

Sample Space and Collections of Sets

A sample space is a set. It is the “universe of discourse” in a given problem. It is often denoted by Ω .

We usually denote collections of sets with upper-case caligraphic letters, e.g., \mathcal{A} , \mathcal{F} , etc.

The usual set operators and set relations are used with collections of sets, and generally have the same meaning.

An important collection of sets is the collection of all open intervals of \mathbb{R} .

σ -Fields

A collection of subsets, \mathcal{F} , of a given sample space, Ω , such that

1. $\emptyset \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$.

Notice that a σ -field is defined with respect to some given sample space.

A σ -field is also called a σ -algebra. The definition of a σ -field sometimes includes the condition that $\Omega \in \mathcal{F}$ rather than $\emptyset \in \mathcal{F}$.

Given any sample space Ω and any collection \mathcal{C} of subsets of Ω , the “smallest” σ -field of which \mathcal{C} is a subset is called the σ -field generated by \mathcal{C} , and is denoted by $\sigma(\mathcal{C})$.

Example: If $\mathcal{A} = \{A\}$, with respect to the sample space Ω , $\sigma(\mathcal{A}) = \{\emptyset, A, A^c, \Omega\}$.

σ -fields grow very rapidly.

If k is the maximum number of sets that partition Ω that can be formed by operations on the sets in \mathcal{A} , then the number of sets in the σ -field is 2^k .

Measurable Space: The Structure (Ω, \mathcal{F})

If Ω is a sample space, and \mathcal{F} is a σ -field over Ω , the double (Ω, \mathcal{F}) is called a *measurable space*.

Notice that no “measure” is required.

We have not yet defined *measure*.

Functions and Images

A function is a set of ordered pairs such that no two pairs have the same first element.

We use “function” and “mapping” synonymously, although the latter term is sometimes interpreted more generally.

To say that f is a mapping from Ω to Λ , written

$$f : \Omega \mapsto \Lambda,$$

means that for every $\omega \in \Omega$ there is a pair in f whose first member is ω .

We use the notation $f(\omega)$ to represent the second member of the pair in f whose first member is ω .

Types include functions that are *onto*, meaning that for every $\lambda \in \Lambda$ there is a pair in f whose second member is λ ; and functions that are *one-to-one*, written $1 : 1$, meaning that no two pairs have the same second member.

A function that is one-to-one and onto is called a *bijection*.

A function f that is one-to-one has an inverse, written f^{-1} .

If $(a, b) \in f$, we may write $a = f^{-1}(b)$, although sometimes this notation is restricted to the cases in which f is one-to-one.

(There are some subtleties here; if f is not one-to-one, if the members of the pairs in f are reversed, the resulting set is not a function. We say f^{-1} does not exist; yet we may write $a = f^{-1}(b)$, with the meaning above.)

If $A \subset \Omega$, the *image* of A , denoted by $f[A]$, or just by $f(A)$, is the set of all $\lambda \in \Lambda$ for which $\lambda = f(\omega)$ for some $\omega \in \Omega$.

(The notation $f[A]$ is preferable, but we will often just use $f(A)$ in this course.) Similarly, if \mathcal{C} is a collection of sets, the notation $f[\mathcal{C}]$ denotes the collection of sets $\{f[C] : C \in \mathcal{C}\}$.

For the function f that maps from Ω to Λ , Ω is called the *domain* of the function and $f[\Omega]$ is called the *range* of the function.

For a subset B of Λ , the *inverse image* or the *preimage* of B , denoted by $f^{-1}[B]$, or just by $f^{-1}(B)$, is the set of all $\omega \in \Omega$ such that $f(\omega) \in B$.

The notation f^{-1} used in this sense must not be confused with the inverse function f^{-1} (if the latter exists).

We use this notation for the inverse image whether or not the inverse of the function exists.

Notice that the inverse image of a set may not generate the set; that is, $f[f^{-1}[B]] \subset B$.

We also write $f[f^{-1}[B]]$ as $f \circ f^{-1}[B]$. The set $f[f^{-1}[B]]$ may be a proper subset of B ; that is, there may be an element λ in B for which there is no $\omega \in \Omega$ such that $f(\omega) = \lambda$. If f is bijective, then $f[f^{-1}[B]] = B$.

The following are useful facts about a function f that maps Ω to Λ .

$$(i) f(\emptyset) = \emptyset \text{ and } f^{-1}(\emptyset) = \emptyset$$

This is the only way that makes sense!

(ii) For $B \subset \Lambda$, $f^{-1}(B^c) = (f^{-1}(B))^c$

(where $B^c = \Lambda - B$, and $(f^{-1}(B))^c = \Omega - f^{-1}(B)$).

We see this in the standard way by showing that each is a subset of the other.

Let ω be an arbitrary element of Ω .

Suppose $\omega \in f^{-1}(B^c)$. Then $f(\omega) \in B^c$, so $f(\omega) \notin B$, hence $\omega \notin f^{-1}(B)$, and so $\omega \in (f^{-1}(B))^c$. We have $f^{-1}(B^c) \subset (f^{-1}(B))^c$.

Now suppose $\omega \in (f^{-1}(B))^c$. Then $\omega \notin f^{-1}(B)$, so $f(\omega) \notin B$, hence $f(\omega) \in B^c$, and so $\omega \in f^{-1}(B^c)$. We have $(f^{-1}(B))^c \subset f^{-1}(B^c)$.

(iii) Let $A_1, A_2, \dots \subset \Lambda$ and suppose $(\cup_{i=1}^{\infty} A_i) \subset \Lambda$, then $f^{-1}(\cup_{i=1}^{\infty} A_i) = \cup_{i=1}^{\infty} f^{-1}(A_i)$.

Again, let λ be an arbitrary element of Λ .

Suppose $\lambda \in f^{-1}(\cup_{i=1}^{\infty} A_i)$. Then $f(\lambda) \in \cup_{i=1}^{\infty} A_i$, so for some j , $f(\lambda) \in A_j$ and $\lambda \in f^{-1}(A_j)$; hence $\lambda \in \cup_{i=1}^{\infty} f^{-1}(A_i)$. We have $f^{-1}(\cup_{i=1}^{\infty} A_i) \subset \cup_{i=1}^{\infty} f^{-1}(A_i)$.

Now suppose $\lambda \in \cup_{i=1}^{\infty} f^{-1}(A_i)$. Then for some j , $\lambda \in f^{-1}(A_j)$, so $f(\lambda) \in A_j$ and $f(\lambda) \in \cup_{i=1}^{\infty} A_i$; hence $\lambda \in f^{-1}(\cup_{i=1}^{\infty} A_i)$. We have $\cup_{i=1}^{\infty} f^{-1}(A_i) \subset f^{-1}(\cup_{i=1}^{\infty} A_i)$.

Measurable Functions

If (Ω, \mathcal{F}) and (Λ, \mathcal{G}) are measurable spaces, and f is a mapping from Ω to Λ , with the property that $\forall A \in \mathcal{G}, f^{-1}(A) \in \mathcal{F}$, then f is a *measurable* function with respect to \mathcal{F} and \mathcal{G} . It is also said to be measurable \mathcal{F}/\mathcal{G}

Note that a measurable function $f(\cdot)$ does not depend on a measure.

The domain of $f(\cdot)$ has no relationship to \mathcal{F} , except through the range of $f(\cdot)$ that happens to be in the subsets in \mathcal{G} .

Product Spaces

Two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ can be used to form a *cartesian product measurable space* with sample space $\Omega_1 \times \Omega_2$. The product of the σ -fields is not necessarily a σ -field.

Given $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, we define the cartesian product measurable space as $(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2))$.

This provides us the basis for developing a probability theory for vectors and multivariate distributions.

The Borel σ -field

Consider the sample space \mathbb{R} , and let \mathcal{C} be the collection of all open intervals in \mathbb{R} . The σ -field $\sigma(\mathcal{C})$ is called the Borel σ -field, and is usually denoted by \mathcal{B} .

Borel Sets

Any set in \mathcal{B} is called a Borel set.

The following are Borel sets:

- \mathbb{R}
- \emptyset
- any countable set; in particular, any finite set, \mathbb{Z} , \mathbb{Z}_+ (the *natural numbers*), and the set of all rational numbers
- hence, from the foregoing, the set of all irrational numbers (which is uncountable)
- any interval, open, closed, or neither

- the Cantor set

We see this by writing the Cantor set as $\bigcap_{i=1}^{\infty} C_i$, where

$$C_1 = [0, 1/3] \cup [2/3, 1], \quad C_2 = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1],$$

and realizing that each of these is Borel.

- any union of any of the above

So, are all subsets of \mathbb{R} Borel sets?

No. Interestingly enough, the cardinality of \mathcal{B} can be shown to be the same as that of \mathbb{R} , and the cardinality of the collection of all subsets of \mathbb{R} , that is, the cardinality of the power set, $2^{\mathbb{R}}$, is much larger – which means there are *many* subsets of \mathbb{R} that are not Borel sets. Construction of a non-Borel set uses the Axiom of Choice, which can be used to construct some truly weird sets.

Product Borel σ -fields

For the product measurable space generated by \mathbb{R}^k , a σ -field is $\sigma(\mathcal{B}^k)$. It can be shown that this is the same as the σ -field generated by all open intervals (or “hyperrectangles”) in \mathbb{R}^k . We denote this product measurable space as $(\mathbb{R}^k, \mathcal{B}^k)$.

The σ -field $\mathcal{B}_{[0,1]}$

This is the σ -field generated by all open intervals on $[0, 1]$.

Borel Function

A measurable function from (Ω, \mathcal{F}) to $(\mathbb{R}, \mathcal{B})$, where \mathcal{B} is the Borel σ -field, is said to be *Borel measurable* with respect to \mathcal{F} . A function that is Borel measurable is called a *Borel function*.

Random Variables

A measurable function from any measurable space (Ω, \mathcal{F}) to the measurable space $(\mathbb{R}, \mathcal{B})$ is called a *random variable*. That is, “Borel function” and “random variable” are synonymous.

This is the definition of the phrase *random variable*; the words “random” and “variable” do not carry any separate meaning.

Measures

A measure is a real-valued nonnegative set function whose domain is a σ -field and with the property that the measure of the union of any collection of disjoint sets is the sum of the measures of the sets.

From this simple definition, several properties derive. For example, if ν is a measure with domain \mathcal{F} then

- $\nu(\emptyset) = 0$
- if $A_1 \subset A_2 \in \mathcal{F}$, then $\nu(A_1) \leq \nu(A_2)$ (this is “monotonicity”)
- if $A_1, A_2, \dots \in \mathcal{F}$, then $\nu(\cup_i A_i) \leq \sum_i \nu(A_i)$ (this is “subadditivity”)

- if $A_1 \subset A_2 \subset \dots \in \mathcal{F}$, then $\nu(\cup_{i=1}^{\infty} A_i) = \lim_{i \rightarrow \infty} \nu(A_i)$ (this is “continuity from below”; think of the A_i s as nested intervals).

We can see this by defining a sequence of disjoint sets, B_1, B_2, \dots as $B_j = A_{j+1} - A_j$, so $\cup_{j=1}^i B_j = A_i$, and

$$\cup_{i=1}^{\infty} B_i = \cup_{i=1}^{\infty} A_i$$

Hence,

$$\begin{aligned} \nu(\cup_{i=1}^{\infty} A_i) &= \nu(\cup_{i=1}^{\infty} B_i) \\ &= \sum_{i=1}^{\infty} \nu(B_i) \\ &= \lim_{i \rightarrow \infty} \sum_{j=1}^i \nu(B_j) \\ &= \lim_{i \rightarrow \infty} \nu(\cup_{j=1}^i B_j) \\ &= \lim_{i \rightarrow \infty} \nu(A_i). \end{aligned}$$

Sequences of nested intervals are important. We denote a sequence $A_1 \subset A_2 \subset \dots$ with $A = \bigcup_{i=1}^{\infty} A_i$, as $A_i \nearrow A$. (This same notation is used for a sequence of real numbers x_i such that $x_1 \leq x_2 \leq \dots$ and $\lim x_i = x$; that is, in that case, we write $x_i \nearrow x$.)

Continuity from below is actually a little stronger than what is stated above, because the sequence of values of the measure is also monotonic: for

$$A_i \in \mathcal{F}, A_i \nearrow A \Rightarrow \nu(A_i) \nearrow \nu(A)$$

A similar sequence is $A_1 \supset A_2 \supset \dots$ with $A = \bigcup_{i=1}^{\infty} A_i$. We denote this as $A_i \searrow A$. Continuity from above is the fact that for $A_i \in \mathcal{F}$, if $A_i \searrow A$ and $\nu(A_1) < \infty$, then $\nu(A_i) \searrow \nu(A)$.

To evaluate $\nu(\cup_i A_i)$ we form disjoint sets by intersections. For example, we have $\nu(A_1 \cup A_2) = \nu(A_1) + \nu(A_2) - \nu(A_1 \cap A_2)$. This is the simplest form of the inclusion-exclusion formula.

If there are three sets, we take out all pairwise intersections and then add back in the triple intersection.

We can easily extend this (the proof is by induction) so that, in general, we have

$$\begin{aligned} \nu(\cup_i^n A_i) = & \sum_{1 \leq i \leq n} \nu(A_i) - \\ & \sum_{1 \leq i < j \leq n} \nu(A_i \cap A_j) + \\ & \sum_{1 \leq i < j < k \leq n} \nu(A_i \cap A_j \cap A_k) - \\ & \dots + \\ & (-1)^{n+1} \nu(A_1 \cap \dots \cap A_n). \end{aligned}$$

If \mathcal{F} is the Borel σ -field, the most common measure is the *Lebesgue measure*, which is defined by the relation

$$\nu((a, b)) = b - a.$$

If \mathcal{F} is a countable σ -field, the most common measure is the *counting measure*, which is defined by the relation

$$\nu(A) = \#(A).$$

If \mathcal{F} is the collection of all subsets of Ω , a useful measure at a fixed point $\omega \in \Omega$ is the *Dirac measure* concentrated at ω , usually denoted by δ_ω , and defined by

$$\delta_\omega(A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

Measurable Sets

If ν is a measure with domain \mathcal{F} then every set in \mathcal{F} is said to be ν -*measurable*, or just *measurable*.

Note that unlike other terms above that involve “measurable”, this term is defined in terms of a given measure.

Measure Space: The Structure $(\Omega, \mathcal{F}, \nu)$

If Ω is a sample space, \mathcal{F} is a σ -field over Ω , and ν is a measure with domain \mathcal{F} , the triple $(\Omega, \mathcal{F}, \nu)$ is called a *measure space* (compare *measurable space*, above).

The elements in the measure space can be any kind of objects. They do not need to be numbers.

σ -Finite Measure

A measure ν is σ -finite on (Ω, \mathcal{F}) iff there exists a sequence A_1, A_2, \dots in \mathcal{F} such that $\cup_i A_i = \Omega$ and $\nu(A_i) < \infty$ for all i .

The Lebesgue measure is σ -finite, as can be seen from the sequence of open intervals $(-i, i)$.

The counting measures is σ -finite iff Ω is countable.

Almost Everywhere (a.e.)

Given a measure space, a property that holds for any element of the σ -field with positive measure is said to hold *almost everywhere*, or a.e.

Probability Measure

A measure whose domain is a σ -field defined on the sample space Ω with the property that $\nu(\Omega) = 1$ is called a *probability measure*. We often use P to denote such a measure.

A property that holds a.e. with respect to a probability measure is said to hold *almost surely*, or a.s.

Support of a Probability Measure

If the probability measure P is defined with respect to the σ -field \mathcal{F} , $S \in \mathcal{F}$, and $P(S) = 1$, S is called a *support* of the probability measure (or of the probability space, or of the “distribution”). If S_1 and S_2 are supports of P , then $S_1 = S_2$ a.s.

Probability Space

If P in the measure space (Ω, \mathcal{F}, P) is a probability measure, the triple (Ω, \mathcal{F}, P) is called a *probability space*.

The elements in the probability space can be any kind of objects. They do not need to be numbers.

Cumulative Distribution Function (CDF)

If $(\mathbb{R}, \mathcal{B}, P)$ is a probability space, and F is defined by

$$F(x) = P((-\infty, x]) \quad \forall x \in \mathbb{R},$$

then F is called a cumulative distribution function.

The probability space completely determines F , and likewise, F completely determines P a.s.

Induced Measure

If $(\Omega, \mathcal{F}, \nu)$ is a measure space and (Λ, \mathcal{G}) is a measurable space, and f is a function from Ω to Λ that is measurable with respect to \mathcal{F} , then $\nu \circ f^{-1}$ is an *induced measure* on \mathcal{G} .

Note that the domain and range of the function $\nu \circ f^{-1}$ is \mathcal{G} .
Note, further, that it is a measure.

Simple Functions

If A_1, \dots, A_k are measurable subsets of Ω and a_1, \dots, a_k are constant real numbers, a function φ is a *simple function* if for $\omega \in \Omega$,

$$\varphi(\omega) = \sum_{i=1}^k a_i \mathbf{I}_{A_i}(\omega),$$

where $\mathbf{I}_S(x)$ is the *indicator function*, i.e., $\mathbf{I}_S(x) = 1$ if $x \in S$ and $\mathbf{I}_S(x) = 0$ otherwise.

Real-Valued Functions over Real Domains

Random variables are real-valued functions

Measurable real-valued functions are called Borel functions.

Now we will define integrals and derivatives of real-valued functions.

For real-valued functions over real domains we identify some additional properties.

We have already defined similar properties for functions that are measures (real-valued set functions), and we could define them for other functions from one measure space to another. The definitions for real-valued functions over real domains are simpler and more useful.

Continuous Functions

Let f be a real-valued function whose domain is a set $D \in \mathbb{R}$. We say that f is *continuous at the point* $x \in D$ if, given $\epsilon > 0$, $\exists \delta \ni \forall y \in D \ni |x - y| < \delta, |f(x) - f(y)| < \epsilon$.

If f is continuous at each point in a subset of its domain, we say it is continuous on that subset.

If f is continuous at each point in its domain, we say it is *continuous*.

From this definition we have an immediate useful fact about continuous functions: the inverse image of an open set is open.

Absolutely Continuous Functions

Let f be a real-valued function defined on $[a, b]$ (its domain may be larger). We say that f is *absolutely continuous* on $[a, b]$ if, given $\epsilon > 0$, there exists a δ such that for every finite collection of nonoverlapping intervals $(x_i, y_i) \subset [a, b]$ with

$$\sum_{i=1}^n |x_i - y_i| < \delta,$$

$$\sum_{i=1}^n |f(x_i) - f(y_i)| < \epsilon.$$

If f is absolutely continuous on $[a, b]$, it is continuous on $[a, b]$.

An absolutely continuous function is of bounded variation; it has a derivative almost everywhere; and if the derivative is 0 a.e., the function is constant.

Simple Functions

Let f be real and measurable. If $f(\omega) \geq 0$, there exists a sequence $\{f_n\}$ of simple functions such that

$$0 \leq f_n(\omega) \nearrow f(\omega) \quad \text{a.s.},$$

and if $f(\omega) \leq 0$, there exists a sequence $\{f_n\}$ of simple functions such that

$$0 \geq f_n(\omega) \searrow f(\omega) \quad \text{a.s.}$$

The sequence is

$$f_n(\omega) = \begin{cases} -n & \text{if } f(\omega) \leq -n, \\ -(k-1)2^{-n} & \text{if } -k2^{-n} < f(\omega) \leq -(k-1)2^{-n}, \text{ for } 1 \leq k \leq n, \\ (k-1)2^{-n} & \text{if } (k-1)2^{-n} < f(\omega) < k2^{-n}, \text{ for } 1 \leq k \leq n, \\ n & \text{if } n \leq f(\omega). \end{cases}$$

In particular, if X is a nonnegative random variable, there exists a sequence of simple (degenerate) random variables $\{X_n\}$ such that

$$0 \leq X_n \nearrow X \quad \text{a.e.}$$

Integration

Integrals are some of the most important functionals of real-valued functions. Integrals and the action of integration are defined using measures.

Integrals of nonnegative functions are themselves measures.

There are various types of integrals, Lebesgue, Riemann, Riemann-Stieltjes, Ito, and so on. The most important is the Lebesgue, and when we use the term “integral” without qualification that will be the integral meant.

The (Lebesgue) Integral of a Function with Respect to a Given Measure

An *integral of a function f with respect to a given measure ν* , if it exists, is a functional whose value is an average of the function weighted by the measure. It is denoted by $\int f d\nu$.

We build the definition of an integral of a function in three steps.

1. Simple function.

If f is a simple function defined as $f(\omega) = \sum_{i=1}^k a_i I_{A_i}(\omega)$, where the A_i s are measurable with respect to ν , then

$$\int f d\nu = \sum_{i=1}^k a_i \nu(A_i).$$

(Note that a simple function over measurable A_i s is measurable.)

2. Nonnegative Borel function.

We define the integral of a nonnegative Borel function in terms of the supremum of a collection of simple functions. Let f be a nonnegative Borel function with respect to ν on Ω , and let S_f be the collection of all nonnegative simple functions such that

$$\varphi \in S_f \Rightarrow \varphi(\omega) \leq f(\omega) \forall \omega \in \Omega$$

We define the integral of f with respect to ν as

$$\int f \, d\nu = \sup \left\{ \int \varphi \, d\nu \mid \varphi \in S_f \right\}.$$

3. General Borel function.

For a general Borel function f , we form two nonnegative Borel functions f_+ and f_- such that $f = f_+ - f_-$:

$$f_+(\omega) = \max\{f(\omega), 0\}$$

$$f_-(\omega) = \max\{-f(\omega), 0\}$$

We define the integral of f with respect to ν as the difference of the integrals of the two nonnegative functions:

$$\int f \, d\nu = \int f_+ \, d\nu - \int f_- \, d\nu,$$

so long as either $\int f_+ \, d\nu$ or $\int f_- \, d\nu$ is finite. ($\infty - \infty$ is not defined.)

The (Lebesgue) Integral of a Function with Respect to a Given Measure

This definition of an integral immediately yields some important properties of integrals:

- linearity: for real a and Borel f and g ,
$$\int af + g \, d\nu = a \int f \, d\nu + \int g \, d\nu.$$
- $\int |f| \, d\nu$ is a norm: $\int |f| \, d\nu = 0 \Rightarrow f = 0$ a.e.
This fact together with the linearity means that $\int |f| \, d\nu$ is a *norm* for functions. A more general form norm is $(\int |f|^p \, d\nu)^{1/p}$ for $1 \leq p$.
- finite monotonicity: for integrable f and g ,
 $f \leq g$ a.e. $\Rightarrow \int f \, d\nu \leq \int g \, d\nu.$

Integration

There are some conditions for interchange of an integration operation and a limit operation that are not so obvious.

- monotone convergence: if $0 \leq f_1 \leq f_2$ and $\lim_{n \rightarrow \infty} f_n = f$ a.e., then $\int \lim_{n \rightarrow \infty} f_n \, d\nu = \lim_{n \rightarrow \infty} \int f_n \, d\nu$.
- Fatou's lemma (follows from finite monotonicity and monotone convergence): if $0 \leq f_n$ then $\int \lim_n \inf f_n \, d\nu \leq \lim_n \inf \int f_n \, d\nu$.
- dominated convergence: if $\lim_{n \rightarrow \infty} f_n = f$ a.e. and there exists an integrable function g such that $|f_n| \leq g$ a.e., then $\int \lim_{n \rightarrow \infty} f_n \, d\nu = \lim_{n \rightarrow \infty} \int f_n \, d\nu$.

Shao gives proofs of these (but he should prove them in the order given above).

Integration

If the measure is a probability measure P with associated CDF F , all of the following notations are equivalent:

$$\int f \, dP, \int f(x) \, dP, \int f \, dF, \int f(x) \, dF(x)$$

The Riemann integral is one of the simplest integrals. We can define the Riemann integral of a real function f over the interval $(a, b]$ in terms of the Lebesgue measure λ as the real number r such that for any $\epsilon > 0$, there exists a δ such that

$$\left| r - \sum_i f(x_i) \lambda(I_i) \right| < \epsilon$$

where $\{I_i\}$ is any finite partition of $(a, b]$ such that for each i , $\lambda(I_i) < \delta$ and $x_i \in I_i$. If the Riemann integral exists, it is the same as the Lebesgue integral.

A classic example for which the Lebesgue integral exists, but the Riemann integral does not, is the function g defined over $(0, 1]$ as $g(x) = 1$ if x is rational, and $g(x) = 0$ otherwise.

The Lebesgue integral $\int_0^1 g(x) dx$ exists and equals 0, because $g(x) = 0$ a.e.

The Riemann integral, on the other hand does not exist because for an arbitrary partition $\{I_i\}$, the integral is 1 if $x_i \in I_i$ is taken as a rational, and the integral is 0 if $x_i \in I_i$ is taken as an irrational.

Integrable Function

If $f = f_+ - f_-$, where f_+ and f_- are nonnegative Borel functions, and both $\int f_+ d\nu$ and $\int f_- d\nu$ are finite, we say f is *integrable*.

Note that being Borel does not imply that a function is integrable.

A random variable is not necessarily integrable.

Change of Variables

Consider two measurable spaces (Ω, \mathcal{F}) and (Λ, \mathcal{G}) , let f be a measurable function from (Ω, \mathcal{F}) to (Λ, \mathcal{G}) , and let ν be a measure on \mathcal{F} . As we have seen, $\nu \circ f^{-1}$ is an induced measure on \mathcal{G} .

Now let g be a Borel function on (Λ, \mathcal{G}) .

Then the integral of $g \circ f$ over Ω with respect to ν is the same as the integral of g over Λ with respect to $\nu \circ f^{-1}$:

$$\int_{\Omega} g \circ f \, d\nu = \int_{\Lambda} g \, d(\nu \circ f^{-1})$$

Integration in a Product Space (Fubini's Theorem)

Given two measure spaces $(\Omega_1, \mathcal{F}_1, \nu_1)$ and $(\Omega_2, \mathcal{F}_2, \nu_2)$ and a Borel function f on $\Omega_1 \times \Omega_2$, the integral over Ω_1 , if it exists, is a function of $\omega_2 \in \Omega_2$ a.e., and likewise, the integral over Ω_2 , if it exists, is a function of $\omega_1 \in \Omega_1$ a.e.

Fubini's theorem shows that if one of these marginal integrals, say

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1,$$

exists a.e., then the natural extension of an integral to a product space, resulting in the *double integral*,

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) d\nu_1 \times d\nu_2$$

is the same as the *iterated integral*,

$$\int_{\Omega_2} \left(\int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1 \right) d\nu_2.$$

The Integral over a Domain

For $A \subset \Omega$, we define

$$\int_A f \, d\nu = \int \mathbf{I}_A f \, d\nu$$

If f is a nonnegative Borel function, then $\int_A f \, d\nu$ for $A \subset \Omega$ is a measure over Ω .

[Probability of an Event

Given a probability space (Ω, \mathcal{F}, P) , a set $A \in \mathcal{F}$ is called an “event”, and $\int_A dP$ is the *probability of A*, written $P(A)$ or $\text{Pr}(A)$.

Expected Value

Given a probability space (Ω, \mathcal{F}, P) and a random variable with respect to \mathcal{F} , X , we define the *expected value* of X with respect to P as

$$\int X \, dP,$$

and denote it as $E(X)$ or for clarity, $E_P(X)$.

Sometimes we limit this definition to integrable random variables X .

Conditional Expectation over a Sub- σ -Field

Let (Ω, \mathcal{F}, P) be a probability space, let \mathcal{A} be a sub- σ -field of \mathcal{F} , and let X be an integrable random variable over Ω .

The *conditional expectation* of X given \mathcal{A} , denoted by $E(X|\mathcal{A})$ is a random variable such that $E(X|\mathcal{A})$ is a measurable function from (Ω, \mathcal{F}) to $(\mathbb{R}, \mathcal{B})$ and $\int_A E(X|\mathcal{A}) dP = \int_A X dP$ for any $A \in \mathcal{A}$. (The existence and uniqueness of this random variable follows from the Radon-Nikodym theorem (Theorem 1.4)).

Conditional Probability

Let (Ω, \mathcal{F}, P) be a probability space, let \mathcal{A} be a sub- σ -field of \mathcal{F} , and let $B \in \mathcal{F}$. The *conditional probability* of B given \mathcal{A} , denoted by $\Pr(B|\mathcal{A})$ is defined as $E(I_B|\mathcal{A})$.

Conditional Expectation with Respect to Another Measurable Function

Let (Ω, \mathcal{F}, P) be a probability space, let \mathcal{A} be a sub- σ -field of \mathcal{F} , let X be an integrable random variable over Ω , and let Y be a measurable function from (Ω, \mathcal{F}, P) to any measurable space (Λ, \mathcal{G}) . Then the *conditional expectation* of X given Y , denoted by $E(X|Y)$ is defined as the conditional expectation of X given the sub- σ -field generated by Y , that is, $E(X|\sigma(Y))$.

Continuity

Given two measures ν and λ on the same measurable space, (Ω, \mathcal{F}) , if $\forall A \in \mathcal{F}$

$$\nu(A) = 0 \quad \Rightarrow \quad \lambda(A) = 0,$$

then λ is said to be *absolutely continuous* with respect to ν . We denote that λ is said to be *absolutely continuous* with respect to ν by

$$\lambda \ll \nu.$$

As an example, let $(\Omega, \mathcal{F}, \nu)$ be a measure space and let f be a nonnegative Borel function on Ω . Define the measure λ by

$$\lambda(A) = \int_A f \, d\nu$$

for any $A \in \mathcal{F}$. Then $\nu(A) = 0 \Rightarrow \lambda(A) = 0$, and so λ is absolutely continuous with respect to ν .

Radon-Nikodym Theorem

Given two measures ν and λ on the same measurable space, (Ω, \mathcal{F}) , such that $\lambda \ll \nu$ and ν is σ -finite. Then there exists a unique a.e. nonnegative Borel function f on Ω such that

$$\lambda(A) = \int_A f d\nu \quad \forall A \in \mathcal{F}.$$

Uniqueness a.e. means that if also, for some g ,

$$\lambda(A) = \int_A g d\nu \quad \forall A \in \mathcal{F} \text{ then } f = g \text{ a.e.}$$

The proof of the Radon-Nikodym theorem is complicated.

Radon-Nikodym Derivative

If $\lambda(A) = \int_A f d\nu \forall A \in \mathcal{F}$, then f is called the Radon-Nikodym derivative of λ with respect to ν , and we write $f = d\lambda/d\nu$.

Some Important Probability Facts

A probability distribution is built from a σ -field, \mathcal{F} , defined on a sample space Ω and a σ -finite probability measure P . In practice, the probability measure is usually based on the counting measure (defined on countable sets as their cardinality), or on the Lebesgue measure (the length of intervals).

A random variable is a measurable function, $X(\omega)$ or just X , defined on Ω . We will assume X is real. (When X is a vector, we will sometimes emphasize this by writing \underline{X} – but not always!)

For a given random variable X , a probability distribution determines $\Pr(X \in A)$ for $A \in \mathcal{A}$.

A probability family or family of distributions, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, is a set of probability distributions of a random variable that is defined over Ω . We will assume θ is real. (When θ is a vector, we will sometimes emphasize this by writing $\underline{\theta}$ – but not always!)

For a given random variable X , the function $F(x) = \Pr(X \leq x)$ is called the cumulative distribution function, CDF, of X . If the random variable is assumed to be in a family of distributions indexed by θ , we may use the notation $F_\theta(x)$ or $F(x; \theta)$.

The derivative of a CDF with respect to the appropriate measure is called the probability density function, PDF. (We use this term for either discrete random variables and a counting measure or continuous random variables and Lebesgue measure. We can also think of densities in the context of a continuous random variable with positive mass points.)

Expected Value

The expected value of a measurable function g of a random variable X with CDF F is defined as $E(g(X)) = \int g(x)dF(x)$.

An Interesting Relationship

For a positive random variable X ,

$$E(X) = \int_0^{\infty} \Pr(X > t) dt.$$

This is a simple application of Fubini's theorem:

$$\begin{aligned} E(X) &= \int_0^{\infty} x dF(x) \\ &= \int_0^{\infty} \int_{(0,x)} dt dF(x) \\ &= \int_0^{\infty} \int_{(t,\infty)} dF(x) dt \\ &= \int_0^{\infty} (1 - F(t)) dt \\ &= \int_0^{\infty} \Pr(X > t) dt \end{aligned}$$

This leads in general to the fact that for any given random variable X , if $E(X)$ exists, then

$$E(X) = \int_0^{\infty} (1 - F(t)) dt - \int_{-\infty}^0 F(t) dt.$$

Independence

Marginal distribution equals conditional distribution (factor the joint PDF or CDF).

$$\forall A, \Pr(X \in A|Y) = \Pr(X \in A).$$

Transformations of Random Variables

If \underline{X} has density $p_{\underline{X}}(\underline{x}|\theta)$ and $\underline{Y} = h(\underline{X})$, where h is a full-rank transformation, that is there is a function h^{-1} such that $\underline{X} = h^{-1}(\underline{Y})$, then the density of \underline{Y} is $p_{\underline{Y}}(h^{-1}(\underline{y})|\theta)|J_{h^{-1}}(\underline{y})|$, where $J_{h^{-1}}(\underline{y})$ is the Jacobian of the inverse transformation, and $|\cdot|$ is the determinant.

Why the inverse transformation? Think of the density as a differential; that is, it has a factor $d\underline{x}$, so in the density for \underline{Y} , we want a factor $d\underline{y}$. Under pressure you may forget exactly how this goes, or want a quick confirmation of the transformation. You should be able to construct a simple example quickly. An easy one is the right-triangular distribution; that is, the distribution with density $p_X(x) = 2x$, for $0 < x < 1$. Let $y = 2x$, so $x = \frac{1}{2}y$. Sketch the density of Y , and think of what transformations are necessary to get the expression $p_Y(y) = \frac{1}{2}y$, for $0 < y < 2$.

Order Statistics

The joint density of the i^{th} and j^{th} ($i < j$) order statistics is

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} (F(x_{(i)}))^{i-1} (F(x_{(j)}) - F(x_{(i)}))^{j-i-1} (1 - F(x_{(j)}))^{n-j} f(x_{(i)}) f(x_{(j)})$$

Understand the heuristics that lead to this formula.

Order statistics are not i.i.d.

Moment-Generating Functions and Characteristic Functions

The moment-generating function (MGF) and characteristic function (CF) are *transforms* of the density function. The moment-generating function for the random variable X , which may not be well-defined (in which case we say it does not exist), is

$$\psi_X(t) = E(e^{tX}),$$

and the characteristic function, which is always well-defined is

$$\phi_X(t) = E(e^{itX}).$$

MGF and CF

Both functions are nonnegative. If the MGF is finite for some $t \neq 0$, the CF can be obtained by replacing t in $\psi_X(t)$ by it (where $i = \sqrt{-1}$).

The characteristic function is the Fourier transform of the density with argument $t/(2\pi)$.

Both transforms are defined for a vector-valued random variable \underline{X} similarly, and the corresponding transforms are functions of a vector-valued variable \underline{t} .

The expression tX in the definitions above is replaced by $\underline{t}^T \underline{X}$.

MGF and CF

An interesting property of the MGF and the CF is that the moments of X can be obtained from their derivatives; for example,

$$\left. \frac{d^k \phi_X(t)}{dt^k} \right|_{t=0} = (-1)^{k/2} \mathbb{E}(X^k).$$

For vector-valued random variables, the moments become tensors, but the first two are very simple: $\nabla \phi_{\underline{X}}(t)|_{t=0} = \mathbb{E}(\underline{X})$ and $\nabla \nabla \phi_{\underline{X}}(t)|_{t=0} = \mathbb{E}(\underline{X}^T \underline{X})$.

The CF or MGF completely determines the distribution. Also, the limit of a sequence of CFs or MGFs determines the limiting distribution.

MGF and CF

A nice use of CFs (or MGFs, if we are willing to assume that the MGF exists) is in the proof of a simple form of the central limit theorem that states that if X_1, \dots, X_n are i.i.d. with mean μ and variance $0 < \sigma^2 < \infty$, then $Y_n = (\sum X_i - n\mu)/\sqrt{n}\sigma$ has limiting distribution $N(0, 1)$.

Proof: It will be convenient first to define a function related to the CF.

Let $h(t) = e^{\mu t} \phi_X(t)$; hence $h(0) = 1$, $h'(0) = 0$, and $h''(0) = \sigma^2$.

Now expand h in a Taylor series about 0:

$$h(t) = h(0) + h'(0)t - \frac{1}{2}h''(\xi)t^2,$$

for some ξ between 0 and t .

Substituting for $h(0)$ and $h'(0)$, and adding and subtracting $\sigma^2 t/2$ to this, we have

$$h(t) = 1 - \frac{\sigma^2 t^2}{2} - \frac{(h''(\xi) - \sigma^2)t^2}{2}.$$

This is the form we will find useful.

Now, consider the CF of Y_n :

$$\begin{aligned}\phi_{Y_n}(t) &= \mathbb{E} \left(\exp \left(it \left(\frac{\sum X_i - n\mu}{\sqrt{n}\sigma} \right) \right) \right) \\ &= \left(\mathbb{E} \left(\exp \left(it \left(\frac{X - \mu}{\sqrt{n}\sigma} \right) \right) \right) \right)^n \\ &= \left(h \left(\frac{it}{\sqrt{n}\sigma} \right) \right)^n.\end{aligned}$$

From the expansion of h , we have

$$h \left(\frac{it}{\sqrt{n}\sigma} \right) = 1 - \frac{t^2}{2n} - \frac{(h''(\xi) - \sigma^2)t^2}{2n\sigma^2}.$$

So,

$$\phi_{Y_n}(t) = \left(1 - \frac{t^2}{2n} - \frac{(h''(\xi) - \sigma^2)t^2}{2n\sigma^2} \right)^n.$$

Now we need a well-known (but maybe forgotten) result: If $\lim_{n \rightarrow \infty} f(n) = 0$, then

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a}{n} + \frac{f(n)}{n} \right)^n = e^{ab}.$$

Therefore, because $\lim_{n \rightarrow \infty} h''(\xi) = h''(0) = \sigma^2$, $\lim_{n \rightarrow \infty} \phi_{Y_n}(t) = e^{-t^2/2}$, which is the CF of the $N(0, 1)$ distribution.

This completes the proof.

This simple form of the CLT together with its proof is an *easy piece* that you should be able to prove relatively quickly.

Conditional Expectations

The concept of conditional expectation is one of the most important in statistics.

It is the basis for the analysis of relationships among variables.

The definition of conditional expectation of one random variable given another random variable is developed in two stages.

First, we define conditional expectation over a sub- σ -field.

Then we define conditional expectation with respect to another measurable function (a random variable, for example) in terms of the conditional expectation over the sub- σ -field generated by the inverse image of the function.

Conditional Expectations

A major difference in conditional expectations and unconditional expectations is that conditional expectations are random variables.

Relations involving conditional expectations, therefore, must be qualified by such conditions with probability 1.

Two Important Conditional Expectations

There are some conditional expectations that arise often, and which we should immediately recognize. These equalities are a.e.

- $E(E(Y|X)) = E(Y)$.

The student should realize that the expectation operator is based on a probability distribution, and so anytime we see “E”, we need to ask “with respect to what probability distribution?” In notation such as that above, the distribution is implicit. The inner expectation on the left is with respect to the conditional distribution of X given Y , and so is a function of Y . The outer expectation is with respect to the marginal distribution of Y .

- $V(Y) = V(E(Y|X)) + E(V(Y|X))$.

This is intuitive, although you should be able to prove it formally. The intuitive explanation is: the total variation in Y is the sum of the variation of its mean given X and its average variation about X (or given X). (Think of $SST = SSR + SSE$ in regression analysis.)

This equality implies the Rao-Blackwell inequality (drop the second term on the right).

Inequalities

Inequalities involving functions of events and random variables are important throughout the field of probability and statistics.

Two important uses are for showing that one procedure is better than another and for showing that some sequence converges to a given object (a constant, a function, or a set).

Important Types of Inequalities

Four important types of inequalities involve relations between

- $\Pr(X \in A)$ and $E(f(X))$, e.g., Chebyshev
- $E(f(X))$ and $f(E(X))$, e.g., Jensen's
- $E(f_1(X, Y))$ and $E(f_2(X))$ and $E(f_3(Y))$, e.g., covariance, Cauchy-Schwarz, information
- $V(Y)$ and $V(E(Y|X))$, e.g., Rao-Blackwell

Any special case of these involves an appropriate definition of A or f (e.g., nonnegative, convex, etc.)

A more general case of the inequalities is to replace distributions, and hence expected values, by conditioning on a sub- σ -field, \mathcal{A} .

For each type of inequality there is essentially a straightforward method of proof, which is important to know.

Some of these inequalities involve absolute values of the random variable.

To work with these inequalities, it is useful to recall the triangle inequality for the absolute value of real numbers:

$$|x + y| \leq |x| + |y|.$$

We can prove this merely by considering all four cases for the signs of x and y .

This inequality generalizes immediately to $|\sum x_i| \leq \sum |x_i|$.

Expectations of absolute values of functions of random variables are norms. (A *norm* is a function of x that (1) is positive unless $x = 0$ a.e., that (2) is equivariant to scalar multiplication, and that (3) satisfies the triangle inequality.) Some of the inequalities given below involving expectations of absolute values of random variables are essentially triangle inequalities and their truth establishes the expectation as a norm.

Some of the expectations I'll discuss are recognizable as familiar norms over vector spaces.

For example, the expectation in Minkowski's inequality is essentially the L_p norm of a vector, which is defined for an n -vector x in a finite-dimensional vector space as

$$\|x\|_p \equiv (\sum |x_i|^p)^{1/p}.$$

Minkowski's inequality in this case is $\|x + y\|_p \leq \|x\|_p + \|y\|_p$. For $p = 1$, this is the triangle inequality for absolute values given above.

- $\Pr(X \in A)$ and $E(f(X))$

The important general form is Markov's inequality. Several others are special cases of it.

– **Markov's inequality**

For $\epsilon > 0$, $k > 0$, and r.v. $X \ni E(|X|^k)$ exists,

$$\Pr(|X| \geq \epsilon) \leq \frac{1}{\epsilon^k} E(|X|^k)$$

Proof: For nonnegative r.v. Y ,

$E(Y) \geq \int_{y \geq \epsilon} y dP(y) \geq \epsilon \int_{y \geq \epsilon} dP(y) = \epsilon \Pr(Y \geq \epsilon)$. Now let $Y = |X|^k$.

– **Chebyshev's inequality**

$$\Pr(|X - E(X)| \geq \epsilon) \leq \frac{1}{\epsilon^2} V(X)$$

Proof: In Markov's inequality, let $k = 2$, and replace X by $X - E(X)$.

– **Chebyshev's inequality (another form)**

For $g \ni g(x) \geq 0$,

$$\Pr(g(X) \geq \epsilon) \leq \frac{1}{\epsilon} E(g(X))$$

Proof: Same as Markov's inequality; start with $E(g(X))$.

• $E(f(X))$ and $f(E(X))$

– **Jensen's inequality**

For g convex over support of r.v. X (and all expectations shown exist),

$$g(E(X)) \leq E(g(X)).$$

Proof: By the definition of convexity, g convex

$\Rightarrow \exists c \ni \forall t$ in domain of convexity,

$c(x - t) + g(t) \leq g(x)$. (Notice that $L(x) = c(x - t) + g(t)$ is a straight line through the point $(t, g(t))$. By the

definition of convexity, g is convex if its value at the weighted average of two points does not exceed the weighted average of the function at those two points.) Now, given this, let $t = E(X)$ and take expectations of both sides of the inequality.

If g is strictly convex, $g(E(X)) < E(g(X))$ unless $g(X) = E(g(X))$ with probability 1.

For a concave function, the inequality is reversed. (The negative of a concave function is convex.)

Some simple examples for a nonconstant positive random variable X :

* Reciprocals

$$\frac{1}{E(X)} < E\left(\frac{1}{X}\right)$$

* Logs

$$E(\log(X)) < \log(E(X)).$$

Some other consequences:

- * Entropy distance (Kullback-Leibler information): If f and g are probability densities, $E_f(\log(f(X)/g(X)))$, is the *entropy distance* between f and g with respect to g . It is also called the *Kullback-Leibler information* or *Kullback-Leibler distance*. It is nonnegative:

$$E_f(\log(f(X)/g(X))) \geq 0.$$

Proof:

$$\begin{aligned} E_f(\log(f(X)/g(X))) &= -E_f(\log(g(X)/f(X))) \\ &\geq -\log(E_f(g(X)/f(X))) \\ &= 0. \end{aligned}$$

A related fact applies to any positive integrable functions f and g on a measure space with a σ -finite measure ν , for which $\int f d\nu \geq \int g d\nu > 0$:

$$\int f(\log(f/g)) d\nu \geq 0.$$

This can be proved as above by forming densities.

- * An inequality important in showing the convergence of the EM algorithm:

$$E_f(\log(f(X))) \geq E_f(\log(g(X))).$$

We see this by use of the entropy distance.

- $E(f_1(X, Y))$ and $E(f_2(X))$ and $E(f_3(Y))$

The important form $(E(|X|^p))^{1/p}$, for $1 \leq p$ is an L_p norm, $\|X\|_p$. Many of the inequalities in this section hold for general L_p norms. For example, Minkowski's inequality is just the triangle inequality for an L_p norm:

$\|f + g\|_p \leq \|f\|_p + \|g\|_p$. (The fact that this inequality holds makes $(E(|\cdot|^p))^{1/p}$, or $\|\cdot\|_p$, a norm.)

An important general form is Hölder's inequality:

$$\|fg\|_1 \leq \|f\|_p \|g\|_p.$$

Several other inequalities are special cases of it.

- **Hölder's inequality** (a general inequality relating $E(f_1(X, Y))$ to $E(f_2(X))$ and $E(f_3(Y))$)

For $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$ (and all expectations shown exist),

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

p and q as in this inequality are called *dual* indices. Note that $q = p/(p - 1)$.

Proof: If $E(|X|^p) = 0$ or $E(|Y|^q) = 0$, then true because both sides = 0 wp1. Hence, assume both > 0 .

Now, for p and q as in hypothesis,

$\forall a, b > 0, \exists s, t \ni a = e^{s/p}$ and $b = e^{t/q}$. Now e^x is convex, so $e^{s/p+t/q} \leq \frac{1}{p}e^s + \frac{1}{q}e^t$, or $ab \leq a^p/p + b^q/q$.

Now let

$$a = \left| \frac{X(\omega)}{(E(|X|^p))^{1/p}} \right| \quad \text{and} \quad b = \left| \frac{Y(\omega)}{(E(|Y|^q))^{1/q}} \right|$$

and so

$$|X(\omega)Y(\omega)| \leq (\mathbb{E}(|X|^p))^{1/p} (\mathbb{E}(|Y|^q))^{1/q} \left(\frac{|X(\omega)|^p}{\mathbb{E}(|X|^p)} \frac{1}{p} + \frac{|Y(\omega)|^q}{\mathbb{E}(|Y|^q)} \frac{1}{q} \right).$$

Take expectations. (The notation $X(\omega), Y(\omega)$ is meant to emphasize how to take expectation of XY .)

We note a special case by letting $Y \equiv 1$:

$$\mathbb{E}(|X|) \leq (\mathbb{E}(|X|^p))^{1/p},$$

and with $p = 2$, we have a special case of the Cauchy-Schwarz inequality:

$$\mathbb{E}(|X|) \leq (\mathbb{E}(X^2))^{1/2}.$$

– Liapounov's inequality

If $1 \leq r \leq s$

$$(\mathbb{E}(|X|^r))^{1/r} \leq (\mathbb{E}(|X|^s))^{1/s}$$

Proof: First, we observe this is true for $r = s$, and for $r = 1$. If $1 < r < s$, replace $|X|$ in the special case of

Hölder's inequality above with $|X|^r$, and let $s = pr$ for $1 < p$. This yields $(\mathbb{E}(|X|^r))^{1/r} \leq (\mathbb{E}(|X|^s))^{1/s}$.

– **Minkowski's inequality (triangle inequality for L_p norms and related functions)**

For $1 \leq p$,

$$(\mathbb{E}(|X + Y|^p))^{1/p} \leq (\mathbb{E}(|X|^p))^{1/p} + (\mathbb{E}(|Y|^p))^{1/p}$$

Proof: First, observe the truth for $p = 1$ using the triangle inequality for the absolute value,

$$|x + y| \leq |x| + |y|, \text{ giving } \mathbb{E}(|X + Y|) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|).$$

Now assume $p > 1$. Now,

$$\begin{aligned} \mathbb{E}(|X + Y|^p) &= \mathbb{E}(|X + Y||X + Y|^{p-1}) \\ &\leq \mathbb{E}(|X||X + Y|^{p-1}) + \mathbb{E}(|Y||X + Y|^{p-1}), \end{aligned}$$

where the inequality comes from the triangle inequality for absolute values. From Hölder's inequality on the terms above with $q = p/(p - 1)$, we have

$$\mathbb{E}(|X + Y|^p) \leq (\mathbb{E}(|X|^p))^{1/p} (\mathbb{E}(|X + Y|^p))^{1/q} + (\mathbb{E}(|Y|^p))^{1/p} (\mathbb{E}(|X + Y|^p))^{1/q}$$

Now, if $E(|X + Y|^p) = 0$, Minkowski's inequality holds. If $E(|X + Y|^p) \neq 0$, divide through by $(E(|X + Y|^p))^{1/q}$, recalling again that $q = p/(p - 1)$.

– **Other inequalities similar to the triangle inequality**

* For $0 \leq p$,

$$|X + Y|^p \leq 2^p(|X|^p + |Y|^p)$$

This is true because $\forall \omega \in \Omega$,

$\|X(\omega) + Y(\omega)\| \leq 2 \max\{\|X(\omega)\|, \|Y(\omega)\|\}$, and so

$$\begin{aligned} \|X(\omega) + Y(\omega)\|^p &\leq \max\{2^p\|X(\omega)\|^p, 2^p\|Y(\omega)\|^p\} \\ &\leq 2^p\|X(\omega)\|^p + 2^p\|Y(\omega)\|^p. \end{aligned}$$

– **Schwarz inequality, or Cauchy-Schwarz inequality**

$$E(|XY|) \leq (E(X^2)E(Y^2))^{1/2}$$

Proof: Let $p = q = 2$ in Hölder's inequality.

Another proof: For nonnegative r.v. X and Y and all t (real),

$$E((tX + Y)^2) = t^2E(X^2) + 2tE(XY) + E(Y^2) \geq 0.$$

Hence the discriminant of the quadratic formula ≤ 0 .

Now, for any r.v., take absolute value.

– **Covariance inequality**

If the second moments of X and Y are finite, then

$$(E((X - E(X))(Y - E(Y))))^2 \leq E((X - E(X))^2) E((Y - E(Y))^2)$$

or

$$(\text{Cov}(X, Y))^2 \leq V(X) V(Y)$$

– **Information inequality**

Subject to some “regularity conditions”, if X had PDF $p(x; \theta)$,

$$V(f(X)) \geq \frac{\left(\frac{\partial E(f(X))}{\partial \theta}\right)^2}{E_{\theta} \left(\left(\frac{\partial \log p(X; \theta)}{\partial \theta}\right)^2 \right)}$$

- $V(Y)$ and $V(E(Y|X))$

- **Rao-Blackwell inequality**

$$V(E(Y|X)) \leq V(Y)$$

Recall the equality above,

$$V(Y) = V(E(Y|X)) + E(V(Y|X)).$$

There are multivariate extensions of most of these inequalities. The ones involving simple inequalities are extended by conditions on vector norms, and the ones involving variances are usually extended by positive (or semi-positive) definiteness of the difference of two variance-covariance matrices.

Conditional Distributions

The concept of conditional distributions provides the basis for the analysis of relationships among variables.

A simple way of developing the ideas begins by defining the conditional probability of event A , given event B . If $\Pr(B) \neq 0$, the standard definition in the standard notation is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)},$$

which leads to the useful multiplication rule

$$\Pr(A \cap B) = \Pr(B)\Pr(A|B).$$

If we interpret all of this in the context of the probability space (Ω, \mathcal{F}, P) , we can define a new “conditioned” probability space, $(\Omega, \mathcal{F}, P_B)$, where we define P_B by

$$P_B(A) = \Pr(A \cap B),$$

for any $A \in \mathcal{F}$.

This approach, however, is not entirely satisfactory because of the requirement that $\Pr(B) \neq 0$.

Another approach is to make use of a concept of conditional expectation.

The definition of conditional expectation of one random variable given another random variable is developed in two stages.

First, we define conditional expectation over a sub- σ -field, and then we define conditional expectation with respect to another measurable function (a random variable, for example) in terms of the conditional expectation over the sub- σ -field generated by the inverse image of the function.

A major difference in conditional expectations and unconditional expectations is that conditional expectations may be random variables. When the expectation is conditioned on a random variable, relations involving the conditional expectations must be qualified by such conditions with probability 1.

Conditional expectation over a sub- σ -field

Let (Ω, \mathcal{F}, P) be a probability space, let \mathcal{A} be a sub- σ -field of \mathcal{F} , and let X be an integrable random variable over Ω . The *conditional expectation* of X given \mathcal{A} , denoted by $E(X|\mathcal{A})$ is a random variable such that $E(X|\mathcal{A})$ is a measurable function from (Ω, \mathcal{F}) to $(\mathbb{R}, \mathcal{B})$ and $\int_A E(X|\mathcal{A}) dP = \int_A X dP$ for any $A \in \mathcal{A}$. (The existence and uniqueness of this random variable follows from the Radon-Nikodym theorem (Theorem 1.4)).

Conditional expectation with respect to another measurable function

Let (Ω, \mathcal{F}, P) be a probability space, let \mathcal{A} be a sub- σ -field of \mathcal{F} , let X be an integrable random variable over Ω , and let Y be a measurable function from (Ω, \mathcal{F}, P) to any measurable space (Λ, \mathcal{G}) . Then the *conditional expectation* of X given Y , denoted by $E(X|Y)$ is defined as the conditional expectation of X given the sub- σ -field generated by Y , that is, $E(X|\sigma(Y))$.

Sub- σ -fields generated by random variables, such as $\sigma(Y)$, play an important role in statistics. We can think of $\sigma(Y)$ as being the “information provided by Y ”. In an important type of time series, Y_1, Y_2, \dots , we encounter a sequence $\sigma(Y_1) \subset \sigma(Y_2) \subset \dots$.

Some common conditional expectations

There are some conditional expectations that arise often, and which we should immediately recognize. These equalities are a.e.

- $E(E(Y|X)) = E(Y)$.

The student should realize that the expectation operator is based on a probability distribution, and so anytime we see “E”, we need to ask “with respect to what probability distribution?” In notation such as that above, the distribution is implicit. The inner expectation on the left is with respect to the conditional distribution of X given Y , and so is a function of Y . The outer expectation is with respect to the marginal distribution of Y .

- $V(Y) = V(E(Y|X)) + E(V(Y|X))$.

This is intuitive, although you should be able to prove it formally. The intuitive explanation is: the total variation in Y is the sum of the variation of its mean given X and its average variation about X (or given X). (Think of $SST = SSR + SSE$ in regression analysis.)

This equality implies the Rao-Blackwell inequality (drop the second term on the right).

Conditional probability distributions

We can now develop important concepts of joint and conditional probability distributions in terms of conditional expectations.

- Conditional probability.

Let (Ω, \mathcal{F}, P) be a probability space, let \mathcal{A} be a sub- σ -field of \mathcal{F} , and let $B \in \mathcal{F}$. The *conditional probability* of B given \mathcal{A} , denoted by $\Pr(B|\mathcal{A})$ is defined as $E(I_B|\mathcal{A})$.

- Independence.

We define *independence* in a probability space (Ω, \mathcal{F}, P) in **three steps:**

1. Independence of events within a collection of events.

Let \mathcal{C} be a collection of events; that is, a collection of subsets of \mathcal{F} . The events in \mathcal{C} are *independent* iff for any positive integer n and distinct events A_1, \dots, A_n in \mathcal{C} ,

$$P(A_1 \cap \dots \cap A_n) = P(A_1) \cdots P(A_n).$$

We can have the situation in which all pairs within the collection are independent, but the collection is not independent; for example, in an experiment of tossing a coin twice, let

A be “heads on the first toss”

B be “heads on the second toss”

C be “exactly one head and one tail on the two tosses” We see immediately that any pair is independent, but that the intersection is \emptyset .

Sometimes people use the phrase “mutually independent” to try to emphasize that we are referring to independence of *all* events. The phrase “mutually independent”, however, could be interpreted as “pairwise independent”, so it really does not clarify anything.

2. Independence of collections of events (and, hence, of σ -fields)

For any index set \mathcal{I} , let \mathcal{C}_i be a collection of sets with $\mathcal{C}_i \subset \mathcal{F}$. The collections \mathcal{C}_i are independent iff the events in any collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are independent.

3. Independence of functions (and, hence, of random variables).

(This also defines independence of any generators of σ -fields.)

The elements X_i for $i \in \mathcal{I}$ are independent iff $\sigma(X_i)$ for $i \in \mathcal{I}$ are independent.

Conditional probability distributions

For distributions with PDFs we can define conditional distributions very simply. The concept of a joint distribution with a PDF comes from the Radon-Nikodym derivative of a CDF over a product space. This is the familiar

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)},$$

which may be taken as a definition so long as $f_Y(y) > 0$. With this, we can work back to define a conditional expectation in agreement with that above.

Conditional probability distributions

Another approach for defining joint and conditional distributions is given in Shao's "Theorem 1.7". In this we start with a probability space $(\mathbb{R}^m, \mathcal{B}^m, P_1)$ and define a probability measure on the measurable space $(\mathbb{R}^n \times \mathbb{R}^m, \sigma(\mathcal{B}^n \times \mathcal{B}^m))$. The existence of such a probability measure is given in the first part of this multi-theorem (which is proved in Billingsley).

For a random variable Y in \mathbb{R}^m , its (marginal) distribution is determined by P_1 , which we denote as $P_Y(y)$. For $B \in \mathcal{B}^n$ and $C \in \mathcal{B}^m$, the conditional distribution is defined by identifying a probability measure, denoted as $P_{X|Y}(\cdot|y)$, on $(\mathbb{R}^n, \sigma(\mathcal{B}^n))$ for any fixed $y \in \mathbb{R}^m$.

The joint probability measure of (X, Y) over $\mathbb{R}^n \times \mathbb{R}^m$ is defined as

$$P_{XY} = \int_C P_{X|Y}(\cdot|y) dP_Y(y),$$

where $C \in \mathcal{B}^m$.

Some properties of conditional expectations

We have the simple relationship with the unconditional expectation:

$$E(E(X|\mathcal{A})) = E(X).$$

We can establish conditional versions of the three theorems that relate to the interchange of an integration operation and a limit operation, monotone convergence, Fatou's lemma, and dominated convergence. These extensions are fairly straightforward.

$$X \leq Y \text{ a.s.} \quad \Rightarrow \quad E(X|\mathcal{A}) \leq E(Y|\mathcal{A}) \text{ a.s.}$$

and

$$E(\liminf_n X_n|\mathcal{A}) \leq \liminf_n E(X_n|\mathcal{A}) \text{ a.s.}$$

We also have the linearity of the conditional expectation operator:

$$\text{for } a \in \mathbb{R}, E(aX + Y|\mathcal{A}) = aE(X|\mathcal{A}) + E(Y|\mathcal{A}) \text{ a.s.}$$

Conditional expectation can be viewed as a projection in a linear space defined by the square-integrable random variables over a given probability space and the inner product $\langle X, Y \rangle = E(XY)$ and its induced norm. (Ω, \mathcal{F}, P) be a probability space, let \mathcal{A} be a sub- σ -field of \mathcal{F} , and let $B \in \mathcal{F}$. The *conditional probability* of B given \mathcal{A} , denoted by $\Pr(B|\mathcal{A})$ is defined as $E(I_B|\mathcal{A})$

Martingales

The concept of conditional expectation is important in developing a theory of martingales. These are special sequences of random variables that have applications in various processes that evolve over time.

As a concrete example, we can think of a random variable X_1 as an initial sum (say, of money), and a sequence of events in which X_2, X_3, \dots represents a sequence of sums with the property that each event is a “fair game”; that is, $E(X_2|X_1) = X_1$ a.s., $E(X_3|X_1, X_2) = X_2$ a.s., \dots . We can generalize this somewhat by letting $\mathcal{D}_n = \sigma(X_1, \dots, X_n)$, and requiring that the sequence be such that $E(X_n|\mathcal{D}_{n-1}) = X_{n-1}$ a.s.

We say that $\{X_t : t \in T\}$ is a *martingale* relative to $\{\mathcal{D}_t : t \in T\}$ in some probability space (Ω, \mathcal{F}, P) , if $X_s = E(X_t|\mathcal{D}_t)$ for s, t .

Martingales

An alternate definition is in terms of the pairs (X_t, \mathcal{F}_t) ; that is, the definition is for the random variable and an associated σ -field, rather than the random variable relative to some sequence of σ -fields.

We say the sequence $\{(X_t, \mathcal{F}_t) : t \in T\}$, where $\mathcal{F}_t \subset \mathcal{F}_{t+1} \subset \dots$, is a *martingale* if $E(X_n | \mathcal{F}_{n-1}) = X_{n-1}$ a.s.

We say that $\{X_t : t \in T\}$ is a *submartingale* relative to $\{\mathcal{D}_t : t \in T\}$ if $X_s \leq E(X_t | \mathcal{D}_t)$ for s, t .

A common application of martingales is as a model for stock prices.

Asymptotic Properties

Countably infinite sequences play the main role in the definition of the basic concept of a σ -field, and consequently, in the development of a theory of probability.

Sequences of sets correspond to sequences of events and, consequently, of sequences of random variables. Unions, intersections, and complements of sequences of sets are important for studying sequences of random variables. Important relationships between unions, intersections, and complements are

$$(A \cup B)^c = A^c \cap B^c \quad \text{and} \quad (A \cap B)^c = A^c \cup B^c$$

This is how we get the important fact for a σ -field \mathcal{F} that if $A_1, A_2, \dots \in \mathcal{F}$ then $\bigcap_i A_i \in \mathcal{F}$.

Often we want to show that two sets or collections are equal.

Two important types of unions and intersections of sequences of sets

$$\limsup_n A_n \equiv \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i$$

and

$$\liminf_n A_n \equiv \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i.$$

We often write

$$A^* = \limsup_n A_n$$

and

$$A_* = \liminf_n A_n.$$

We also use the notation $\limsup_n \Pr(A_n)$ and $\liminf_n \Pr(A_n)$; that is, for sequences of real numbers, or values of real-valued functions.

These are defined as:

$$\limsup_n \Pr(A_n) \equiv \inf_n \sup_{i \geq n} \Pr(A_i)$$

$$\liminf_n \Pr(A_n) \equiv \sup_n \inf_{i \geq n} \Pr(A_i).$$

\limsup_n is often written as $\overline{\lim}_n$

\liminf_n is often written as $\underline{\lim}_n$

We can interpret A^* and A_* in an intuitive fashion.

An element ω is in A^* iff for each n , there is some $i \geq n$ for which $\omega \in A_i$. This means that ω must lie in infinitely many of the A_n .

An element ω is in A_* iff there is some n such that for all $i \geq n$, $\omega \in A_i$. This means that ω must lie in all but finitely many of the A_n .

Convergence of sequences of sets

We define convergence of a sequence of sets in terms of lim sups and lim infs.

The sequence of sets $\{A_n\}$ is said to *converge* if $\limsup_n A_n = \liminf_n A_n$, and this set is said to be the *limit* of the sequence.

A sequence of sets $\{A_n\}$ is said to be *increasing* if $A_n \subset A_{n+1}$ for all n , and is said to be *decreasing* if $A_{n+1} \subset A_n$ for all n . In either case, the sequence is said to be *monotone*.

An increasing sequence $\{A_n\}$ converges to $\cup_{n=1}^{\infty} A_n$.

A decreasing sequence $\{A_n\}$ converges to $\cap_{n=1}^{\infty} A_n$.

Some basic facts about lim sups and lim infs

First, we note that for a σ -field \mathcal{F} if $A_1, A_2, \dots \in \mathcal{F}$ then $\limsup_n A_n \in \mathcal{F}$ and $\liminf_n A_n \in \mathcal{F}$. This is because for any n , $\bigcup_{i=n}^{\infty} A_i \in \mathcal{F}$ and $\bigcap_{i=n}^{\infty} A_i \in \mathcal{F}$.

Two simple relationships that follow immediately from the definitions:

$$\limsup_n A_n \subset \bigcup_{i=n}^{\infty} A_i$$

and

$$\bigcap_{i=n}^{\infty} A_i \subset \liminf_n A_n.$$

We also have the relationships:

$$\Pr(\limsup_n A_n) \leq \limsup_n \Pr(A_n)$$

and

$$\Pr(\liminf_n A_n) \leq \liminf_n \Pr(A_n)$$

We see this by considering $B_n = \cup_{i=n}^{\infty} A_i$, so that $B_n \searrow \limsup_n A_n$, and likewise $C_n = \cap_{i=n}^{\infty} A_i$, so that $C_n \nearrow \liminf_n A_n$.

We use the continuity of the measure to get

$$\Pr(A_n) \leq \Pr(B_n) \rightarrow \Pr(\limsup_n A_n) \text{ and}$$

$$\Pr(A_n) \geq \Pr(C_n) \rightarrow \Pr(\liminf_n A_n)$$

Another important fact is

$$\liminf_n A_n \subset \limsup_n A_n.$$

To see this, consider any $\omega \in \liminf_n A_n$:

$$\omega \in \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i \iff \exists n \text{ such that } \forall i \geq n, \omega \in A_i,$$

so $\omega \in \limsup_n A_n$.

A similar relation for any $\omega \in \limsup_n A_n$ is

$$\omega \in \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \iff \forall n \exists i \geq n \text{ such that } \omega \in A_i.$$

Examples

1. Consider the alternating-constant series: $A_{2n} = B$ and $A_{2n+1} = C$. Then $\liminf_n A_n = B \cap C$ and $\limsup_n A_n = B \cup C$.
2. Let the sample space be \mathbb{R} , and let $A_{2n} = (-n, n)$ and $A_{2n+1} = (0, 1/n)$. Then $\liminf_n A_n = \emptyset$ and $\limsup_n A_n = \mathbb{R}$.
3. Consider

$$A_n = \begin{cases} \left(\frac{1}{n}, \frac{3}{4} - \frac{1}{n}\right) & \text{for } n = 1, 3, 5, \dots \\ \left(\frac{1}{4} - \frac{1}{n}, 1 + \frac{1}{n}\right) & \text{for } n = 2, 4, 6, \dots \end{cases}$$

We have $\liminf_n A_n = \left[\frac{1}{4}, \frac{3}{4}\right)$ and $\limsup_n A_n = (0, 1]$.

The Borel-Cantelli lemmas

Let A_n be a sequence of events and P be a probability measure.

- If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$.

Proof: (First, notice that $P(\cup_{i=n}^{\infty} A_i)$ can be arbitrarily small if n is large enough.)

From $\limsup_n A_n \subset \cup_{i=n}^{\infty} A_i$, we have

$$\begin{aligned} P(\limsup_n A_n) &\leq P(\cup_{i=n}^{\infty} A_i) \\ &\leq \sum_{i=n}^{\infty} P(A_i) \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{because } \sum_{n=1}^{\infty} P(A_n) < \infty. \end{aligned}$$

- If $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.

We can see this by a similar argument as above.

Sets in \mathbb{R}

The material above applies to any kind of sets. Subsets of the field \mathbb{R} that retain the operations of that field have some special interesting properties.

First, recall some important definitions:

A set A of real numbers is called *open* if for each $x \in A$, there exists a $\delta > 0$ such that for each y with $|x - y| < \delta$ belongs to A .

A real number x is called a *point of closure* of a set A of real numbers if for every $\delta > 0$ there exists a y in A such that $|x - y| < \delta$. (Notice that every $y \in A$ is a point of closure of A .)

We denote the set of points of closure of A by \overline{A} .

A set A is called *closed* if $A = \overline{A}$.

Some simple fact follow:

- The intersection of a finite collection of open sets is open.
- The union of a countable collection of open sets is open.
- The union of a finite collection of closed sets is closed.
- The intersection of a countable collection of closed sets is closed.

Notice what is *not* said above (where we use the word “finite”).

Intervals in \mathbb{R}

A very important type of set is an *interval* in \mathbb{R} . Intervals are the basis for building important structures on \mathbb{R} .

All intervals are Borel sets.

The main kinds of intervals have forms such as $(-\infty, a)$, $(-\infty, a]$, (a, b) , $[a, b]$, $(a, b]$, $[a, b)$, (b, ∞) , and $[b, \infty)$.

$(-\infty, a)$, (a, b) , and (b, ∞) are open;

$[a, b]$ is closed;

$(a, b]$ and $[a, b)$ are neither (they are “half-open”).

$(-\infty, a]$ and $[b, \infty)$ are closed, although in a special way that sometimes requires special treatment.

$$\overline{(a, b)} = [a, b].$$

Some simple facts:

- $\bigcap_{i=1}^n (a_i, b_i) = (a, b)$ (some open interval)
- $\bigcup_{i=1}^{\infty} (a_i, b_i)$ is an open set
- $\bigcup_{i=1}^n [a_i, b_i]$ is a closed set
- $\bigcap_{i=1}^{\infty} [a_i, b_i] = [a, b]$ (some closed interval)

Notice what is *not* said above (where limit is n rather than ∞).
(Also note that intersection of intervals are intervals.)

Two types of interesting intervals are

$$\left(a - \frac{1}{i}, b + \frac{1}{i}\right)$$

and

$$\left[a + \frac{1}{i}, b - \frac{1}{i}\right].$$

Notice, of course, that

$$\lim_{i \rightarrow \infty} \left(a - \frac{1}{i}, b + \frac{1}{i}\right) = [a, b]$$

and

$$\lim_{i \rightarrow \infty} \left[a + \frac{1}{i}, b - \frac{1}{i}\right] = [a, b].$$

Some basic facts about sequences of unions

Infinite intersections and unions behave differently with regard to collections of open and closed sets. Recall our earlier statements that were limited to finite intersections and unions: $\bigcap_{i=1}^n (a_i, b_i)$ is an open interval, and $\bigcup_{i=1}^n [a_i, b_i]$ is a closed set.

Now, with our open and closed intervals of the special forms, for infinite intersections and unions, we have the important facts:

$$[a, b] = \bigcap_{i=1}^{\infty} \left(a - \frac{1}{i}, b + \frac{1}{i} \right)$$

$$(a, b) = \bigcup_{i=1}^{\infty} \left[a + \frac{1}{i}, b - \frac{1}{i} \right]$$

Iff $x \in A_i$ for some i , then $x \in \bigcup A_i$.

So if $x \notin A_i$ for any i , then $x \notin \bigcup A_i$. (This is why the union of the closed intervals above is not a closed interval.)

Likewise, we have

$$(a, b] = \bigcup_{i=1}^{\infty} \left[a + \frac{1}{i}, b \right] = \bigcap_{i=1}^{\infty} \left(a, b + \frac{1}{i} \right).$$

Recall a basic fact about probability (which we will discuss again from time to time):

$$\lim \Pr(A_i) \neq \Pr(\lim A_i).$$

Compare this with the fact from above:

$$\lim_{n \rightarrow \infty} \bigcup_{i=1}^n \left[a + \frac{1}{i}, b - \frac{1}{i} \right] \neq \bigcup_{i \rightarrow \infty} \left[a + \frac{1}{i}, b - \frac{1}{i} \right].$$

Equivalent definitions of the Borel σ -field

The facts that unions of closed sets may be open and that intersections of open intervals may be closed allow us to characterize the Borel σ -field \mathcal{B} in various ways. The canonical definition is that $\mathcal{B} = \sigma(\mathcal{C})$, where \mathcal{C} is the collection of all finite open intervals.

If \mathcal{D} is the collection of all finite *closed* intervals $\mathcal{B} = \sigma(\mathcal{D})$.

If \mathcal{A} is the collection of all intervals of the form (a, ∞) , then $\mathcal{B} = \sigma(\mathcal{A})$.

Types of Convergence

The first important point to understand about asymptotic theory is that there are different kinds of convergence of a sequence of random variables, $\{X_n\}$.

One type of convergence applies directly to the function (the random variable). This is the strongest convergence.

One type of convergence applies to expected values of powers of the random variable. This is also a very strong convergence.

One type of convergence applies to probabilities of the random variable being within a range of another random variable.

One type of convergence applies to the distribution of the random variable. This is the weakest convergence.

Almost sure (a.s.) We say that $\{X_n\}$ converges to X almost surely if

$$\lim_{n \rightarrow \infty} X_n = X \text{ a.s.}$$

We write

$$X_n \rightarrow_{\text{a.s.}} X.$$

In r^{th} moment (in L_r) We say that $\{X_n\}$ converges to X in r^{th} moment if

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

We write

$$X_n \rightarrow_{L_r} X.$$

In probability We say that $\{X_n\}$ converges to X in probability if for every $\epsilon > 0$,

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

We write

$$X_n \rightarrow_p X.$$

In distribution (in law) We say that $\{X_n\}$ with sequence of CDFs $\{F_n\}$ converges to X with CDF F in distribution or in law if at each point of continuity t of F ,

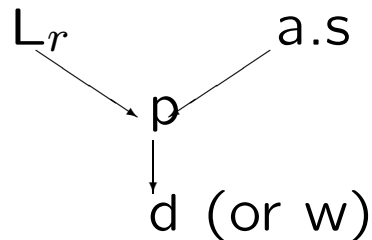
$$\lim_{n \rightarrow \infty} F_n(t) = F(t).$$

We write

$$X_n \rightarrow_d X.$$

If the sequence $\{X_n\}$ in distribution to X , we say that the sequence of CDFs $\{F_n\}$ converges weakly to the CDF of X , F , and we write $F_n \rightarrow_w F$.

We have the logical relations (parts of “Theorem 1.8”):



Convergence in distribution allows us to construct an a.s. convergent sequence. This is stated in Skorohod’s theorem (part of “Theorem 1.8”), whose proof is beyond the scope of this course. (It’s not hard, it’s just long and complicated.)

Skorohod’s Theorem: If for the random variables (vectors!) X_1, X_2, \dots , we have $X_n \rightarrow_p X$, then there exist random variables Y_1, Y_2, \dots on the same probability space with $P_{Y_n} = P_{X_n}$ and $P_Y = P_X$, such that $Y_n \rightarrow_{\text{a.s.}} Y$.

Big O and Little o Notation

We have the standard definitions for sequences of real numbers:

Big O, $O(a_n)$. $b_n = O(a_n)$ means $b_n/a_n \rightarrow c$ as $n \rightarrow \infty$, where c is a nonzero finite constant.

In particular, $b_n = O(1)$ means b_n is bounded.

Little o, $o(a_n)$. $b_n = o(a_n)$ means $b_n/f(n) \rightarrow 0$ as $n \rightarrow \infty$.

In particular, $b_n = o(1)$ means $b_n \rightarrow 0$.

A slightly different definition requires that the ratios never exceed a given c . They are equivalent for sequences all of whose elements are finite.

For sequences of functions (or of random variables) we need to qualify the types of convergence. For random variables X_n and Y_n , we say

$$X_n = O(Y_n) \text{ a.s. iff } \Pr(\|X_n\| = O(\|Y_n\|)) = 1$$

and

$$X_n = o(Y_n) \text{ a.s. iff } \|X_n\|/(\|Y_n\|) \rightarrow_{\text{a.s.}} 0.$$

(Compare $X_n/Y_n \rightarrow_{\text{a.s.}} 0$ for $X_n \in \mathbb{R}^m$ and $Y_n \in \mathbb{R}$.)

Now for a weaker convergence, we say

$$X_n = O_P(Y_n) \text{ iff } \forall \epsilon > 0 \exists \text{ constant } C_\epsilon > 0 \ni \sup_n \Pr(\|X_n\| \geq C_\epsilon \|Y_n\|) < \epsilon.$$

and

$$X_n = o_P(Y_n) \text{ iff } \|X_n\|/\|Y_n\| \rightarrow_p 0.$$

Convergence of Functions

The next issue has to do with functions of convergent sequences. For a sequence X_1, X_2, \dots in \mathbb{R}^k and a measurable function g from $(\mathbb{R}^k, \mathcal{B}^k)$ to $(\mathbb{R}^k, \mathcal{B}^k)$, the simple facts are

$$X_n \rightarrow_{\text{a.s.}} X \Rightarrow g(X_n) \rightarrow_{\text{a.s.}} g(X)$$

$$X_n \rightarrow_{\text{p}} X \Rightarrow g(X_n) \rightarrow_{\text{p}} g(X)$$

$$X_n \rightarrow_{\text{d}} X \Rightarrow g(X_n) \rightarrow_{\text{d}} g(X)$$

Slutsky's theorem is convergence in distribution, for the case that one sequence converges to a random variable and another sequence converges to a fixed real number. It tells us that sums, products, and quotients behave like we would expect (or hope):

$$X_n + Y_n \rightarrow_d X + c$$

$$X_n Y_n \rightarrow_d X + cX$$

$$X_n/Y_n \rightarrow_d X/c \text{ if } c \neq 0.$$

This is Theorem 1.11 in Shao.

Another useful fact is given as Theorem 1.12(i) in Shao:
Let X_1, X_2, \dots and Y be random variables (k -vectors) such that

$$a_n(X_n - c) \rightarrow_d Y,$$

where c is a constant (k -vector) and a_1, a_2, \dots is a sequence of constant scalars such that $\lim_{n \rightarrow \infty} a_n = \infty$. Now let g be a function from \mathbb{R}^k to \mathbb{R} that is differentiable at c . Then

$$a_n(g(X_n) - g(c)) \rightarrow_d (\nabla g(c))^T Y.$$

There is an extension of this given as Theorem 1.12(ii) in Shao for powers of the a_n that has applications in the covariance of the random vector Y .

The most common application of Theorem 1.12(i) arises from the simple corollary (called “Corollary 1.1” in Shao) for the case when Y has the multivariate normal distribution $N_k(0, \Sigma)$:

$$a_n(g(X_n) - g(c)) \rightarrow_d Z,$$

where $Z \sim N_k(0, (\nabla g(c))^T \Sigma \nabla g(c))$

One reason limit theorems are important is that they can provide approximations useful in statistical inference. For example, the convergence of the sequence above provides a method for setting approximate confidence sets using the normal distribution. This method in asymptotic inference is called the *delta method*.

Limit Theorems

There are two general types of important limit theorems: laws of large numbers and central limit theorems. Laws of large numbers give limits for probabilities or for expectations of sequences of random variables. Central limit theorems also do this and more; they specify a limiting *normal distribution*. The first versions of both laws of large numbers and central limit theorem applied to sequences of binomial random variable.

The first law of large numbers was **Bernoulli's** (Jakob) **theorem**: If X_n has a binomial distribution with parameters n and π , then

$$X_n/n \rightarrow_p \pi.$$

This follows from $\int_{\Omega} (X_n/n - \pi)^2 dP = \pi(1 - \pi)/n$, which means X_n/n converges in mean square to π , which in turn means that it converges in probability to π . This is a *weak law* because the convergence is in probability.

The first central limit theorem, called the de Moivre Laplace central limit theorem followed soon after this. It applies to X_n that has a binomial distribution with parameters n and π .

The **de Moivre Laplace central limit theorem** states that

$$\frac{X_n - n\pi}{\sqrt{n\pi(1 - \pi)}}$$

has a normal (0,1) distribution. This central limit theorem, called the is a special case of the classical central limit theorem for i.i.d. random variables with finite mean and variance.

Notice that these two limit theorems, which are stated in terms of binomial random variables, apply to normalized limits of sums of Bernoulli random variables. This is the usual form of these kinds of limit theorems; that is, they apply to normalized limits of sums of random variables.

The first generalizations apply to sums of i.i.d. random variables, and then further generalizations apply to sums of just independent random variables.

A generalization of the Bernoulli's theorem is the **weak law of large numbers for i.i.d. random variables:**

Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables. There exists a sequence of real numbers a_1, a_2, \dots such that

$$\frac{1}{n} \sum_{i=1}^n X_i - a_n \rightarrow_p 0$$

iff $n\Pr(|X_1| > n) \rightarrow 0$. If this condition holds, we can choose $a_n = E(X_i \mathbf{I}_{\{|X_1| \leq n\}})$.

We will not prove this or the following limit theorems in this course.

If $E(|X_1|) < \infty$, we can form a strong law in terms of a $E(X_1)$; the

weak law of large numbers for i.i.d. random variables:

Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables such that $E(|X_1|) < \infty$. Then

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow_{\text{a.s.}} E(X_1).$$

(Shao states this in a slightly different form.)

A slight generalization is the alternate conclusion

$$\frac{1}{n} \sum_{i=1}^n c_i (X_i - E(X_1)) \rightarrow_{\text{a.s.}} 0,$$

for any bounded sequence of real numbers c_1, c_2, \dots

We can generalize these two limit theorems to the case of independence but not necessarily identical distributions, by putting limits on normalized p^{th} moments.

The **weak law of large numbers for independent random variables with finite expectation**: Let X_1, X_2, \dots be a sequence of independent random variables such for some constant $p \in [1, 2]$,

$$\lim_{n \rightarrow \infty} \frac{1}{n^p} \sum_{i=1}^n \mathbb{E}(|X_i|^p) = 0,$$

then

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X_i)) \rightarrow_p 0.$$

The **strong law of large numbers for independent random variables with finite expectation**: Let X_1, X_2, \dots be a sequence of independent random variables such for some constant $p \in [1, 2]$,

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}(|X_i|^p)}{i^p} < \infty,$$

then

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X_i)) \xrightarrow{\text{a.s.}} 0.$$

The **central limit theorem for i.i.d. scalar random variables with finite mean and variance:**

Let X_1, X_2, \dots be a sequence of independent random variables that are identically distributed with mean μ and variance $\sigma^2 > 0$. Then

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

converges in distribution to a random variable that is $N(0, 1)$.

The proof of this uses a limit of a characteristic function and the uniqueness of the characteristic function.

A more general central limit theorem is called Lindeberg's central limit theorem. It is stated in terms of a sequence of finite subsequences.

The central limit theorem for independent scalar random variables with finite mean and variance:

Let $\{X_{nj}, j = 1, 2, \dots, k_n\}$ be independent random variables with $0 < \sigma_n^2$, where $\sigma_n^2 = \text{V}(\sum_{j=1}^{k_n} X_{nj})$ and $k_n \rightarrow \infty$ as $n \rightarrow \infty$. If the *Lindeberg condition*,

$$\sum_{j=1}^{k_n} \text{E} \left((X_{nj} - \text{E}X_{nj})^2 \mathbf{I}_{(X_{nj} - \text{E}X_{nj}) > \epsilon \sigma_n} \right) = o(\sigma_n^2) \text{ for any } \epsilon > 0,$$

holds, then

$$\frac{1}{\sigma_n} \sum_{j=1}^{k_n} (X_{nj} - \text{E}X_{nj}) \rightarrow_d \text{N}(0, 1).$$