x)$ is the $N(0, 1)$ c.d.f.

2. Let F be a distribution function and define

   $$G(x) = 1 - F(-x-)$$

   where $x-$ denotes the limit from the left. Prove that $F * G$ is symmetric.

3. Prove that $F * G = G * F$.

4. Prove using characteristic functions that if $F_n \Rightarrow F$ and $G_n \Rightarrow G$, then $F_n * G_n \Rightarrow F * G$.

5. Prove that convolution is *associative*. That

   $$(F * G) * H = F * (G * H).$$

6. Prove that if φ is a characteristic function, so is $|\varphi(t)|^2$.

7. Prove that any characteristic function is non-negative definite:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi(t_i-t_j)z_i\bar{z}_j \ \geq 0$$

for all real $t_1,...t_n$ and complex $z_1,...,z_n$.

8. Find the characteristic function of the Laplace distribution with density on $\Re$

$$f(x)=\frac{1}{2}e^{-|x|}. \tag{6.1}$$

What is the characteristic function of X_1+X_2 where X_i are independent with the probability density function (6.1)?

9. (Stable Laws) A family of distributions of importance in financial modelling is the *symmetric stable family.* These are unimodal densities, symmetric about their mode, and roughly similar in shape to the normal or Cauchy distribution (both special cases $\alpha=2$ or 1 respectively). They are of considerable importance in finance as an alternative to the normal distribution, because they tend to fit observations better in the tail of the distribution than does the normal. However, this is a more complicated family of densities to work with; neither the density function nor the cumulative distribution function can be expressed in a simple closed form. Both require a series expansion. They are most easily described by their *characteristic function*, which, upon setting location equal to 0 and scale equal to 1 is $Ee^{iXt} \ = e^{-|t|^{\alpha}}$. The parameter $0<\alpha\leq 2$ indicates what moments exist, for except in the special case $\alpha=2$ (the normal distribution), moments of order less than α exists while moments of order α or more do not. Of course, for the normal distribution, moments of all orders exist. The stable laws are useful for modelling in situations in which variates are thought to be approximately normalized sums of independent identically distributed random variables. To determine robustness against heavy-tailed departures from the normal distribution, tests and estimators can be computed with data simulated from a symmetric stable law with α near 2. The probability density function does not have a simple closed form except in the case $\alpha=1$ (the Cauchy distribution) and $\alpha=2$ (the Normal distribution) but can, at least theoretically, be determined from the series expansion of the probability density

$$f_c(x)=\sum_{k=0}^{\infty}(-1)^k\frac{\Gamma((k+1)/2)}{\pi c\alpha k!}\cos(\frac{k\pi}{c})(\frac{x}{c})^k.$$

Let $X_1,...,X_n$ be independent random variables all with a symmetric stable (α) distribution. Show that $n^{-1/\alpha}\sum_{i=1}^{n}X_i$ has the same Stable distribution. (If a stock price process follows a stable random walk with $\alpha<2$, large jumps in the process are more likely than in the case of normal returns. See for example the graph below of a stable random walk, $\alpha=1.7$).

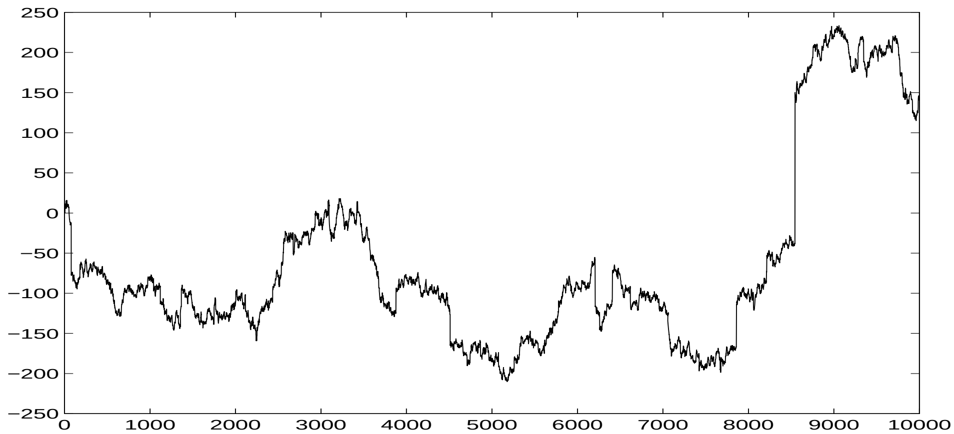

Figure 6.1: Stable Random Walk with $\alpha = 1.7$

10. Let Ω be the unit interval and P the uniform distribution and suppose we express each $\omega \in [0,1]$ in the binary expansion which does not terminate with finitely many terms. If $\omega = .\omega_1\omega_2....$, define $R_n(\omega) = 1$ if $\omega_n = 1$ and otherwise $R_n(\omega) = -1$. These are called the *Rademacher functions.* Prove that they are independent random variables with the same distribution.

11. For the Rademacher functions R_n defined on the unit interval, Borel sets and Lebesgue measure, let

    $Y_1 = R_1/2 + R_3/2^2 + R_6/2^3....$

    $Y_2 = R_2/2 + R_4/2^2 + R_7/2^3 + ...$

    $Y_3 = R_5/2 + R_8/2^2 + R_{12}/2^3 + ...$

    Prove that the Y_i are independent identically distributed and find their distribution.

12. Find the characteristic function of:

    (a) The Binomial distribution

    (b) The Poisson distribution

    (c) The geometric distribution

    Prove that suitably standardized, both the binomial distribution and the Poisson distribution approaches the standard normal distribution as one of the parameters $\rightarrow \infty$.

13. (*Families Closed under convolution.*) Show that each of the following families of distributions are closed under convolution. That is suppose X_1, X_2 are independent and have a distribution in the given family. Then show that the distribution of $X = X_1 + X_2$ is also in the family and identify the parameters.

(a) $Bin(n,p)$, with p fixed.

(b) Poisson (λ).

(c) Normal (μ, σ^2).

(d) Gamma (α, β), with β fixed.

(e) Chi-squared.

(f) Negative Binomial, with p fixed.

14. Suppose that a sequence of random variables X_n has characteristic functions $\varphi_n(t) \to \varphi(t)$ for each t and φ is a continuous function at $t = 0$. Prove that the distribution of X_n is tight, i.e. for all $\epsilon > 0$ there exists a $c < \infty$ such that $P[|X_n| > c] \le \epsilon$ for all n.

15. Prove, using the central limit theorem, that

$$\sum_{i=0}^{n} \frac{n^i e^{-n}}{i!} \to \frac{1}{2}, \quad \text{as } n \to \infty.$$

16. (*Negative binomial*) Suppose we decide in advance that we wish a fixed number (k) of successes in a sequence of Bernoulli trials, and sample repeatedly until we obtain exactly this number. Then the number of trials X is random and has probability function

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots.$$

Use the central limit theorem to show that this distribution can be approximated by a normal distribution when k is large. Verify the central limit theorem by showing that the characteristic function of the standardized Negative binomial approaches that of the Normal.

17. Consider a random walk built from independent Bernoulli random variables $X_i = 1$ with probability $p = \mu/\sqrt{n}$ and otherwise $X_i = 0$. Define the process

$$B_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} X_i$$

for all $0 \le t \le 1$. Find the limiting distribution of $B(t)$ and the limiting joint distribution of $B(s)$, $B(t) - B(s)$ for $0 < s < t < 1$.

Chapter 7

CONDITIONAL EXPECTATION AND MARTINGALES

7.1 Conditional Expectation.

Throughout this section we will assume that random variables X are defined on a probability space $(\Omega, \mathcal{F}, P)$ and have finite second moments so $E(X^2) < \infty$. This allows us to define conditional expectation through approximating one random variable by another, measurable with respect to a more coarse (i.e. made up of larger sets) or less informative sigma-algebra. We begin with the coursest sigma algebra of all, the trivial one $\{\Omega, \varphi\}$, with respect to which only constants are measurable.

What constant is the best fit to a random variable in the sense of smallest mean squared error? In other words, what is the value of c solving

$$\min_c E[(X-c)^2]?$$

Expanding,

$$E[(X-c)^2] = var(X) + (EX-c)^2$$

and so the minimum is achieved when we choose $c = EX$.

A constant is, of course, a random variable but a very basic one, measurable with respect to the trivial sigma-field $\{\Omega, \varphi\}$. Now suppose that we wished to approximate the value of a random variable X, not with a constant, but with another random variable Z, measurable with respect to some other sigma field $\mathcal{G} \subset \sigma(X)$. How course or fine the sigma algebra $\mathcal{G}$ is depends on how much information we have pertinent to the approximation of X. How good is our approximation will be measured using the mean squared error

$$E[(X-Z)^2]$$

and we wish to minimize this over all possible $\mathcal{G}-$random variables Z. The minimizing value of Z is the conditional expected value of X.

Theorem 115 *(conditional expectation as a projection)*

Let $\mathcal{G} \subset \mathcal{F}$ be sigma-algebras and X a random variable on $(\Omega, \mathcal{F}, P)$. Assume $E(X^2) < \infty$. Then there exists a $\mathcal{G}$-measurable Y such that

$$E[(X-Y)^2] = inf_Z E(X-Z)^2 \tag{7.1}$$

where the infimum is over all $\mathcal{G}$-measurable random variables.

Definition 116 *We denote the minimizing Y by $E(X|\mathcal{G})$.*

The next result assures us that the conditional expectation is unique, almost surely. In other words two random variables Y which solve the above minimization problem differ on a set of probability zero.

Theorem 117 *For two such minimizing Y_1, Y_2 , i.e. random variables Y which satisfy (7.1), we have $P[Y_1 = Y_2] = 1$. This implies that conditional expectation is almost surely unique.*

Proof. Suppose both Y_1 and Y_2 are $\mathcal{G}$-measurable and both minimize $E[(X-Y)^2]$. Then for any $A \in \mathcal{G}$ it follows from property (d) below that

$$\int_A Y_1 dP = \int_A Y_2 dP$$

or

$$\int_A (Y_1 - Y_2) dP = 0.$$

Choose $A = [Y_1 - Y_2 \geq 0]$ and note that

$$\int (Y_1 - Y_2) I_A dP = 0$$

and the integrand $(Y_1 - Y_2) I_A$ is non-negative together imply that $(Y_1 - Y_2) I_A = 0$ almost surely. Similarly on the set $A = [Y_1 - Y_2 < 0]$ we can show that $(Y_1 - Y_2) I_A = 0$ almost surely. It follows that $Y_1 = Y_2$ almost surely. ■

Example 118 *Suppose $\mathcal{G} = \{\varphi, \Omega\}$. What is $E(X|\mathcal{G})$?*

The only random variables which are measurable with respect to the trivial sigma-field are constants. So this leads to the same minimization discussed above, $\min_c E[(X-c)^2] = \min_c \{var(X) + (EX-c)^2\}$ which results in $c = E(X)$.

Example 119 *Suppose $\mathcal{G} = \{\varphi, A, A^c, \Omega\}$ for some event A. What is $E(X|\mathcal{G})$? Consider the special case: $X = I_B$.*

In this case suppose the random variable Z takes the value a on A and b on the set A^c. Then

$$\begin{aligned} E[(X-Z)^2] &= E[(X-a)^2 I_A] + E[(X-b)^2 I_{A^c}] \\ &= E(X^2 I_A) - 2aE(XI_A) + a^2 P(A) \\ &+ E(X^2 I_{A^c}) - 2bE(XI_{A^c}) + b^2 P(A^c). \end{aligned}$$

Minimizing this with respect to both a and b results in

$$\begin{aligned} a &= E(XI_A)/P(A) \\ b &= E(XI_{A^c})/P(A^c). \end{aligned}$$

These values a and b are usually referred to in elementary probability as $E(X|A)$ and $E(X|A^c)$ respectively. Thus, the conditional expectated value can be written

$$E(X|\mathcal{G})(\omega) = \begin{cases} E(X|A) & \text{if } \omega \in A \\ E(X|A^c) & \text{if } \omega \in A^c \end{cases}$$

As a special case consider X to be an indicator random variable $X = I_B$. Then we usually denote $E(I_B|\mathcal{G})$ by $P(B|\mathcal{G})$ and

$$P(B|\mathcal{G})(\omega) = \begin{cases} P(B|A) & \text{if } \omega \in A \\ P(B|A^c) & \text{if } \omega \in A^c \end{cases}$$

Note: *Expected value is a constant, but the conditional expected value $E(X|G)$ is a random variable measurable with respect to G. Its value on the atoms of $\mathcal{G}$ is the average of the random variable X over these atoms.*

Example 120 *Suppose $\mathcal{G}$ is generated by a finite partition $\{A_1, A_2, ..., A_n\}$ of the probability space Ω.. What is $E(X|\mathcal{G})$?*

In this case, any $\mathcal{G}$-measurable random variable is constant on the sets in the partition $A_j, j = 1, 2, ..., n$ and an argument similar to the one above shows that the conditional expectation is the simple random variable:

$$E(X|\mathcal{G})(\omega) = \sum_{i=1}^{n} c_i I_{A_i}(\omega)$$

$$\text{where } c_i = E(X|A_i) = \frac{E(XI_{A_i})}{P(A_i)}$$

Example 121 *Consider the probability space $\Omega = (0, 1]$ together with $P =$ Lebesgue measure and the Borel Sigma Algebra. Suppose the function $X(\omega)$ is Borel measurable. Assume that $\mathcal{G}$ is generated by the intervals $(\frac{j-1}{n}, \frac{j}{n}]$ for $j = 1, 2, ..., n$. What is $E(X|\mathcal{G})$?*

In this case

$$\begin{aligned} E(X|\mathcal{G})(\omega) &= n \int_{(j-1)/n}^{j/n} X(s)ds \quad \text{when } \omega \in (\frac{j-1}{n}, \frac{j}{n}] \\ &= \text{average of } X \text{ values over the relevant interval.} \end{aligned}$$

For the purpose of the following properties, we say that a random variable X and a sigma-field $\mathcal{G}$ are independent if for any Borel set B the event $[X \in B]$ is independent of all events in $\mathcal{G}$. We also remind the reader that all random variables appearing in this section are assumed to have finite variance although most of these properties can be extended to more general integrable random variables.

7.1.1 Properties of Conditional Expectation.

(a) If a random variable X is $\mathcal{G}$-measurable, $E(X|\mathcal{G}) = X$.

(b) If a random variable X and a sigma-field $\mathcal{G}$ are independent, then $E(X|\mathcal{G}) = E(X)$.

(c) For any square integrable $\mathcal{G}$-measurable Z, $E(ZX) = E[ZE(X|\mathcal{G})]$.

(d) (special case of (c)): $\int_A X dP = \int_A E(X|\mathcal{G}] dP$ for all $A \in \mathcal{G}$.

(e) $E(X) = E[E(X|\mathcal{G})]$.

(f) If a $\mathcal{G}$-measurable random variable Z satisfies $E[(X - Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y, then $Z = E(X|\mathcal{G})$.

(g) If Y_1, Y_2 are distinct $\mathcal{G}-$measurable random variables both minimizing $E(X - Y)^2$, then $P(Y_1 = Y_2) = 1$.

(h) *Additive* $E(X + Y|\mathcal{G}) = E(X|\mathcal{G}) + E(Y|\mathcal{G})$.
Linearity $E(cX + d|\mathcal{G}) = cE(X|\mathcal{G}) + d$.

(i) If Z is $\mathcal{G}-$measurable, $E(ZX|\mathcal{G}) = ZE(X|\mathcal{G})$ a.s.

(j) If $\mathcal{H} \subset \mathcal{G}$ are sigma-algebras, $E[E(X|\mathcal{G})|\mathcal{H}] = E(X|\mathcal{H})$.

(k) If $X \leq Y$, $E(X|\mathcal{G}) \leq E(Y|\mathcal{G})$ a.s.

(l) *Conditional Lebesgue Dominated Convergence.* If $X_n \to X$ a.s. and $|X_n| \leq Y$ for some integrable random variable Y , then $E(X_n|\mathcal{G}) \to E(X|\mathcal{G})$ in distribution

Proof. (Proof of the above properties) ∎

(a) Notice that for any random variable Z that is $\mathcal{G}$-measurable, $E(X-Z)^2 \geq E(X - X)^2 = 0$ and so the minimizing Z is X (by definition this is $E(X|\mathcal{G})$).

(b) Consider a random variable Y measurable with respect $\mathcal{G}$ and therefore independent of X. Then

$$\begin{aligned} E(X-Y)^2 &= E[(X - EX + EX - Y)^2] \\ &= E[(X - EX)^2] + 2E[(X - EX)(EX - Y)] + E[(EX - Y)^2] \\ &= E[(X - EX)^2] + E[(EX - Y)^2] \text{ by independence} \\ &\geq E[(X - EX)^2]. \end{aligned}$$

It follows that $E(X-Y)^2$ is minimized when we choose $Y = EX$ and so $E(X|\mathcal{G}) = E(X)$.

(c) for any $\mathcal{G}-$measurable square integrable random variable Z, we may define a quadratic function of λ by

$$g(\lambda) = E[(X - E(X|\mathcal{G}) - \lambda Z)^2]$$

By the definition of $E(X|\mathcal{G})$, this function is minimized over all real values of λ at the point $\lambda = 0$ and therefore $g'(0) = 0$. Setting its derivative $g'(0) = 0$ results in the equation

$$E(Z(X - E(X|\mathcal{G}))) = 0$$

or $E(ZX) = E[ZE(X|\mathcal{G})]$.

(d) If in (c) we put $Z = I_A$ where $A \in \mathcal{G}$, we obtain $\int_A XdP = \int_A E(X|\mathcal{G}]dP$.

(e) Again this is a special case of property (c) corresponding to $Z = 1$.

(f) Suppose a $\mathcal{G}$-measurable random variable Z satisfies $E[(X-Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y. Consider in particular $Y = E(X|\mathcal{G}) - Z$ and define

$$\begin{aligned} g(\lambda) &= E[(X - Z - \lambda Y)^2] \\ &= E((X-Z)^2 - 2\lambda E[(X-Z)Y] + \lambda^2 E(Y^2) \\ &= E(X-Z)^2 + \lambda^2 E(Y^2) \\ &\geq E(X-Z)^2 = g(0). \end{aligned}$$

In particular $g(1) = E[(X - E(X|\mathcal{G}))^2] \geq g(0) = E(X-Z)^2$ and by Theorem 117, $Z = E(X|\mathcal{G})$ almost surely.

(g) This is just deja vu (Theorem 117) all over again.

(h) Consider, for an arbitrary $\mathcal{G}-$measurable random variable Z,

$$\begin{aligned} E[Z(X + Y - E(X|\mathcal{G}) - E(Y|\mathcal{G}))] &= E[Z(X - E(X|\mathcal{G}))] + E[Z(Y - E(Y|\mathcal{G}))] \\ &= 0 \text{ by property (c).} \end{aligned}$$

It therefore follows from property (f) that $E(X+Y|\mathcal{G}) = E(X|\mathcal{G}) + E(Y|\mathcal{G})$.

By a similar argument we may prove $E(cX + d|\mathcal{G}) = cE(X|\mathcal{G}) + d$.

(i) This is Problem 2.

(j) This is Problem 4 (sometimes called the tower property of conditional expectation: If $\mathcal{H} \subset \mathcal{G}$ are sigma-algebras, $E[E(X|\mathcal{G})|\mathcal{H}] = E(X|\mathcal{H})$).

(k) If $X \leq Y$, $E(X|\mathcal{G}) \leq E(Y|\mathcal{G})$ a.s.

(l) *Conditional Lebesgue Dominated Convergence.* If $X_n \to X$ a.s. and $|X_n| \leq Y$ for some integrable random variable Y, then $E(X_n|\mathcal{G}) \to E(X|\mathcal{G})$ *in probability.*

Notes. In general, we define $E(X|Z) = E(X|\sigma(Z))$ and conditional variance $var(X|\mathcal{G}) = E\{(X - E(X|\mathcal{G}))^2|\mathcal{G}\}$. For results connected with property (l) above providing conditions under which the conditional expectations converge, see Convergence in distribution of conditional expectations, (1994) E.M. Goggin, *Ann. Prob* 22, 2. 1097-1114.

7.2 Conditional Expectation for integrable random variables.

We have defined conditional expectation as a projection only for random variables with finite variance. It is fairly easy to extend this definition to random variables X on a probability space $(\Omega, \mathcal{F}, P)$ for which $E(|X|) < \infty$. We wish to define $E(X|\mathcal{G})$ where the sigma algebra $\mathcal{G} \subset \mathcal{F}$. First, for non-negative integrable X choose simple random variables $X_n \uparrow X$. Since simple random variables have only finitely many values, they have finite variance, and we can use the definition above for their conditional expectation. Then $E(X_n|\mathcal{G}) \uparrow$ and so it converges. Define $E(X|\mathcal{G})$ to be the limit. In general, for random variables taking positive and negative values, we define $E(X|\mathcal{G}) = E(X^+|\mathcal{G}) - E(X^-|\mathcal{G})$. There are a number of details that need to be ironed out. First we need to show that this new definition is consistent with the old one when the random variable happens to be square integrable. We can also show that the properties (a)-(i) above all hold under this new definition of conditional expectation. We close with the more common definition of conditional expectation found in most probability and measure theory texts, essentially property (d) above. It is, of course, equivalent to the definition as a projection in section 7.1 and the definition above as a limit of the conditional expectation of simple functions.

Theorem 122 *Consider a random variable X defined on a probability space $(\Omega, \mathcal{F}, P)$ for which $E(|X|) < \infty$. Suppose the sigma algebra $\mathcal{G} \subset \mathcal{F}$. Then there is a unique (almost surely P) $\mathcal{G}-$measurable random variable Z satisfying*

$$\int_A X dP = \int_A Z dP \quad \text{for all } A \in \mathcal{G}$$

Any such Z we call the conditional expectation and denote by $E(X|\mathcal{G})$.

7.3 Martingales in Discrete Time

In this section all random variables are defined on the same probability space $(\Omega, \mathcal{F}, P)$. Partial information about these random variables may be obtained from the observations so far, and in general, the "history" of a process up to time

t is expressed through a sigma-algebra $H_t \subset \mathcal{F}$. We are interested in stochastic processes or sequences of random variables called martingales, intuitively, the total fortune of an individual participating in a "fair game". In order for the game to be "fair", the expected value of your future fortune given the history of the process up to and including the present should be equal to your present wealth. In a sense you are neither tending to increase or decrease your wealth over time- any fluctuations are purely random. Suppose your fortune at time s is denoted X_s. The values of the process of interest and any other related processes up to time s generate a sigma-algebra H_s. Then the assertion that the game is fair implies that the expected value of our future fortune given this hisstory of the process up to the present is exactly our present wealth $E(X_t|H_s) = X_s$ for $t > s$.

Definition 123 $\{(X_t, H_t); t \in T\}$ *is a* martingale *if*

(a) H_t is increasing (in t) family of sigma-algebras

(b) Each X_t is H_t- measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $s, t \in T$, $E(X_t|H_s) = X_s$ a.s.

Example 124 *Suppose Z_t are independent random variables with expectation 0. Define $H_t = \sigma(Z_1, Z_2, . . . Z_t)$ and $S_t = \sum_{i=1}^t Z_i$. Then $\{(S_t, H_t), t = 1, 2, ...\}$ is a martingale. Suppose that $E(Z_t^2) = \sigma^2 < \infty$. Then $\{(S_t^2 - t\sigma^2, H_t), t = 1, 2, ...\}$ is a martingale.*

Example 125 *Suppose Z_t are independent random variables with $Z_t \geq 0$. Define $H_t = \sigma(Z_1, Z_2, . . . Z_t)$ and $M_t = \prod_{i=1}^t Z_i$. Suppose that $E(Z_i^\lambda) = \phi(\lambda) < \infty$. Then*

$$\{(\frac{M_t^\lambda}{\phi^t(\lambda)}, H_t), t = 1, 2, ...\}$$

is a martingale.

This is an example of a parametric family of martingales indexed by λ obtained by multiplying independent random variables.

Example 126 *Let X be any integrable random variable, and H_t an increasing family of sigma-algebras. Put $X_t = E(X|H_t)$. Then (X_t, H_t) is a martingale.*

Definition 127 *Let $\{(M_n, H_n); n = 1, 2, ...\}$ be a martingale and A_n be a sequence of random variables measurable with respect to $H_{n-1.}$ Then the sequence A_n is called **non-anticipating**. (an alternate term is **predictable**)*

In gambling, we must determine our stakes and our strategy on the $n'th$ play of a game based on the information available to use at time $n-1$. Similarly, in investment, we must determine the weights on various components in our portfolio at the end of day (or hour or minute) $n-1$ *before* the random marketplace determines our profit or loss for that period of time. In this sense gambling and investment strategies must be determined by non-anticipating sequences of random variables (although both gamblers and investors often dream otherwise).

Definition 128 *(Martingale Transform). Let* $\{(M_t, H_t), t = 0, 1, 2, ...\}$ *be a martingale and let* A_n *be a bounded non-anticipating sequence with respect to* H_n. *Then the sequence*

$$\tilde{M}_t = A_1(M_1 - M_0) + ... + A_t(M_t - M_{t-1}) \tag{7.2}$$

is called a Martingle transform of M_t.

The martingale transform is sometimes denoted $A \circ M$.

Theorem 129 $\{(\tilde{M}_t, H_t), t = 1, 2, ...\}$ *is a martingale.*

Proof.

$$\begin{aligned} E[\widetilde{M}_j - \widetilde{M}_{j-1}|H_{j-1}] &= E[A_j(M_j - M_{j-1})|H_{j-1}] \\ &= A_j E[(M_j - M_{j-1})|H_{j-1}] \text{ since } A_j \text{ is } H_{j-1} \text{ measurable} \\ &= 0 \text{ a.s.} \end{aligned}$$

Therefore

$$E[\widetilde{M}_j|H_{j-1}] = \widetilde{M}_{j-1} \quad \text{a.s.}$$

■

Consider a random variable τ that determines when we stop betting or investing. Its value can depend arbitrarily on the outcomes in the past, as long as the decision to stop at time $\tau = n$ depends only on the results at time $n, n-1,$ ...etc. Such a random variable is called an optional stopping time.

Definition 130 *A random variable* τ *taking values in* $\{0, 1, 2, ...\} \cup \{\infty\}$ *is a (optional)* stopping time *for a martingale* (X_t, H_t) *if for each* n , $[\tau \leq n] \in H_n$.

If we stop a martingle at some random stopping time, the result continues to be a martingale as the following theorem shows.

Theorem 131 *Suppose that* $\{(M_t, H_t), t = 0, 1, 2, ...\}$ *is a martingale and* τ *is an optional stopping time with values on* $\{0, 1, 2, ...\}$. *Define* $Y_n = M_{n\wedge\tau} = M_{\min(n,\tau)}$. *Then* $\{(Y_n, H_n), n = 0, 1, 2, ..\}$ *is a martingale.*

Proof. Notice that

$$M_{n\wedge\tau} = M_0 + \sum_{j=1}^{n}(M_j - M_{j-1})I(\tau \geq j).$$

Letting $A_j = I(\tau \geq j)$ this is a bounded $H_{j-1}-$measurable sequence and therefore $\sum_{j=1}^{n}(M_j - M_{j-1})I(\tau \geq j)$ is a martingale transform. By Theorem 128 it is a martingale. ■

Example 132 *(Ruin probabilities). Consider a random walk* $S_n = \sum_{i=1}^n X_i$ *where the random variables* X_i *are independent identically distributed with* $P(X_i = 1) = p$, $P(X_i = -1) = q$, $P(X_i = 0) = 1-p-q$ *for* $0 < p+q \le 1, p \neq q$. *Then* $M_n = (q/p)^{S_n}$ *is a martingale. Suppose that* $A < S_0 < B$ *and define the optional stopping time* τ *as the first time we hit either of two barriers at* A *or* B. *Then* $M_{n\wedge\tau}$ *is a martingale. Since* $E(M_\tau) = \lim_{n\to\infty} E(M_{n\wedge\tau}) = (q/p)^{S_0}$ *by dominated covergence, we have*

$$(q/p)^A p_A + (q/p)^B p_B = (q/p)^{S_0} \tag{7.3}$$

where p_A *and* $p_B = 1 - p_A$ *are the probabilities of hitting absorbing barriers at* A *or* B *respectively. Solving, it follows that*

$$((q/p)^A - (q/p)^B)p_A = (q/p)^{S_0} - (q/p)^B \tag{7.4}$$

or that

$$p_A = \frac{(q/p)^{S_0} - (q/p)^B}{(q/p)^A - (q/p)^B}. \tag{7.5}$$

In the case $p = q$, *a similar argument provides*

$$p_A = \frac{B - S_0}{B - A}. \tag{7.6}$$

Definition 133 *For an optional stopping time* τ *define*

$$H_\tau \;=\; \{A \in H; A \cap [\tau \le n] \in H_n, \quad for\ all\ n\}. \tag{7.7}$$

Theorem 134 H_τ *is a sigma-algebra.*

Proof. Clearly since the empty set $\varphi \in H_n$ for all n, so is $\varphi \cap [\tau \le n]$ and so $\varphi \in H_\tau$. We also need to show that if $A \in H_\tau$ then so is the complement A^c. Notice that for each n,

$$\begin{aligned}&[\tau \le n] \cap \{A \cap [\tau \le n]\}^c \\ &\quad = [\tau \le n] \cap \{A^c \cup [\tau > n]\} \\ &\quad = A^c \cap [\tau \le n]\end{aligned}$$

and since each of the sets $[\tau \le n]$ and $A \cap [\tau \le n]$ are H_n-measurable, so must be the set $A^c \cap [\tau \le n]$. Since this holds for all n it follows that whenever $A \in H_\tau$ then so A^c. Finally, consider a sequence of sets $A_m \in H_\tau$ for all $m = 1, 2, ...$ We need to show that the countable union $\cup_{m=1}^\infty A_m \in H_\tau$. But

$$\{\cup_{m=1}^\infty A_m\} \cap [\tau \le n] = \cup_{m=1}^\infty \{A_m \cap [\tau \le n]\}$$

and by assumption the sets $\{A_m \cap [\tau \le n]\} \in H_n$ for each n. Therefore

$$\cup_{m=1}^\infty \{A_m \cap [\tau \le n]\} \in H_n$$

and since this holds for all n, $\cup_{m=1}^\infty A_m \in H_\tau$. ■

Definition 135 $\{(X_t, H_t); t \in T\}$ *is a* submartingale *if*

(a) H_t is increasing (in t) family of sigma-algebras.

(b) Each X_t is H_t measurable and $E|X_t| < \infty$.

(c) For each $s < t,$, $E(X_t|H_s) \geq X_s$ a.s.

Note that every martingale is a submartingale. There is a version of Jensen's inequality for conditional expectation as well as the one proved before for ordinary expected value.

Theorem 136 *(Jensen's Inequality) Let ϕ be a convex function. Then for any random variable X and sigma-field H,*

$$\phi(E(X|H)) \leq E(\phi(X)|H). \tag{7.8}$$

Proof. Consider the set $\mathcal{L}$ of linear function $L(x) = a + bx$ that lie entirely below the graph of the function $\phi(x)$. It is easy to see that for a convex function

$$\phi(x) = \sup\{L(x); L \in \mathcal{L}\}.$$

For any such line,

$$\begin{aligned} E(\phi(X)|H) &\geq E(L(X)|H) \\ &\geq L(E(X)|H)). \end{aligned}$$

If we take the supremum over all $L \in \mathcal{L}$, we obtain

$$E(\phi(X)|H) \geq \phi(E(X)|H)).$$

■

Example 137 *Let X be any random variable and H be a sigma-field. Then for $1 \leq p \leq k < \infty$*

$$\{E(|X|^p|H)\}^{1/p} \leq \{E(|X|^k|H)\}^{1/k}. \tag{7.9}$$

In the special case that H is the trivial sigma-field, this is the inequality

$$||X||_p \leq ||X||_k. \tag{7.10}$$

Proof. Consider the function $\phi(x) = |x|^{k/p}$. This function is convex provided that $k \geq p$ and by the conditional form of Jensen's inequality,

$$E(|X|^k|H) = E(\phi(|X|^p)|H) \geq \phi(E(|X|^p|H)) = |E(|X|^p|H)|^{k/p} \text{ a.s.}$$

■

Example 138 *(Constructing Submartingales). Let S_n be a martingale with respect to H_n. Then $(|S_n|^p, H_n)$ is a submartingale for any $p \geq 1$ provided that $E|S_n|^p < \infty$.*

Proof. Since the function $\phi(x) = |x|^p$ is convex for $p \geq 1$, it follows from the conditional form of Jensen's inequaltiy that

$$E(|S_{n+1}|^p|H_n) = E(\phi(S_{n+1})|H_n) \geq \phi(E(S_{n+1}|H_n)) = \phi(S_n) = |S_n|^p \text{ a.s.}$$

■

Theorem 139 *Let X_n be a submartingale and suppose ϕ is a convex nondecreasing function with $E\phi(X_n) < \infty$. Then $\phi(X_n)$ is a submartingale.*

Proof. Since the function $\phi(x)$ is convex ,

$$E(\phi(S_{n+1})|H_n) \geq \phi(E(S_{n+1}|H_n)) \geq \phi(S_n) \text{ a.s.}$$

since $E(S_{n+1}|H_n) \geq S_n$ a.s. and the function ϕ is non-decreasing. ■

Corollary 140 *Let (X_n, H_n) be a submartingale. Then $((X_n - a)^+, H_n)$ is a submartingale.*

Proof. The function $\phi(x) = (x - a)^+$ is convex and non-decreasing. ■

Theorem 141 *(Doob's Maximal Inequality) Suppose (M_n, H_n) is a non-negative submartingale. Then for $\lambda > 0$ and $p \geq 1$,*

$$P(\sup_{0 \leq m \leq n} M_m \geq \lambda) \leq \lambda^{-p} E(M_n^p)$$

Proof. We prove this in the case $p = 1$. The general case we leave as a problem. Define a stopping time

$$\tau = \min\{m; M_m \geq \lambda\}$$

so that $\tau \leq n$ if and only if the maximum has reached the value λ by time n or

$$P[\sup_{0 \leq m \leq n} M_m \geq \lambda] = P[\tau \leq n].$$

Now on the set $[\tau \leq n]$, the maximum $M_\tau \geq \lambda$ so

$$\lambda I(\tau \leq n) \leq M_\tau I(\tau \leq n) = \sum_{i=1}^{n} M_i I(\tau = i). \tag{7.11}$$

By the submartingale property, for any $i \leq n$ and $A \in H_i$,

$$E(M_i I_A) \leq E(M_n I_A).$$

Therefore, taking expectations on both sides of (7.11), and noting that for all $i \leq n$,

$$E(M_i I(\tau = i)) \leq E(M_n I(\tau = i))$$

we obtain

$$\lambda P(\tau \leq n) \leq E(M_n I(\tau \leq n)) \leq E(M_n).$$

■

Theorem 142 *(Doob's L^p Inequality) Suppose (M_n, H_n) is a non-negative submartingale and put $M_n^* = \sup_{0 \leq m \leq n} M_n$. Then for $p > 1$, and all n*

$$||M_n^*||_p \leq \frac{p}{p-1} ||M_n||_p$$

One of the main theoretical properties of martingales is that they converge under fairly general conditions. Conditions are clearly necessary. For example consider a simple random walk $S_n = \sum_{i=1}^n Z_i$ where $Z_i$are independent identically distributed with $P(Z_i = 1) = P(Z_i = -1) = \frac{1}{2}$. Starting with an arbitrary value of S_0, say $S_0 = 0$ this is a martingale, but as $n \to \infty$ it does not converge almost surely or in probability.

On the other hand, consider a Markov chain with the property that $P(X_{n+1} = j|X_n = i) = \frac{1}{2i+1}$ for $j = 0, 1, ..., 2i$. Notice that this is a martingale and beginning with a positive value, say $X_0 = 10$, it is a non-negative martingale. Does it converge almost surely? If so the only possible limit is $X = 0$ because the nature of the process is such that $P[|X_{n+1} - X_n| \geq 1|X_n = i] \geq \frac{2}{3}$ unless $i = 0$. The fact that it does converge a.s. is a consequence of the martingale convergence theorem. Does it converge in L_1 i.e. in the sense that $E[|X_n - X|] \to 0$ as $n \to \infty$? If so, then clearly $E(X_n) \to E(X) = 0$ and this contradicts the martingale property of the sequence which implies $E(X_n) = E(X_0) = 10$. This is an example of a martingale that converges almost surely but not in L_1.

Lemma 143 *If $(X_t, H_t), t = 1, 2, ..., n$ is a (sub)martingale and if α, β are optional stopping times with values in $\{1, 2, ..., n\}$ such that $\alpha \leq \beta$ then*

$$E(X_\beta | H_\alpha) \geq X_\alpha$$

with equality if X_t is a martingale.

Proof. It is sufficient to show that

$$\int_A (X_\beta - X_\alpha) dP \geq 0$$

for all $A \in H_\alpha$. Note that if we define $Z_i = X_i - X_{i-1}$ to be the submartingale differences, the submartingale condition implies

$$E(Z_j | H_i) \geq 0 \text{ a.s. whenever } i < j.$$

Therefore for each $j = 1, 2, ...n$ and $A \in H_\alpha$,

$$\begin{aligned}\int_{A\cap[\alpha=j]}(X_\beta - X_\alpha)dP &= \int_{A\cap[\alpha=j]}\sum_{i=1}^{n} Z_i I(\alpha < i \le \beta)dP \\ &= \int_{A\cap[\alpha=j]}\sum_{i=j+1}^{n} Z_i I(\alpha < i \le \beta)dP \\ &= \int_{A\cap[\alpha=j]}\sum_{i=j+1}^{n} E(Z_i|H_{i-1})I(\alpha < i)I(i \le \beta)dP \\ &\ge 0 \text{ a.s.}\end{aligned}$$

since $I(\alpha < i)$, $I(i \le \beta)$ and $A \cap [\alpha = j]$ are all measurable with respect to H_{i-1} and $E(Z_i|H_{i-1}) \ge 0$ a.s. If we add over all $j = 1, 2, ..., n$ we obtain the desired result. ■

The following inequality is needed to prove a version of the submartingale convergence theorem.

Theorem 144 *(Doob's upcrossing inequality) Let M_n be a submartingale and for $a < b$, define $N_n(a, b)$ to be the number of complete upcrossings of the interval (a, b) in the sequence $M_j, j = 0, 1, 2, ..., n$. This is the largest k such that there are integers $i_1 < j_1 < i_2 < j_2... < j_k \le n$ for which*

$$M_{i_l} \le a \quad \text{and } M_{j_l} \ge b \quad \text{for all } l = 1, ..., k.$$

Then

$$(b - a)EN_n(a, b) \le E\{(M_n - a)^+ - (M_0 - a)^+\}$$

Proof. By Corollary 140, we may replace M_n by $X_n = (M_n - a)^+$ and this is still a submartingale. Then we wish to count the number of upcrossings of the interval $[0, b']$ where $b' = b - a$. Define stopping times for this process by $\alpha_0 = 0$,

$$\begin{aligned}\alpha_1 &= \min\{j; 0 \le j \le n, X_j = 0\} \\ \alpha_2 &= \min\{j; \alpha_1 \le j \le n, X_j \ge b'\} \\ &... \\ \alpha_{2k-1} &= \min\{j; \alpha_{2k-2} \le j \le n, X_j = 0\} \\ \alpha_{2k} &= \min\{j; \alpha_{2k-1} \le j \le n, X_j \ge b'\}.\end{aligned}$$

In any case, if α_k is undefined because we do not again cross the given boundary, we define $\alpha_k = n$. Now each of these random variables is an optional stopping time. If there is an upcrossing between X_{α_j} and $X_{\alpha_{j+1}}$ (where j is odd) then the distance travelled

$$X_{\alpha_{j+1}} - X_{\alpha_j} \ge b'.$$

If X_{α_j} is well-defined (i.e. it is equal to 0) and there is no further upcrossing, then $X_{\alpha_{j+1}} = X_n$ and

$$X_{\alpha_{j+1}} - X_{\alpha_j} = X_n - 0 \ge 0.$$

Similarly if j is even, since by the above lemma, $(X_{\alpha_j}, H_{\alpha_j})$ is a submartingale,

$$E(X_{\alpha_{j+1}} - X_{\alpha_j}) \geq 0.$$

Adding over all values of j, and using the fact that $\alpha_0 = 0$ and $\alpha_n = n$,

$$E\sum_{j=0}^{n}(X_{\alpha_{j+1}} - X_{\alpha_j}) \geq b' EN_n(a,b)$$

$$E(X_n - X_0) \geq b' EN_n(a,b).$$

In terms of the original submartingale, this gives

$$(b-a)EN_n(a,b) \leq E(M_n - a)^+ - E(M_0 - a)^+.$$

■

Doob's martingale convergence theorem that follows is one of of the nicest results in probability and one of the reasons why martingales are so frequently used in finance, econometrics, clinical trials and lifetesting.

Theorem 145 *(Sub)martingale Convergence Theorem. Let* (M_n, H_n); $n = 1, 2, \ldots$ *be a submartingale such that* $sup_{n\to\infty} EM_n^+ < \infty$. *Then there is an integrable random variable* M *such that* $M_n \to M$ *a.s.*

Proof. The proof is an application of the upcrossing inequality. Consider any interval $a < b$ with rational endpoints. By the upcrossing inequality,

$$E(N_a(a,b)) \leq \frac{1}{b-a} E(M_n - a)^+ \leq \frac{1}{b-a}[|a| + E(M_n^+)]. \tag{7.12}$$

Let $N(a,b)$ be the total number of times that the martingale completes an upcrossing of the interval $[a,b]$ over the infinite time interval $[1,\infty)$ and note that $N_n(a,b) \uparrow N(a,b)$ as $n \to \infty$. Therefore by monotone convergence $E(N_a(a,b)) \to EN(a,b)$ and by (7.12)

$$E(N(a,b)) \leq \frac{1}{b-a} \limsup[a + E(M_n^+)] < \infty.$$

This imples

$$P[N(a,b) < \infty] = 1.$$

Therefore,

$$P(\liminf M_n \leq a < b \leq \limsup M_n) = 0$$

for every rational $a < b$ and this implies that M_n converges almost surely to a (possibly infinite) random variable. Call this limit $M.We$ need to show that this random variable is almost surely finite. Because $E(M_n)$ is non-decreasing,

$$E(M_n^+) - E(M_n^-) \geq E(M_0)$$

and so

$$E(M_n^-) \le E(M_n^+) - E(M_0).$$

But by Fatou's lemma

$$E(M^+) = E(\liminf M_n^+) \le \liminf EM_n^+ < \infty$$

Therefore $E(M^-) < \infty$ and consequently the random variable M is finite almost surely. ■

Theorem 146 *(L^p martingale Convergence Theorem) Let (M_n, H_n); $n = 1, 2, \ldots$ be a martingale such that $sup_{n\to\infty} E|M_n|^p < \infty, p > 1$. Then there is an random variable M such that $M_n \to M$ a.s. and in L^p.*

Example 147 *(The Galton-Watson process). Consider a population of Z_n individuals in generation n each of which produces a random number ξ of offspring in the next generation so that the distribution of Z_{n+1} is that of $\xi_1 + ... + \xi_{Z_n}$ for independent identically distributed ξ. This process $Z_n, n = 1, 2, ...$ is called the Galton-Watson process. Let $E(\xi) = \mu$. Assume we start with a single individual in the population $Z_0 = 1$ (otherwise if there are j individuals in the population to start then the population at time n is the sum of j independent terms, the offspring of each). Then*

- The sequence Z_n/μ^n is a martingale.
- If $\mu < 1$, $Z_n \to 0$ and $Z_n = 0$ for all sufficiently large n.
- If $\mu = 1$ and $P(\xi \neq 1) > 0$, then $Z_n = 0$ for all sufficiently large n.
- If $\mu > 1$, then $P(Z_n = 0$ for some $n) = \rho$ where ρ is the unique value < 1 satisfying $E(\rho^\xi) = \rho$.

Definition 148 $\{(X_t, H_t); t \in T\}$ *is a* supermartingale *if*

(a) H_t is increasing (in t) family of sigma-algebras.

(b) Each X_t is H_t measurable and $E|X_t| < \infty$.

(c) For each $s < t$, with $s, t \in T$,we have $E(X_t|H_s) \le X_s$ a.s.

Theorem 149 *Suppose $A_n \ge 0$ is a predictable (non-anticipating) bounded sequence and X_n is a supermartingale, $n = 0, 1, ...$ Then the supermartingale transform $\widetilde{X} = A \circ X$ defined by*

$$\widetilde{X}_t = A_1(X_1 - X_0) + ... + A_t(X_t - X_{t-1}) \tag{7.13}$$

is a supermartingale.

Theorem 150 *Let (M_n, H_n); $n = 0, 1, 2, \ldots$ be a supermartingale such that $M_n \ge 0$. Then there is a random variable M such that $M_n \to M$ a.s. with $E(M) \le E(M_0)$.*

Example 151 *Let S_n be a simple symmetric random walk with $S_0 = 1$ and define the optional stopping time $N = \inf\{n; S_n = 0\}$. Then*

$$X_n = S_{n \wedge N}$$

is a non-negative (super)martingale and therefore $X_n \to$ almost surely. The limit must be 0 since otherwise, $|X_{n+1} - X_n| = 1$ and so convergence is impossible. However, in this case, $E(X_n) = 1$ whereas $E(X) = 0$ so the convergence is not in L_1.

Definition 152 $\{(X_t, H_t); t \in T\}$ *is a* reverse martingale *if*

(a) H_t is decreasing (in t) family of sigma-algebras.

(b) Each X_t is H_t- measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $E(X_s|H_t) = X_t$ a.s.

Example 153 *Let X be any integrable random variable, H_t be any decreasing family of sigma-algebras. Put $X_t = E(X|H_t)$. Then (X_t, H_t) is a reverse martingale.*

Theorem 154 *(Reverse martingale convergence Theorem). If (X_n, H_n); $n = 1, 2, \ldots$ is a reverse martingale,*

$$X_n \to E(X_1| \cap_{n=1}^{\infty} H_n) \quad a.s. \tag{7.14}$$

Example 155 *(The Strong Law of Large Numbers) Let Y_i be independent identically distributed, $H_n = \sigma(\bar{Y}_n, Y_{n+1}, Y_{n+2}, ...)$, where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then H_n is a decreasing family of sigma fields and $\bar{Y}_n = E(Y_1|H_n)$ is a reverse martingale. It follows from the reverse martingale convergence theorem that $\bar{Y}_n \to Y$ where Y is a random variable measurable with respect to $\cap_{n=1}^{\infty} H_n$. But $\cap_{n=1}^{\infty} H_n$ is in the tail sigma-field and so by the Hewitt-Savage 0-1 Law, Y is a constant almost surely and $Y = E(Y_i)$.*

Example 156 *(Hewitt-Savage 0-1 Law) Suppose Y_i are independent identically distributed and A is an event in the tail sigma-field. Then $P(A) = 0$ or $P(A) = 1$.*

7.4 Uniform Integrability

Definition 157 *A set of random variables $\{X_i, i = 1, 2, ...\}$ is uniformly integrable if*

$$\sup_i E(|X_i| I(|X_i| > c) \to 0 \quad as\ c \to \infty$$

7.4.1 Some Properties of uniform integrability:

1. Any finite set of integrable random variables is uniformly integrable.

2. Any infinite sequence of random variables which converges in L^1 is uniformly integrable.

3. Conversely if a sequence of random variables converges almost surely and is uniformly integrable, then it also converges in L^1.

4. If X is integrable on a probability space (Ω, H) and H_t any family of sub-sigma fields, then $\{E(X|H_t)\}$ is uniformly integrable.

5. If $\{X_n, n = 1, 2, ...\}$is uniformly integrable, then $\sup_n E(X_n) < \infty$.

Theorem 158 *Suppose a sequence of random variables satisfies $X_n \to X$ in probability. Then the following are all equivalent:*

1. $\{X_n, n = 1, 2, ...\}$ is uniformly integrable

2. $X_n \to X$ in L^1.

3. $E(|X_n|) \to E(|X|)$

Theorem 159 *Suppose X_n is a submartingale. Then the following are all equivalent:*

1. $\{X_n, n = 1, 2, ...\}$ is uniformly integrable

2. $X_n \to X$ almost surely and in L^1.

3. $X_n \to X$ in L^1.

Theorem 160 *Suppose X_n is a martingale. Then the following are all equivalent:*

1. $\{X_n, n = 1, 2, ...\}$ is uniformly integrable

2. $X_n \to X$ almost surely and in L^1.

3. $X_n \to X$ in L^1.

4. There exists some integrable X such that $X_n = E(X|H_n)$ a.s.

7.5 Martingales and Finance

Let $S(t)$ denote the price of a security at the beginning of period $t = 0, 1, 2, ...T$. We assume that the security pays no dividends. Define the *(cumulative) returns process* associated with this security by R_S where

$$\Delta R_S(t) = R_S(t) - R_S(t-1) = \frac{\Delta S(t)}{S(t-1)} = \frac{S(t) - S(t-1)}{S(t-1)}, \quad R_S(0) = 0.$$

Then $100\Delta R_S(t)\%$ is the percentage return in an investment in the stock in the $t-1'$st period. The returns process is a more natural characterisation of stock prices than the original stock price process since it is invariant under artificial scale changes such as stock splits etc. Note that we can write the stock price in terms of the returns process;

$$S(t) = S(0)\prod_{i=1}^{t}(1 + \Delta R_S(i)).$$

Now consider another security, a *riskless discount bond* which pays no coupons. Assume that the price of this bond at time t is $B(t)$, $B(0) = 1$ and $R_B(t)$ is the return process associated with this bond. Then $\Delta R_B(t) = r(t)$ is the interest rate paid over the $t-1$'st period. It is usual that the interest paid over the $t-1$st period should be declared in advance, i.e. at time $t-1$ so that if $S(t)$ is adapted to a filtration $\mathcal{F}_t$, then $r(t)$ is *predictable* , i.e. is $\mathcal{F}_{t-1}-$measurable. The *discounted stock price process* is the process given by

$$S^*(t) = S(t)/B(t).$$

Consider a *trading strategy* of the form $(\beta(t), \alpha(t))$ representing the total number of shares of bonds and stocks respectively held at the beginning of the period $(t-1, t)$. Since our investment strategy must be determined by using only the present and the past values of this and related processes, both $\beta(t)$ and $\alpha(t)$ are predictable processes. Then the value of our investment at time $t-1$ is $V_{t-1} = \beta(t)B(t-1) + \alpha(t)S(t-1)$ and at the end of this period, this changes to $\beta(t)B(t) + \alpha(t)S(t)$ with the difference $\beta(t)\Delta B(t) + \alpha(t)\Delta S(t)$ representing the *gain* over this period. An investment strategy is *self-financing* if the value after rebalancing the portfolio is the value before- i.e. if all investments are paid for by the above gains. In other words if $V_t = \beta(t)B(t) + \alpha(t)S(t)$ for all t. An *arbitrage opportunity* is a trading strategy that makes money with no initial investment; i.e. one such that $V_0 = 0$, $V_t \geq 0$ for all $t = 1, \ldots T$ and $E(V_T) > 0$. The basic theorem of no-arbitrage pricing is the following:

7.5.1 Theorem

There are no arbitrage opportunities in the above economy if and only if there is a measure Q equivalent to the underlying measure P i.e. $P << Q$ and $Q << P$ such that under Q the discounted process is a martingale; i.e. $E_Q(S^*(t)|\mathcal{F}_{t-1}] = S^*(t-1)$ a.s. for all $t \leq T$.

Proof; See Pliska (3.19)) page 94.

Note: The measure Q is called the equivalent martingale measure and is used to price derivative securities. For any attainable contingent claim X; (a for any random variable X which can be written as a linear function of the available investments), the arbitrage-free price at time t is given by the conditional expected value under Q of the discounted return X given $\mathcal{F}_\sqcup$.

7.6 Problems

1. Let $(\Omega, \mathcal{F}, P)$ be the unit interval with the Borel sigma-algebra and Lebesgue measure defined thereon. Define $\mathcal{F}_n$ to be the sigma field generated by the intervals $(\frac{j-1}{2^n}, \frac{j}{2^n}]$, $j = 1, 2, \ldots 2^n$. Let X be a bounded continuous function on the unit interval.
   (a) Find $E(X|\mathcal{F}_n)$.
   (b) Show $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all n .
   (c) Verify that $E(X|\mathcal{F}_n)$ converges pointwise and identify the limit.
   (d) Verify directly that $E\{E(X|\mathcal{F}_n)\} = E(X)$.
   (e) What could you conclude if X had countably many points of discontinuity?
2. Prove property (i), that if Z is $\mathcal{G}-$measurable, $E(ZX|\mathcal{G}) = ZE(X|\mathcal{G})$ a.s.
3. Suppose that X is integrable so that $E(|X|) < \infty$. Prove for constants c, d that $E(cX+d|\mathcal{G}) = cE(X|\mathcal{G})+d$ (First give the proof in case $E(X^2) < \infty$).
4. Prove property (j): if $\mathcal{H} \subset \mathcal{G}$ are sigma-algebras, $E[E(X|\mathcal{G})|\mathcal{H}] = E(X|\mathcal{H})$. Does the same hold if $\mathcal{G} \subset \mathcal{H}$?
5. Prove: if $X \le Y$, $E(X|\mathcal{G}) \le E(Y|\mathcal{G})$ a.s.
6. Prove: $var(X) = E\{var(X|\mathcal{G})\} + var\{E(X|\mathcal{G})\}$.
7. Prove that if X and Y are simple random variables, $X = \sum c_i I_{A_i}$ and $Y = \sum_j d_j I_{B_j}$ then
$$E(X|Y)(\omega) = \sum_j \sum_i c_i P(A_i|B_j) I_{B_j}(\omega).$$
8. Suppose X is a normal$(0,1)$ variate and $Y = XI(X \le c)$. Find $E(X|Y)$.
9. Suppose X and Y are independent exponentially distributed random variables each with mean 1. Let I be the indicator random variable $I = I(X > Y)$. Find the conditional expectations
   (a) $E(X|I)$

(b) $E(X+Y|I)$

10. Suppose X is a random variable having the Poisson(λ) distribution and define the indicator random variable $I = I(X$ is even$)$. Find $E(X|I)$.

11. Consider the pair of random variables (X_n, Y_n) where $X_n = X$, $Y_n = (1/n)X$ for all $n = 1, 2, \ldots$. Show that (X_n, Y_n) converges almost surely to some (X, Y) but it is NOT true in general that $E(X_n|Y_n) \to E(X|Y)$ almost surely or that $E(X_n|Y_n) \to E(X|Y)$ weakly.

12. Suppose Y_i are independent identically distributed. Define $\mathcal{F}_n = \sigma(Y_{(1)}, ..., Y_{(n)}, Y_{n+1}, Y_{n+2}, ...)$, where $(Y_{(1)}, ..., Y_{(n)})$ denote the order statistics. Show $\mathcal{F}_n$ is a decreasing family of sigma fields, find $s_n^2 = E(\frac{1}{2}(Y_1 - Y_2)^2|\mathcal{F}_n)$ and show it is a reverse martingale. Conclude a limit theorem.

13. Let X be an arbitrary absolutely continuous random variable with probability density function $f(x)$. Let $\alpha(s) = f(s)/P[X \geq s]$ denote the hazard function. Show

$$X_t = I(X \geq t) - \int_{-\infty}^{min(X,t)} \alpha(s)ds$$

is a martingale with respect to a suitable family of sigma-algebras.

14. Suppose $(X_t, \mathcal{F}_t)$ is a martingale and a random variable Y is independent of every $\mathcal{F}_t$. Show that we continue to have a martingale when $\mathcal{F}_t$ is replace by $\sigma(Y, \mathcal{F}_t)$.

15. Suppose τ is an optional stopping time taking values in a interval $\{1, 2, ..., n\}$. Suppose $\{(X_t, \mathcal{F}_t); t = 1, 2, ..., n\}$ is a martingale. Prove $E(X_\tau) = E(X_1)$.

16. Prove the general case of Doob's maximal inequality, that for $p > 1, \lambda > 0$ and a non-negative submartingale M_n,

$$P(\sup_{0 \leq m \leq n} M_m \geq \lambda) \leq \lambda^{-p} E(M_n^p)$$

17. Consider a stock price process $S(t)$ and a riskless bond price process $B(t)$ and their associated returns process $\Delta R_S(t)$ and $\Delta R_B(t) = r(t)$. Assume that the stock price takes the form of a binomial tree; $S(t) = S(t-1)[d + (u-d)X_t]$ where X_t are independent Bernoulli random variables adapted to some filtration $\mathcal{F}_t$ and where $d < 1 < 1 + r(t) < u$ for all t. We assume that under the true probability measure P, $P(X_t = 0)$ and $P(X_t = 1)$ are positive for all t.

    Determine a measure Q such that the discounted process $S^*(t) = \frac{S(t)}{B(t)}$ is a martingale under the new measure Q and such that Q is equivalent to P i.e. $P << Q$ and $Q << P$. Is this measure unique? What if we were to replace the stock price process by one which had three branches at each step, i.e. it either stayed the same, increased by a factor u or decreased by factor d at each step (a trinomial tree)?

18. Prove that if, under a measure Q, the expected return from a stock is the risk-free interest rate; i.e. if

$$E_Q[\Delta R_S(t)|\mathcal{F}_{t-1}] = r(t) \quad a.s.$$

then the discounted price process $S^*(t)$ is a martingale under Q.

19. Prove that for an optional stoping time τ, $\sigma(\tau) \subset H_\tau$.

20. Let $X_1, X_2, ...$ be a sequence of independent random variables all with the same expected value μ. Suppose τ is an optional stopping time with respect to the filtration $H_t = \sigma(X_1, X_2, ..., X_t), t = 1, 2, ...$ and assume that

$$E(\sum_{i=1}^{\tau} |X_i|) < \infty.$$

Prove that

$$E(\sum_{i=1}^{\tau} X_i) = \mu E(\tau).$$

21. Find an example of a martingale $X_t, t = 1, 2, ...$ and an optional stopping time τ such that

$$P[\tau < \infty] = 1$$

but X_τ is not integrable.