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

Proof. Suppose $X = \sum_i c_i I_{A_i} \leq \sum_j d_j I_{B_j}$ 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 $c_i \leq d_j$ whenever $A_i B_j \neq \phi$. Therefore

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

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 $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then we define

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

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 $X(\omega) \leq Y(\omega)$ for all ω, then $E(X) \leq E(Y)$.
3. If the sequence of non-negative random variables X_n is increasing to a random variable 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 $Z \leq X$ and Z is a simple function, then $E(Z) \leq E(X)$. 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 $Z \leq X$. Then $Z \leq Y$. It follows that the set of Z satisfying $Z \leq X$ is a subset of the set satisfying $Z \leq Y$ and therefore $sup\{E(Z); Z \text{ is simple}, Z \leq X\} \leq sup\{E(Z); Z \text{ is simple}, Z \leq Y\}$.

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

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

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

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

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

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

whenever Y is a simple function satisfying $Y \leq X$. 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 $\epsilon > 0$, $\lim E(X_n) \geq (1-\epsilon)E(X)$ 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 $P(A) = 0, \quad \int_A X(\omega)dP = 0.$

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 $limY_n = lim\ inf\ X_n = X$, say. Therefore by monotone convergence, $E(Y_n) \to E(X)$. Since $Y_n \le X_n$ for all n ,

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

■

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. $f(x) \leq g(x)$ for all x implies $\int f(x)d\mu \leq \int g(x)d\mu$.
2. For real numbers $\alpha,\ \beta,\ \int(\alpha f + \beta g)d\mu = \alpha \int f d\mu + \beta \int g d\mu$.
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*

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

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 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,

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

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 $\mu_s(B^c) \leq \mu(B^c) = 0$ which shows that the measure μ_s is identically 0. This now shows that

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

■

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, $F(x) = \sum_{\{n;x_n \leq x\}} p_n$. 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 $P(X \in B) = \int_B dF(x) = 1$. (For example, the uniform distribution on the Cantor set).

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 $|E(X)| \le E(|X|)$. Similarly, show $|E(X)| \le \sqrt{E(|X|^2)}$.

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 $\int x dF$ 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 $e^{-rT}E((S_T - K)^+)$ 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 $e^{-rT}E(S_T - K)^+ = V(S_0, T)$ 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 $H(g) \leq H(f)$.

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

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

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

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 $\mu(B) - n^{-1}\lambda(B) \leq 0$ 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 $\frac{1}{n}\lambda(E) \leq \mu(E)$ 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 $E(X)$.