STAT/MATH 6105

MEASURE AND INTEGRATION FOR PROBABILITY

Lecture Notes for Fall 1997

Preface

These notes have evolved through teaching this course over many years. They began mostly as a "record" of the material covered, a list of definitions, theorem statements and examples. Now however they also contain proofs (or outlines of proofs) for many of the main results, as well as a few supplemental topics that past students have asked for. However they are not intended to be a completely self-contained treatment. Rather they and the lectures are meant as compliments to each other. In some places you will need the lectures for a full explanation of details. (Some acknowledged omissions indicated with an ellipsis " . . . ".) At other places we will use our class time for additional examples or clarifying discussion, leaving technical details to the notes alone. Your comments and suggestions are welcome, and will influence future improvements to the notes. For a more thorough treatment we recommend the excellent book by Billingsley [1]. Billingsley's text has had a strong influence on these notes. In particular the credit for a number of the problems is his.

– M. Day, June 1997

Contents

The material is organized into the following units:

Equations, problems and examples are numbered consecutively within a unit. Formally stated results (i.e. theorems and lemmas) will be labeled A, B, ... To refer to equations or results outside the current unit the labels will be preceded by the unit reference. Thus Theorem III.D means Theorem D of Unit III and (IV.12) is equation (12) of Unit IV.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

References

There are many books discussing this material. The list below includes both classics and recent texts, but is only a sampling of the books on the subject. [1] is the former text for the course, though [3] has also been used. [13] is the most common text for Math 5225, though [4] has also been used. [12] and [15] are intended to be advanced undergraduate texts. [9] is the original, though now outdated. [7] is also a classic.

[1] P. Billingsley, Probability and Measure, Wiley, NY, 1986.

[2] L. Breiman, Probability, Addison-Wesley, Reading, MA, 1968.

[3] K. L. Chung, A Course In Probability Theory, Academic Press, 1974.

[4] R. M. Dudley, Real Analysis and Probability, Wadsworth & Brooks/Cole, Pacific Grove, 1989.

[5] W. Feller, Am Introduction to Probability Theory and its Applications, Vol II, 2nd ed., Wiley, 1971.

[6] B. V. Gnedenko and A. N. Kolmogorov, Limit Distributions for Sums of Independent Random Variables, Addison-Wesley, Reading, Mass., 1954.

[7] P. R. Halmos, Measure Theory, Van Nostrand Reinhold, NY, 1950.

[8] J. F. C. Kingman and S. J. Taylor, Introduction to Measure and Probability, Cambridge Univ. Press, 1973.

[9] A. N. Kolmogorov, Foundations of the Theory of Probability, Chelsa, 1956.

[10] M. Loeve, Probability Theory, Springer-Verlag, NY, 1977.

[11] J. Neveu, Mathematical Foundations of the Calculus of Probability, Holden-Day, 1965.

[12] P. E. Pfeiffer, Concepts of Probability Theory, Dover, 1978.

[13] H. L. Royden, Real Analysis, Macmillan, NY, 1968.

[14] A. N. Shiryayev, Probability, Springer-Verlag, NY, 1984.

[15] D. Williams, Probability with Martingales, Cambridge Univ. Press, 1991.

The following references cna be consulted for more on Chernoff's Theorem (IV.E).

[16] R. R. Bahadur, Some Limit Theorems in Statistics, SIAM, Philadelphia, 1971.

[17] S. R. S. Varadhan, Large Deviations and Applications, SIAM, Philadelphia, 1984.

The job of a statistician is to select and apply statistical procedures to analyze “real” data. Understanding the properties of these procedures is of course vital to deciding when their use is appropriate. To understand the properties of these statistical procedures in a careful way we have to study their properties in a precise mathematical setting. For instance, to say an estimator Θ of some parameter θ is unbiased is a statement about its properties as a mathematical object: $E_\theta[\Theta] = \theta$. Probability theory is the mathematical setting in which most statistical procedures are studied. Our goal in this course is to understand the mathematical concepts of modern probability theory.

You are probably familiar with “primitive” probability theory, as illustrated in the following examples.

Example 1. A discrete probability space, consisting of a finite number N of distinct “events” ω_1, ω_2, ... , ω_N, each with an associated probability p_i. (We insist that $0 \le p_i \le 1$ and $\sum_1^N \; p_i = 1$). $\diamond\diamond$

Example 2. The above could be extended to allow an infinite sequence p_i; $i = 1, 2, \ldots$ with $0 \le p_i$ and $\sum_1^\infty p_i = 1$. $\diamond\diamond$

Example 3. A continuous probability density is a continuous function $f(x) \ge 0$ defined for all real numbers x (i.e. $f : \mathbb{R} \to [0, \infty)$) with

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

Then with any interval $A = (a, b]$ (or finite disjoint union of intervals) we can associate a probability

$$P((a, b]) = \int_a^b f(x)\, dx.$$

This might describe the “distribution” of a random variable X; $P(A) =$ the probability that $X \in A$. We would then calculate expected values of functions of X by

$$E[\phi(X)] = \int_{-\infty}^{\infty} \phi(x) f(x)\, dx.$$

$\diamond\diamond$

Example 4. Example 3 can be generalized by describing the probability of an interval using a distribution function $F(\cdot)$:

$$P((a, b]) = F(b) - F(a).$$

($F(\cdot)$ must be non-decreasing and satisfy $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.) For instance the distribution of Poisson or binomial random variables can be described this way, although they do not have continuous densities. In general it may not be clear how to generalize the formula of Example 3:

$$E[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)\, dF(x) \quad ?$$

$\diamond\diamond$

Example 5. In the plane, $\mathbb{R}^2$, we can also consider continuous densities $f(x, y)$. The probability of a subset $A \subseteq \mathbb{R}^2$ could be computed by

$$P(A) = \iint_A f(x, y)\, dx\, dy.$$

This could describe the joint distribution of a pair of random variables; $P(A) =$ the probability that $(X, Y) \in A$. Now we would calculate

$$E[\phi(X)] = \iint \phi(x) f(x, y)\, dx\, dy.$$

$\diamond\diamond$

In each of these examples we have a set of elements Ω and a rule for assigning probabilities $P(A)$ to certain subsets $A \subseteq \Omega$. $P(A)$ is not necessarily defined for all subsets A however. Those $A \subseteq \Omega$ for which $P(A)$ is defined form a special class $\mathcal{A}$ of subsets of Ω. One eventually realizes that simple settings such as these are not always adequate, that something more sophisticated mathematically is needed. Example 4 hints at this but the example we develop next will make the point more strongly.

A Game of Ten-Sided Dice

Imagine that we have a dice with 10 sides, numbered 0, 1, ... , 9. The dice is fair, i.e. each side has probability $\frac{1}{10}$ of landing up. We roll the dice repeatedly and independently, obtaining a sequence of digits

$$d_1 \ d_2 \ d_3 \ \dots \ d_n \ \dots,$$

each digit being one of the integers 0, ... , 9. We want to assign probabilities to sets of such sequences. For instance we would say that the set of all sequences with first digit $d_1 = 3$, i.e. all sequences of the form

$$3 \ d_2 \ d_3 \ \dots \ d_n \ \dots,$$

should have probability $\frac{1}{10}$

There is a nice way to assign probabilities to sets of such sequences, by thinking of the sequence of digits as the decimal representation of a real number ω:

$$\omega = .d_1 d_2 d_3 \cdots = \sum_{n=1}^{\infty} d_n/10^n. \tag{1}$$

For instance, $\omega = 123/999 = .123123123123\ldots$. This association of digit sequences with numbers is not perfect however because some numbers have two different decimal representations, such as

$$1/2 = .5000\cdots = .4999\ldots$$

For such ω we will insist on using the decimal representation with trailing 0's, not trailing 9's. (This excludes those outcomes of our dice game which are all 9's after some point, but the probability of these should be 0 anyway.) Thus we will take $\Omega = [0, 1)$ as our set of basic elements. Each $\omega \in [0, 1)$ determines a full digit sequence $d_1(\omega)$, $d_2(\omega)$, ... from the decimal representation (1), with no trailing 9's. Each d_n is viewed as a function of ω, $d_n : \Omega \to \{0, 1, \dots 9\}$.

Example 6. The set of ω for which $d_1(\omega) = 3$ is just an interval of real numbers:

$$\{\omega : \ d_1(\omega) = 3\} = [.3, .4).$$

We want to assign probability $\frac{1}{10}$ to this set of ω. Observe that $\frac{1}{10}$ is precisely the length of the interval! ◇◇

We take this example as our lead, and assign the length of a subset $A \subset [0, 1)$ as its probability $P(A)$:

$$P(A) = b - a \quad \text{if } A = [a, b) \text{ with } 0 \le a \le b \le 1,$$

and if $A = \cup_1^n [a_i, b_i)$ with $[a_i, b_i)$ disjoint (i.e. no two intersect) then define

$$P(A) = \sum_1^n (b_i - a_i).$$

Note that we can only calculate $P(A)$ if A is a set of the proper type. We could certainly make some obvious extensions (to include $A = (a, b), [c, b], (a, b]$ and finite unions of such) but there remain many sets $A \subset [0, 1)$ for which $P(A)$ is not defined!

We can confirm that this set-up does accurately describe the probabilities of our dice game by checking the probability it produces for a specified outcome of the first N rolls. Suppose we pick digits u_1, u_2, ... , u_N and ask for the probability that the first dice produces value u_1, the second dice produces u_2, ... and dice N produces u_N. In our set-up we calculate this by finding the set A of ω's which meet the prescribed requirements and then evaluate $P(A)$:

$$\begin{aligned} A &= \{\omega : \ d_n(\omega) = u_n \text{ each } n = 1, \cdots, N\} \\ &= [a, a + 10^{-N}), \quad \text{where } a = \sum_1^N u_n/10^n, \end{aligned}$$

and so we do get the correct value:

$$P(A) = 1/10^N.$$

The really interesting questions come from considering events that involve the whole sequence of dice rolls, not just some finite number of them. Consider in particular

$$\frac{1}{n}\sum_1^n d_i(w),$$

which is the average of the values rolled in the first n plays. As $n \to \infty$ we expect this average to converge to 4.5 with probability 1. (This is the Strong Law of Large Numbers.) To be precise, we want to say that $P(H) = 1$, where

$$H = \{\omega : \ \lim_{n\to\infty} \frac{1}{n}\sum_1^n d_i(w) = 4.5\}.$$

But now we are faced with a serious problem: H is not a finite disjoint union of intervals, so $P(H)$ is not defined! In problem 1 below you will show that every interval $(a, b) \subseteq [0, 1)$ contains points from both H and H^c. Thus both H and H^c are spread throughout $[0, 1)$, having no segments or gaps of any positive length. Thus our intuition does not tell us what the length $P(H)$ of a complicated set like H should even mean, much less why $P(H) = 1$ is the correct value.

The point is that the definition of P needs to be extended from the simple type of sets (finite unions of intervals) for which the value of P is clear, to sets of more complicated type (like H) for which the definition of P is not clear at the outset. This extended definition of P must satisfy certain properties in order to be a reasonable "measure of probability". Any proof that $P(H) = 1$ must make essential use of the properties which govern this extension. In other words, the assignment of probabilities by our intuition alone is not adequate. In general what we need to do is set down the mathematical principles that govern the assignment of probabilities (the definitions of what are called "probability measures" and "sigma-fields"), and then study how these principles determine those probabilities that are beyond our intuition.

Overview

The basic setup of probability theory is similar to what we constructed above, based on a "probability space" $(\Omega, \mathcal{F}, P)$. The "master set" Ω (like $[0, 1)$ above) encompasses all possibilities. $\mathcal{F}$ is the collection (class) of subsets $A \subset \Omega$ for which probabilities $P(A)$ will be defined. P is the "probability measure" which assigns probabilities to $A \in \mathcal{F}$. In addition we often have random variables X (like our $d_1, d_2, \cdots$) which are functions taking $w \in \Omega$ to $X(w) \in \mathbb{R}$, or some other set of outcomes like our $\{0, 1, 2, \ldots 9\}$. Our first goal

(Unit I) is to discuss the properties that $\Omega, \mathcal{F}$ and P must have to be mathematically adequate (to surmount the type of problem we encountered with H above instance). Then (Unit II) we need to talk about the properties that random variables must have in order to be compatible with $(\Omega, \mathcal{F}, P)$.

The expected value "$E[X]$" is what mathematicians call the "integral of X with respect to the measure P", usually written

$$E[X] = \int_\Omega X \, dP.$$

This notion of integration is more general and powerful than that of calculus. If $P([a,b)) = b - a$ gives the measure of length as above (called "Lebesgue measure") then $\int_{[a,b)} f(x)\, dP$ extends the Riemann integral $\int_a^b f(x)\, dx$. An infinite series $\sum_1^\infty f(n)$ is another special case of this new notion of integral, this time using "counting measure" on $\Omega = \mathbb{N}$. After discussing this concept of integration (Unit III) we take a quick overview of the various notions of convergence common in probability theory and some of the most famous limit laws (Unit IV).

These measure-theoretic foundations also allow a unified understanding of conditional probabilities. The various conditioning formulas of elementary probability theory, such as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad f_{X|Y}(x|y) = \frac{f(x,y)}{\int f(x,y)\, dx}$$

will all be seen as different expressions of the same idea. By understanding what conditioning really is mathematically we will be able to explain the important rules for manipulating conditional probabilities and expectations (Unit V). This is fundamental to understanding several important classes of stochastic processes, such as martingales and Markov processes.

Problem **1** ....

Let $H \subseteq [0,1)$ be the set described above. Show that every non-empty open interval $(a,b) \subseteq [0,1)$ contains a point from H and a point from H^c. As a hint, notice that if we take ω and alter the terms of its decimal expansion after d_n, we obtain a new $\tilde{\omega}$with

$$\begin{aligned} |\omega - \tilde{\omega}| &= \left| \sum_{k=n+1}^\infty (d_k(\omega) - d_k(\tilde{\omega}))/10^k \right| \\ &\leq \sum_{k=n+1}^\infty 9/10^k = 10^{-n}. \end{aligned}$$

Using this you can construct $\tilde{\omega}$of either type (in H or H^c) as close as you like to a given ω.

Problem **2** ....

Let

$$\delta_k(i) = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases}.$$

For each $k = 0, 1, \ldots, 9$ define the set

$$A_k = \{\omega \in [0,1) : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \delta_k(d_i(\omega)) = \frac{1}{10}\}.$$

Show that

$$\cap_{k=0}^9 A_k \subseteq H.$$

Problem **3** ....

Draw graphs of $d_1(\cdot)$ and of $d_2(\cdot)$. Explain why, according to our definitions above,

$$P(\{\omega : \ d_2(\omega) = 7\}) = 1/10.$$

We begin a careful study of "measures of probability" by describing their essential properties. Let Ω be any "space" or master set of points $\omega \in \Omega$.

Example(s) 1.

- $\Omega = \{\omega_1, \dots, \omega_N\}$ – Ex. M.1.
- $\Omega = \{1, 2, 3, \dots\}$ – Ex. M.2.
- $\Omega = \{$ all sequences $\omega = HTTHTHHTHTHTTT\dots\}$ – Coin tossing.
- $\Omega = \{f : [0, \infty) \to \mathbb{R}\}$ – random motions.
- $\Omega = \mathbb{R}$ – probability distributions.
- $\Omega = [0, 1)$ – 10-sided dice, Unit M. ◇◇

We will want to define probabilities $P(A)$ for certain subsets $A \subseteq \Omega$. There will be a collection $\mathcal{F}$ of subsets of Ω consisting of those $A \subseteq \Omega$ for which $P(A)$ is be defined. (A set is a collection of elements but we use the word *class* for a collection of (sub)sets. Thus $\mathcal{F}$ is a class of subsets of Ω; see the Mathematical Supplements.) What properties should $\mathcal{F}$ have? Our calculations are likely to manipulate sets in $\mathcal{F}$ via the usual set theoretic operations: compliment, intersection and union. We want the resulting sets also to be in $\mathcal{F}$. This means we want $\mathcal{F}$ to be a what is called a field of subsets.

Definition. *A class $\mathcal{F}$ of subsets of Ω is called a field when the following properties hold:*

(i) $\Omega \in \mathcal{F}$;
(ii) if $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$;
(iii) if $A, B \in \mathcal{F}$ *then* $A \cup B \in \mathcal{F}$.

If $A \in \mathcal{F}$ we say A is an $\mathcal{F}$ set, or that A is $\mathcal{F}$ measurable.

Example 2. Consider $\Omega = [0, 1)$ and the following classes of subsets:

- $\mathcal{I}$ = class of all intervals, $[a, b) \subseteq \Omega$ — this is not a field.
- $\mathcal{S}$ = class of finite disjoint unions of intervals $A = \cup_1^n [a_i, b_i)$ — this is a field. (We consider $\emptyset = [a, a)$ to be in $\mathcal{S}$.) ◇◇

Notice that the following properties are consequences of (i)—(iii): $\emptyset = \Omega^c \in \mathcal{F}$, and $A \cap B = (A^c \cup B^c)^c \in \mathcal{F}$ if $A, B \in \mathcal{F}$. Moreover if $A_n \in \mathcal{F}$ for each $n = 1, \dots, N$, then $\cap_1^N A_n$ and $\cup_1^N A_n$ must also be in $\mathcal{F}$. In other words any set we can form by a finite number of operations (intersections, unions, or compliments) applied to $\mathcal{F}$ sets must also be an $\mathcal{F}$ set.

We saw in the last section that for $\Omega = [0, 1)$ even the field $\mathcal{S}$ was not adequate for the discussion of infinite sequences of coin tosses, because $H \notin \mathcal{S}$. However,

$$H = \cap_{k=1}^{\infty} \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} \{\omega : \ |\frac{1}{n} \sum_1^n d_i(\omega) - 4.5| < \frac{1}{k}\}.$$

So we also want to be able to take unions and intersections of sequences of sets in $\mathcal{F}$; i.e. "countable" unions and intersections.

Definition. *A class $\mathcal{F}$ of subsets of Ω is called a σ-field (or σ-algebra) if $\mathcal{F}$ is a field and*

$$\cup_{i=1}^{\infty} A_i \in \mathcal{F} \text{ whenever } A_1, A_2, \dots \in \mathcal{F} \text{ is a sequence of sets in } \mathcal{F}.$$

Notice that a σ-field must also contain countable intersections: if $A_i \in \mathcal{F}$ then $\cap_{i=1}^{\infty} A_i = (\cup_{i=1}^{\infty} A_i^c)^c$ must also be in $\mathcal{F}$ since $\mathcal{F}$ is "closed under complementation".

Example(s) 3. Most σ-fields are complicated but here are a few simple ones.

- $\mathcal{F} = \{\emptyset, \Omega\}$,
- $\mathcal{P} = \{$ all subsets $A \subseteq \Omega\}$,
- $\mathcal{C} = \{A \subseteq \Omega :$ either A or A^c is countable $\}$ – see Problem 2. $\diamond\diamond$

The difficulty we encountered in our 10-sided dice discussion is simply that $\mathcal{S}$, the class on which we knew how to define P, is not a a σ-field. To resolve this we need to "extend" the definition of P to a larger class of sets which is a σ- field. What σ-field is appropriate? $\mathcal{P}$ (from the preceding example) turns out to be too big – $P(A)$ cannot be defined for all subsets $A \subseteq [0, 1)$. $\mathcal{C}$ is clearly too small; it doesn't contain bounded intervals. What we want is the smallest σ-field that includes all the sets in $\mathcal{S}$.

Definition. *If $\mathcal{A}$ is a class of subsets of Ω, we define $\sigma(\mathcal{A})$ to be the class of all $A \subseteq \Omega$ such that A belongs to every σ-field containing $\mathcal{A}$:*

$$\sigma(\mathcal{A}) = \{A \subseteq \Omega : \ A \in \mathcal{F} \textit{ for every } \sigma\textit{- field } \mathcal{F} \textit{ with } \mathcal{A} \subseteq \mathcal{F}\},$$

You can check that $\sigma(\mathcal{A})$ is itself a σ-field, called the σ-field *generated by* $\mathcal{A}$. It is the smallest σ- field containing $\mathcal{A}$.

Example 2 (continued). Define $\mathcal{B} = \sigma(\mathcal{S})$. The sets in $\mathcal{B}$ are called the *Borel sets* in $[0, 1)$. It is impossible to give a direct description of the sets which are in $\mathcal{B}$, but virtually every set you can describe is. For instance $\mathcal{B}$ contains all open subsets of $[0, 1)$. This is because

1. $(a, b) = \cup_{n=k}^{\infty}[a + \frac{1}{n}, b) \in \mathcal{B}$, (where $a + \frac{1}{k} < b$), and
2. If G is open then $G = \cup_{\Gamma}(a, b)$ where

$$\Gamma = \{(a, b) : \ a, b \text{ are rational and } (a, b) \subseteq G\}.$$

Since Γ is countable, $G \in \mathcal{B}$.

There are subsets of $(0, 1]$ which are not Borel sets but they are impossible to describe explicitly. (See the discussion later in this section.) $\diamond\diamond$

Measures

Given a set Ω and a σ-field $\mathcal{F}$ of subsets of Ω, the next thing that we need is the "rule" or "set function" P which assigns to each $A \in \mathcal{F}$ its probability $P(A)$. P is what we will call a *probability measure*. We want $0 \leq P(A) \leq 1$ for probabilities. However there are other settings in which we want to assign "sizes" $\mu(A)$ to $A \in \mathcal{F}$ without the restriction $\mu(A) \leq 1$, or even $\mu(A) < \infty$. These are what are called (general) measures. We define them first.

Definition. *Let $\mathcal{F}$ be a field on a Ω. A measure μ on $\mathcal{F}$ is a function $\mu : \mathcal{F} \to [0, +\infty]$ satisfying*

(i) $\mu(\emptyset) = 0$,

(ii) if $A_1, A_2, \cdots$ is a sequence of disjoint sets in $\mathcal{F}$, and if $\cup_1^{\infty} A_n \in \mathcal{F}$, then

$$\mu(\cup_1^{\infty} A_n) = \sum_1^{\infty} \mu(A_n).$$

Remarks.

- We only required $\mathcal{F}$ to be a field in this definition, but we are most interested in the case in which $\mathcal{F}$ is a σ-field. If $\mathcal{F}$ is σ-field, we do not need to say "if $\cup_1^{\infty} A_n \in \mathcal{F}$" in part (ii).

- We allow $\mu(A) = +\infty$. Thus the definition assumes we have some conventions regarding arithmetic on $[0, \infty]$. The conventions are all quite natural; see Unit S.
- Part (ii) is called *countable additivity*. A less demanding property is *finite additivity*: $\mu(A \cup B) = \mu(A) + \mu(B)$ for any disjoint pair $A, B \in \mathcal{F}$. (This implies the natural generalization to any finite number of sets: $\mu(\cup_1^n A_i) = \sum_1^n \mu(A_i)$ when $A_i \in \mathcal{F}$ are disjoint.) Problem 3 shows that it is possible for μ to be finite but not countably additive.
- There are also "signed measures" which allow $\mu(A)$ to be negative, but we do not need to deal with them.

A measure μ is called

- a *probability* measure if $\mu(\Omega) = 1$,
- a *finite* measure if $\mu(\Omega) < \infty$,
- *infinite* if $\mu(\Omega) = \infty$,
- *σ-finite* if there exist $A_1, A_2, \dots \in \mathcal{F}$ with $\Omega = \cup_1^\infty A_n$ and $\mu(A_n) < \infty$ for each n.

When $\mathcal{F}$ is a σ-field on Ω and μ is a measure on $\mathcal{F}$, the triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*. When $P = \mu$ is a probability measure, $(\Omega, \mathcal{F}, P)$ is called a *probability space*. The pair $(\Omega, \mathcal{F})$ (with no measure specified) is a measur*able* space.

Example 4. With $(\Omega, \mathcal{S})$ as in Example 2 we can define P on $A \in \mathcal{S}$ as in Unit M: $P([a, b)) = b - a$. It seems obvious that P is countably additive. This is true, though the proof is more involved than you might think. (See Theorem E below.) We will see that P extends (in a "unique" way) to all of $\mathcal{B} = \sigma(\mathcal{S})$. The extended version is called *Lebesgue* measure, often denoted by ℓ below. ◇◇

Example 5. The above can be carried out on $\Omega = \mathbb{R}$ as well. Let $\mathcal{J}$ consist of all finite (disjoint) unions of intervals of the form $(-\infty, a]$, $(a, b]$, or $(b, +\infty)$. (Note the reversal of open/closed endpoints, compared to Example 2.) $\mathcal{J}$ is a field. Define

$$\ell((-\infty, a]) = +\infty, \quad \ell((a, b]) = b - a, \quad \ell((b, +\infty)) = +\infty$$

and

$$\ell(\cup_1^n J_i) = \sum_1^n \ell(J_i),$$

if the J_i are disjoint intervals of the indicated types. Theorem E below will tell us that $\ell(\cdot)$ can be extended to a measure on $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{J})$ (the Borel sets) called *Lebesgue measure* on $\mathbb{R}$. Here ℓ is infinite, but σ-finite:

$$\mathbb{R} = \cup_1^\infty (-n, n]; \quad \ell((-n, n]) = 2n < \infty \text{ each } n.$$

By arguing as in Example 2 it follows that $\mathcal{B}(\mathbb{R})$ contains all intervals (of any type), all open sets and all closed sets. ◇◇

Example 6. If $f \geq 0$ is a continuous probability density, $\int_{-\infty}^\infty f(t)\,dt = 1$, then there exists (Theorem E again) P on $\mathcal{B}(\mathbb{R})$ with the property that

$$P((a, b]) = \int_a^b f(t)\,dt$$

The function $F(x) = \int_{-\infty}^x f(t)\,dt$ is called the *distribution function* associated with P:

$$F(x) = P((-\infty, x])$$

◇◇

Example 7. Let $\Omega =$ any set, $\mathcal{F} =$ all subsets, and define $\mu(A)=$ the number of elements in A. This is a measure, called *counting measure*. ◇◇

Additional Properties of Measures. The definition of a measure implies a number of other elementary properties. Suppose μ is a measure on a field $\mathcal{F}$ of subsets of Ω. (All sets referred to below are assumed to be in $\mathcal{F}$.) Probability measures satisfy a few extra properties, which are indicated by writing $P(\cdot)$ instead of $\mu(\cdot)$.

- If $A \subseteq B$ then $\mu(A) \le \mu(B)$.
-
$$\mu(A) + \mu(B) = \mu(A \cup B) + \mu(A \cap B)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$
- If $A_1 \subseteq A_2 \subseteq \cdots A_n \subseteq \ldots$ and $A = \cup_1^\infty A_n$ (i.e. "$A_n \uparrow A$") then $\lim \mu(A_n) = \mu(A)$ (i.e. "$\mu(A_n) \uparrow \mu(A)$").
- If $A_1 \supseteq A_2 \supseteq \cdots A_n \supseteq \cdots$ and $A = \cap_1^\infty A_n$ ("$A_n \downarrow A$") and if $\mu(A_n) < \infty$ for some n, then $\mu(A) = \lim \mu(A_n)$ ("$\mu(A_n) \downarrow \mu(A)$"). (Note that we always have $P(A) = \lim P(A_n)$ because $P(A_n) \le 1 < \infty$.)
- $\mu(\cup_1^\infty A_n) \le \sum_1^\infty \mu(A_n)$ (any sequence of $A_n \in \mathcal{F}$ with $\cup A_n \in \mathcal{F}$).
- If μ is σ-finite, then any collection of disjoint $\mathcal{F}$-sets of positive measure is countable.

Proof(s): ... ▮

The existence of distribution functions is not limited to the situation of Example 6. In fact every probability measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ has a distribution function defined by $F(x) = P((-\infty, x])$. The properties of a measure imply that $F(\cdot)$ is a function from $\mathbb{R}$ to $[0,1]$ which satisfies the following:

- if $x \le y$ then $F(x) \le F(y)$ (nondecreasing);
- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$;
- for every x, $F(x) = \lim_{y\downarrow x} F(y)$ (right continuous).

We can recover P for intervals from F by the formula

$$P((a,b]) = F(b) - F(a).$$

Now it is important which endpoints are open/closed! Also note that $P(\{a\}) = F(a) - F(a-)$.

Conversely, for any such function F there is a unique probability measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with $P((a,b]) = F(b) - F(a)$. This is the content of Theorem E below. Notice what we are saying here with the word "unique": the values of $P(A)$ for all $A \in \mathcal{B}(\mathbb{R})$ are determined by the values of $P(J)$ for intervals $J = (a,b]$, even though there may be no way to write A in terms of intervals!

Generation of Sigma-Fields

We have already defined $\sigma(\mathcal{A})$ where $\mathcal{A}$ is any class of subsets of Ω. Here are some facts about this process of "generating" σ-fields.

- $\sigma(\mathcal{A})$ is a σ-field and contains all $A \in \mathcal{A}$.
- If $\mathcal{A} \subseteq \mathcal{F}$ and $\mathcal{F}$ is a σ-field then $\sigma(\mathcal{A}) \subseteq \mathcal{F}$.
- If $\mathcal{A}$ is itself a σ-field then $\sigma(\mathcal{A}) = \mathcal{A}$.
- If $\mathcal{A}_1 \subseteq \mathcal{A}_2$ then $\sigma(\mathcal{A}_1) \subseteq \sigma(\mathcal{A}_2)$.
- If $\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \sigma(\mathcal{A}_1)$ then $\sigma(\mathcal{A}_1) = \sigma(\mathcal{A}_2)$.

Example 8. $\mathcal{A} = \{A, B, C\}$. $\sigma(\mathcal{A})$ consists of those sets made up of the 8 basic sets $A \cap B \cap C$, $A \cap B^c \cap C$, ... Note that some sets like $A^c \cap B^c \cap C^c$ or $A^c \cap B^c \cap C$ may consist of more than one "piece" in a drawing but are inseparable in $\sigma(\mathcal{A})$. ◇◇

Example 9. $\mathcal{V} = \{(-\infty, c] \times \mathbb{R} : \ c \in \mathbb{R}\}$. $\sigma(\mathcal{V})$ consists only of sets of the form $A \times \mathbb{R}$ ($A \in \mathcal{B}(\mathbb{R})$). I.e. if $G \in \sigma(\mathcal{V})$ and $(x, y) \in G$ then the complete vertical line $L_x = \{(x, z) : \ z \in \mathbb{R}\}$ must be $\subseteq G$. Sets in $\sigma(\mathcal{V})$ are "bundles of lines". ◇◇

Example 10. $\mathcal{D} = \{\{(x, y) : \ y = mx, \ a \leq m \leq b\} : \ a, b \in \mathbb{R}\}$. Then $\{(0, 0)\} \in \sigma(\mathcal{D})$, but all other sets in $\sigma(\mathcal{D})$ are "bundles of lines", possibly with $(0, 0)$ removed. ◇◇

Example 2 (continued). In addition to $\mathcal{I}$, $\mathcal{S}$ and $\mathcal{B}$ defined previously, let

$\mathcal{I}_+$ all sub-intervals of $[0, 1)$ of any type:

$$(a, b), \ [a, b), \ (a, b], \text{ or } [a, b].$$

$\mathcal{T}$ = all open subsets $A \subseteq [0, 1)$.

$\mathcal{U}$ = all singleton sets $\{x\}$.

Then $\mathcal{I} \subseteq \mathcal{S} \subseteq \sigma(\mathcal{I})$, which implies $\sigma(\mathcal{I}) = \sigma(\mathcal{S}) = \mathcal{B}$. Also $\mathcal{I} \subseteq \mathcal{I}_+ \subseteq \sigma(\mathcal{I})$ and $\mathcal{I} \subseteq \sigma(\mathcal{T}) \subseteq \mathcal{B}$, so $\mathcal{I}$, $\mathcal{I}_+$, $\mathcal{S}$ and $\mathcal{T}$ all generate the Borel sets $\mathcal{B}$. $\mathcal{U} \subseteq \mathcal{B}$ also, but $\sigma(\mathcal{U}) \neq \mathcal{B}$. In fact $\sigma(\mathcal{U}) = \mathcal{C}$ as defined in Example 3.◇◇

We tend to think of $\sigma(\mathcal{A})$ as what we get by starting with the $A \in \mathcal{A}$, then including all sets we can construct from these by compliments, countable unions and intersections, then repeat the process .... But this is false; not all sets in $\sigma(\mathcal{A})$ can be constructed from the $A \in \mathcal{A}$ by such a process. We must rely on more abstract mathematical arguments to establish properties of generated σ-fields.

Consider again the situation described in Example 4. We have defined $P(\cdot)$ for sets in $\mathcal{I}$ by $P([a, b)) = b - a$ and claim that there exists a unique probability measure on the Borel sets $\mathcal{B}$ which extends P. Lets think about the uniqueness part of this assertion. What we are saying is that if P and Q are both probability measures on $[0, 1)$ with the Borel sets $\mathcal{B}$ and if both $P([a, b)) = Q([a, b))$ for all $[a, b) \in \mathcal{I}$, then $P(A) = Q(A)$ for all $A \in \mathcal{B}$. This fits a logical pattern that we encounter frequently in dealing with σ-fields.

Typical Problem. Suppose $\mathcal{A}$ is a class of subsets of Ω and $\mathcal{F} = \sigma(\mathcal{A})$. We know that a certain property (X) holds for all $A \in \mathcal{A}$ and want to show that (X) holds for all $A \in \mathcal{F}$.

For instance, in the uniqueness question above, $\mathcal{A} = \mathcal{I}$ and (X) would be the property that $P(A) = Q(A)$. We will encounter several other problems that fit the same pattern. To understand how to handle a problem of this general pattern, define the class of $\mathcal{F}$-sets determined by (X):

$$\mathcal{L} = \{A \in \mathcal{F} : \ \text{(X) holds for } A\}.$$

The very definition implies $\mathcal{L} \subseteq \mathcal{F}$. The problem is to show $\mathcal{L} = \mathcal{F}$. One way to proceed is to show directly from the description of (X) that $\mathcal{L}$ is itself a σ-field. This is sometimes feasible (see problem 5 for instance), but often not.

Example 11. Let P and Q both be probability measures on $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-field. Define

$$\mathcal{L} = \{A \in \mathcal{F} : \ P(A) = Q(A)\}.$$

We can check that $\mathcal{L}$ has the following properties:

(i) $\Omega \in \mathcal{L}$;

(ii) if $A \in \mathcal{L}$ then $A^c \in \mathcal{L}$;

(iii) if $A_n \in \mathcal{L}$ are disjoint, then $\cup A_n \in \mathcal{L}$.

However $\mathcal{L}$ can fail to be a σ-field; see problem 6. ◇◇

Definition. *A class $\mathcal{L}$ of subsets of Ω is called a* <u>λ-system</u> *if it satisfies (i), (ii) and (iii) of the preceding example. A class $\mathcal{P}$ of subsets of Ω is called a* <u>π-system</u> *if $A \cap B \in \mathcal{P}$ whenever $A, B \in \mathcal{P}$.*

Note that a class $\mathcal{F}$ is a σ-field if and only if it is both a λ-system and a π-system. ($A_n \in \mathcal{F}$ implies $B_n = A_n \cap A_{n-1}^c \cap \cdots \cap A_1^c \in \mathcal{F}$ so that $\cup_1^\infty A_n = \cup_1^\infty B_n \in \mathcal{F}$.) Thus the two definitions separate the properties of σ-fields into two parts. Here is the important theorem.

The π-λ Theorem (A). *If $\mathcal{P}$ is a π-system, $\mathcal{L}$ is a λ-system and $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

Proof: Let $\lambda(\mathcal{P})$ be the smallest λ-system containing $\mathcal{P}$ (the intersection of all λ-systems containing $\mathcal{P}$). We will show that $\lambda(\mathcal{P})$ is also a π-system.

Consider $A \in \mathcal{P}$ and define $\mathcal{G}_A = \{B : \ A \cap B \in \lambda(\mathcal{P})\}$. The following steps show that $\mathcal{G}_A$ is a λ-system:

1) $\mathcal{P} \subseteq \mathcal{G}_A$, since $B \in \mathcal{P}$ implies $A \cap B \in \mathcal{P} \subseteq \lambda(\mathcal{P})$.
2) $\Omega \in \mathcal{G}_A$, since $\Omega \cap A = A \in \mathcal{P} \subseteq \lambda(\mathcal{P})$.
3) Suppose $B \in \mathcal{G}_A$, then $A \in \lambda(\mathcal{P})$ implies $A^c \in \lambda(\mathcal{P})$ and $A \cap B \in \lambda(\mathcal{P})$ are disjoint, so $A^c \cup (A \cap B) \in \lambda(\mathcal{P})$. But $A^c \cup (A \cap B) = A^c \cup B$, so $(A^c \cup B)^c = A \cap B^c \in \lambda(\mathcal{P})$. I.e. $B^c \in \mathcal{G}_A$.
4) Suppose $B_1, B_2, \dots \in \mathcal{G}_A$ and are disjoint. Then $A \cap (\cup_1^\infty B_i) = \cup_1^\infty (A \cap B_i)$ and the $A \cap B_i \in \lambda(\mathcal{P})$ are disjoint. Thus $\cup_1^\infty B_i \in \mathcal{G}_A$.

Since $\mathcal{G}_A$ is a λ-system containing $\mathcal{P}$, we conclude that $\lambda(\mathcal{P}) \subseteq \mathcal{G}_A$. Since this holds for any $A \in \mathcal{P}$, we have shown that $A \cap B \in \lambda(\mathcal{P})$ for all $A \in \mathcal{P}$, $B \in \lambda(\mathcal{P})$. Now consider any $B \in \lambda(\mathcal{P})$ and define $\mathcal{G}_B = \{A : \ A \cap B \in \lambda(\mathcal{P})\}$, and repeat the above sequence of steps.

1) $\mathcal{P} \subseteq \mathcal{G}_B$, by the preceding.
2) $\Omega \in \mathcal{G}_B$, because $\Omega \cap B = B \in \lambda(\mathcal{P})$.
3) Suppose $A \in \mathcal{G}_B$. Then since $B^c \in \lambda(\mathcal{P})$ and $A \cap B \in \lambda(\mathcal{P})$ are disjoint, $B^c \cup (A \cap B) = B^c \cup A = (A^c \cap B)^c \in \lambda(\mathcal{P})$, so $A^c \cap B \in \lambda(\mathcal{P})$. I.e. $A^c \in \mathcal{G}_B$.
4) If $A_1, A_2, \dots \in \mathcal{G}_B$ are disjoint, then $A_i \cap B \in \lambda(\mathcal{P})$ are disjoint, so that $(\cup_1^\infty A_i) \cap B = \cup_1^\infty (A_i \cap B) \in \lambda(\mathcal{P})$. Thus $\cup_1^\infty A_i \in \mathcal{G}_B$.

Thus $\mathcal{G}_B$ is a λ-system containing $\mathcal{P}$. Therefore $\lambda(\mathcal{P}) \subseteq \mathcal{G}_B$, for any $B \in \lambda(\mathcal{P})$. This means that $A \cap B \in \lambda(\mathcal{P})$ for all $A, B \in \lambda(\mathcal{P})$. We have shown therefore that $\lambda(\mathcal{P})$ is a π-system, and therefore is in fact s σ-field. We conclude that $\mathcal{P} \subseteq \sigma(\mathcal{P}) \subseteq \lambda(\mathcal{P}) \subseteq \mathcal{L}$. ∎

Examples 4 and 11 (continued). On $\Omega = [0, 1)$ as before, notice that $\mathcal{I}$ is a π-system. So if P and Q are both probability measures on $(\Omega, \mathcal{B})$, with $P(A) = Q(A)$ for all $A \in \mathcal{I}$ then the theorem tells us $\mathcal{L} = \sigma(\mathcal{I}) = \mathcal{B}$. In other words the class of A for which $P(A) = Q(A)$ includes all of $\mathcal{B}$. Thus there is at most one extension of P from intervals to a probability measure on the Borel sets. ⋄⋄

This example demonstrates how the π-λ Theorem is applied to our typical problem above: if the class $\mathcal{A}$ of sets for which we know property (X) to hold is a π-system, then we only need to verify that $\mathcal{L}$ is a λ-system, not that it is a σ-field. Also note that whether or not $\mathcal{L}$ is a λ-system depends only what the property (X) is, not on $\mathcal{A}$.

If μ and ν are general measures on $(\Omega, \mathcal{F})$ then we cannot always show that

$$\mathcal{L} = \{A \in \mathcal{F} : \ \mu(A) = \nu(A)\}$$

is a λ-system; we can't use subtraction to establish (ii). However for σ-finite measures we can get around this difficulty to prove the following uniqueness result.

Theorem B. *Suppose $\mathcal{P}$ is a π-system and μ and ν are two measures on $\sigma(\mathcal{P})$ which are σ-finite* <u>on</u> <u>$\mathcal{P}$</u>. *If $\mu = \nu$ for all $\mathcal{P}$-sets, then $\mu = \nu$ on all of $\sigma(\mathcal{P})$.*

By "σ-finite on $\mathcal{P}$" we mean $\Omega = \cup_1^\infty P_n$, $P_n \in \mathcal{P}$ with $\mu(P_n), \nu(P_n) < \infty$.

PROOF: In our typical problem take

$$\text{(X)} \qquad \nu(A \cap P_n) = \mu(A \cap P_n) \text{ for all } n.$$

Then (X) holds for all $A \in \mathcal{P}$. Define

$$\mathcal{L} = \{A \in \sigma(\mathcal{P}) : \ \nu(A \cap P_n) = \mu(A \cap P_n) \text{ for all } n\}.$$

(i) $\Omega \in \mathcal{L}$ because $\nu(P_n) = \mu(P_n)$.

(ii) $A \in \mathcal{L}$ implies

$$\begin{aligned}\mu(A^c \cap P_n) &= \mu(P_n) - \mu(P_n \cap A) \\ &= \nu(P_n) - \nu(P_n \cap A) = \nu(A^c \cap P_n),\end{aligned}$$

which implies $A^c \in \mathcal{L}$. (Note that we are using $\mu(P_n) < \infty$ here.)

(iii) $A_k \in \mathcal{L}$ disjoint implies

$$\begin{aligned}\mu((\cup_1^\infty A_k) \cap P_n) &= \mu(\cup_1^\infty (A_k \cap P_n)) \\ &= \sum_{k=1}^\infty \mu(A_k \cap P_n) \\ &= \sum_{k=1}^\infty \nu(A_k \cap P_n) = \cdots = \nu((\cup_1^\infty A_k) \cap P_n),\end{aligned}$$

which implies $\cup_1^\infty A_k \in \mathcal{L}$.

Now the π-λ Theorem implies $\mathcal{L} = \sigma(\mathcal{P})$.

Finally, let $Q_1 = P_1$, $Q_2 = P_2 \setminus P_1, \ldots, Q_n = P_n \setminus \cup_1^{n-1} P_k$. For any $A \in \sigma(\mathcal{P})$, $A \cap Q_n = (A \cap Q_n) \cap P_n$ so $\mu(A \cap Q_n) = \nu(A \cap Q_n)$ and $A = \cup_1^\infty (A \cap Q_n)$, disjoint. Thus,

$$\mu(A) = \sum_1^\infty \mu(A \cap Q_n) = \sum_1^\infty \nu(A \cap Q_n) = \nu(A)$$

∎

There is another, older result which provides another way to deal with our typical problem.

DEFINITION. *A class $\mathcal{M}$ of subsets of Ω is called a* <u>*monotone*</u> *class if*

(i) whenever $A_1 \supseteq A_2 \ldots$ are sets in $\mathcal{M}$ then $\cap A_n \in \mathcal{M}$;

(ii) whenever $A_1 \subseteq A_2 \ldots$ are sets in $\mathcal{M}$ then $\cup A_n \in \mathcal{M}$.

THE MONOTONE CLASS THEOREM (C). *If $\mathcal{F}$ is a field, $\mathcal{M}$ is a monotone class and $\mathcal{F} \subseteq \mathcal{M}$ then $\sigma(\mathcal{F}) \subseteq \mathcal{M}$.*

To summarize if $\mathcal{A} \subseteq \mathcal{G}$ there are several ways we might show that $\sigma(\mathcal{A}) \subseteq \mathcal{G}$. Any of these may be viable approaches to dealing with our typical problem.

- Show that $\mathcal{G}$ is a σ-field. (See problem 5 for instance.)
- Show that $\mathcal{A}$ is a π-system and $\mathcal{G}$ is a λ-system. (See Example 3 above.)
- Show that $\mathcal{A}$ is a field and $\mathcal{G}$ is a monotone class. (See problem 9 for an example.)

The Extension Theorem

We come now to the main theorem on the existence of measures.

Carathéodory Extension Theorem (D). *Suppose μ is a measure on a field $\mathcal{F}$ of subsets of Ω. Then μ has an extension to a measure on $\sigma(\mathcal{F})$.*

If μ is σ-finite on $\mathcal{F}$, then the extension is unique, by Theorem B above.

Application to Distribution Functions. Suppose $F : \mathbb{R} \to \mathbb{R}$ is nondecreasing and right-continuous:

$$F(x) \leq F(y) \text{ whenever } x \leq y,$$

$$\lim_{y \downarrow x} F(y) = F(x) \text{ for all } x.$$

Let $\mathcal{J}$ be the class of all intervals of the form

$$(-\infty, b], \quad (a, b], \quad \text{or} \quad (a, \infty).$$

(We consider $\emptyset = (a, a]$ to be in $\mathcal{J}$.) Define $F(\pm\infty)$ by

$$F(\infty) = \lim_{x \to \infty} F(x) \leq \infty, \quad F(-\infty) = \lim_{x \to -\infty} F(x) \geq -\infty.$$

We can define μ on $\mathcal{J}$ by

$$\begin{aligned} \mu((a, b]) &= F(b) - F(a) \\ \mu((-\infty, b]) &= F(b) - F(-\infty) \\ \mu((a, \infty)) &= F(\infty) - F(a). \end{aligned} \tag{1}$$

These are all values in $[0, \infty]$ by the monotonicity of $F(\cdot)$.

Let $\mathcal{S}$ consist of all finite disjoint unions $\cup_1^N J_n$ of intervals $J_n \in \mathcal{J}$. $\mathcal{S}$ is a field. We claim, and will prove, that

$$\mu(\cup_1^N J_n) = \sum_1^N \mu(J_n)$$

defines a σ-finite measure on $\mathcal{S}$. Once we do, the following theorem will be a consequence of the Carathéodory Theorem.

Theorem E. *Let $F : \mathbb{R} \to \mathbb{R}$ be a nondecreasing, right continuous function as described above. There exists a unique measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the property that $\mu((a, b]) = F(b) - F(a)$.*

This theorem justifies several assertions made in our discussion up to this point. In particular, using $F(x) = x$ we get Lebesgue measure ℓ on the Borel sets, determined by the property that $\ell((a, b]) = b - a$.

Proof: First we will show that μ is countably additive within the individual intervals $\mathcal{J}$. If $-\infty \leq c_0 < c_1 < c_2 < \cdots < c_n \leq \infty$ then clearly

$$\sum_1^n F(c_k) - F(c_{k-1}) = F(c_n) - F(c_0).$$

By dropping some terms from the left we see that if $I_k \in \mathcal{I}$ are disjoint and $\cup_1^n I_k \subseteq I$, $I \in \mathcal{J}$, then $\sum_1^n \mu(I_k) \leq \mu(I)$. Passing to the limit as $n \to \infty$ we see that the same is true for sequences of disjoint $I_k \in \mathcal{J}$: $\cup I_k \subseteq I$, $I \in \mathcal{J}$ implies

$$\sum_1^\infty \mu(I_k) \leq \mu(I). \tag{2}$$

Next suppose $I, I_k \in \mathcal{J}$ are intervals with finite endpoints and that $I \subseteq \cup_1^K I_k$. Let $I = (a,b]$ and $I_k = (a_k, b_k]$. By throwing out some intervals and renumbering we can assume that $a \in (a_1, b_1]$, $b_{k-1} \in (a_k, b_k]$, ending with $b \in (a_K, b_K]$. Then since $a_1 \le a$, $a_k \le b_{k-1}$ and $b \le b_K$ the monotonicity of F implies $F(b_k) - F(b_{k-1}) \le F(b_k) - F(a_k)$, and so

$$\begin{aligned} F(b) - F(a) &\le F(b_K) - F(a_1) \\ &= F(b_1) - F(a_1) + \sum_2^K F(b_k) - F(b_{k-1}) \\ &\le \sum_1^K F(b_k) - F(a_k). \end{aligned}$$

Thus, after adding back to the right side any intervals we threw out and reverting to the original numbering, we have

$$\mu(I) \le \sum_1^K \mu(I_k). \tag{3}$$

We want to generalize this to a sequence of intervals. I.e. we want to prove that $I \subseteq \cup_1^\infty I_k$ with $I, I_k \in \mathcal{J}$ implies

$$\mu(I) \le \sum_1^\infty \mu(I_k). \tag{4}$$

(There is no disjointness assumed here.) To do this, first notice that we can assume I is bounded, because $\mu(I \cap (-n, n]) \uparrow \mu(I)$ as $n \to \infty$. So suppose $I = (a,b]$ with finite endpoints a, b. We can also assume that $I_k = (a_k, b_k]$ has finite endpoints. (Otherwise replace I_k with $I_k' = I_k \cap (a,b]$, which only makes the right side in (4) smaller.) Moreover assume $a < b$ because otherwise there is nothing to prove. Consider any $\epsilon > 0$. Since F is right continuous we can find

$$\begin{aligned} a < a' < b \text{ with } F(a') \le F(a) + \epsilon, \text{ and} \\ b_k < b_k' \text{ with } F(b_k') \le F(b_k) + \epsilon/2^k. \end{aligned}$$

Then

$$[a', b] \subseteq (a,b] \subseteq \cup_1^\infty (a_k, b_k] \subseteq \cup_1^\infty (a_k, b_k').$$

By the Heine-Borel Theorem there exists $K < \infty$ so that

$$[a', b] \subseteq \cup_1^K (a_k, b_k').$$

But then $(a', b] \subseteq \cup_1^K (a_k, b_k']$ so that the finite case (3) implies

$$\mu((a'b]) \le \sum_1^K \mu((a_k, b_k']).$$

Now we can put some pieces together:

$$\begin{aligned}\mu(I) = F(b) - F(a) &\le F(b) - F(a') + \epsilon \\ &\le \epsilon + \sum_1^K F(b_k') - F(a_k) \\ &\le \epsilon + \sum_1^K \epsilon 2^{-k} + F(b_k) - F(a_k) \\ &\le \epsilon + \sum_1^\infty \epsilon 2^{-k} + \sum_1^\infty F(b_k) - F(a_k) \\ &= 2\epsilon + \sum_1^\infty \mu(I_k)\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, the inequality (4) follows.

Taking (2) and (4) together shows that

$$\mu(I) = \sum_1^\infty \mu(I_k), \tag{5}$$

whenever $I_k \in \mathcal{J}$ are disjoint intervals and $I = \cup_1^\infty I_k$ is also an interval in $\mathcal{J}$. This is the countable additivity on $\mathcal{J}$ that was our first goal.

The next step is to extend this to $\mathcal{S}$ by first showing that μ is well-defined on $\mathcal{S}$. That is if we consider two different representations of $A \in \mathcal{S}$ as a disjoint union of intervals

$$\cup_1^N I_n = \cup_1^M J_m = A$$

where $I_n \in \mathcal{J}$ are disjoint and $J_m \in \mathcal{J}$ are disjoint, both lead to the same value of $\mu(A)$. Indeed, for each fixed n the $I_n \cap J_m$ are disjoint intervals as $m = 1, \ldots, M$ and $I_n = \cup_1^M I_n \cap J_m$. Thus (5) implies that

$$\mu(I_n) = \sum_{m=1}^M \mu(I_n \cap J_m).$$

Likewise,

$$\mu(J_m) = \sum_{n=1}^N \mu(I_n \cap J_m).$$

Hence,

$$\begin{aligned}\sum_1^N \mu(I_n) &= \sum_{n=1}^N \sum_{m=1}^M \mu(I_n \cap J_m) \\ &= \sum_{m=1}^M \sum_{n=1}^N \mu(I_n \cap J_m) = \sum_1^M \mu(J_m)\end{aligned}$$

Suppose next that $A = \cup_1^\infty A_k$ with $A, A_k \in \mathcal{S}$, the A_k being disjoint. Then

$$\begin{gathered} A = \cup_1^N I_n \text{ for disjoint } I_n \in \mathcal{J}; \\ A_k = \cup_{j=1}^{m_k} I_{k,j} \text{ for disjoint } I_{k,j} \in \mathcal{J}\end{gathered}$$

Here, since the A_k are disjoint, the "doubly indexed" collection of $I_{k,j}$ is disjoint. Thus (5) implies that for each n,

$$\mu(I_n) = \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} \mu(I_n \cap I_{k,j}).$$

Now we can establish the countable additivity of μ by writing

$$\begin{aligned} \mu(A) &= \sum_{n=1}^{N} \mu(I_n) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} \mu(I_n \cap I_{k,j}) \\ &= \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} \sum_{n=1}^{N} \mu(I_n \cap I_{k,j}) \\ &= \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} \mu(I_{k,j}) \\ &= \sum_{k=1}^{\infty} \mu(A_k). \end{aligned}$$

Theorem E now applies to extend μ to $\mathcal{B} = \sigma(\mathcal{A})$. The proof is finished by noting that μ is σ-finite on $\mathcal{J}$, because $\mu((-n, n]) < \infty$ and $(-n, n] \uparrow \mathbb{R}$. Hence the uniqueness is a consequence of Theorem B. ▮

Higher Dimensions. This can all be generalized to $\Omega = \mathbb{R}^k$. We start with the class $\mathcal{R}$ of all "bounded rectangles", i.e. sets of the form

$$\begin{aligned} J &= \{x \in \mathbb{R}^k : \ a_i < x_i \leq b_i \text{ for each } i = 1, \ldots, k\} \\ &= (a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_k b_k] = \times_1^k (a_i, b_i], \end{aligned}$$

including $\emptyset$. $\mathcal{B}(\mathbb{R}^k) = \sigma(\mathcal{R})$ (or simply $\mathcal{B}$ when $\Omega = \mathbb{R}^k$ is understood) is the Borel σ-field. We could have allowed unbounded rectangles (i.e. allow $a_i = -\infty$, and ∞ for b_i]) to obtain a class more analogous to what we called $\mathcal{J}$ previously. This, and many other possibilities, would all generate the same σ-field $\mathcal{B}$. In particular $\mathcal{B}$ is generated by the collection of all open subsets of $\mathbb{R}^k$.

The idea of a distribution function can be generalized and a higher dimensional version of Theorem E proved; see Billingsley §12. In particular there exists Lebesgue measure $\ell(\cdot)$ on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ determined uniquely by its values on $\mathcal{R}$:

$$\ell(\times_i^k (a_i, b_i]) = \prod_1^k (b_i - a_i).$$

(Note that $\mathcal{R}$ is a π-system.)

Completeness. A measure space $(\Omega, \mathcal{F}, \mu)$ is called *complete* if $A \in \mathcal{F}$ and $\mu(A) = 0$ implies that all subsets $B \subseteq A$ are also in $\mathcal{F}$. Completeness is important in certain aspects of advanced probability theory. It can be shown that starting with any measure space $(\Omega, \mathcal{F}, \mu)$ there exists a (smallest) complete extension or "completion": $(\Omega, \mathcal{F}^+, \mu^+)$ where $\mathcal{F} \subseteq \mathcal{F}^+$ and $\mu^+(A) = \mu(A)$ for all the original $A \in \mathcal{F}$.

It turns out that $(\mathbb{R}^k, \mathcal{B}, \ell)$ (any $k \geq 1$) is <u>not</u> complete. The completed σ-field $\mathcal{B}^+$ is the class of Lebesgue measurable sets. Every Borel measurable set is Lebesgue measurable, but not conversely. There are still sets which are not Lebesgue measurable. ℓ^+ is still called Lebesgue measure. The completed version, $(\mathbb{R}^k, \mathcal{B}^+, \ell^+)$, is the standard measure space used in treatments of real analysis (such as Math 5225).

Existence of a Non-Measurable Set

Problem 8 discusses the translation invariance of Lebesgue measure. This is what we need to present the usual construction of a subset of $\mathbb{R}$ which is not a Borel set. We start by defining $x \sim y$ $(x, y \in \mathbb{R})$ to mean that $x - y$ is a rational number. This is what is called an *equivalence relation.* We need to consider the *equivalence class* of x:

$$C_x = \{y \in \mathbb{R} : \ x \sim y\} = \{x + q : \ q \text{ rational } \}.$$

If $z \in C_u \cap C_v$ then $z = u + p = v + q$ where p, q are rational, so that $u - v = q - p$ is also rational. I.e. $u \sim v$, so $u \in C_v$ and $v \in C_u$. In fact every $y \in C_u$ is also in C_v: $C_u \subseteq C_v$. Thus whenever two equivalence classes intersect they actually coincide. The equivalence classes form a partition of $\mathbb{R}$.

Given an equivalence class C_x, $C_x \cap [0, 1)$ is not empty. (Let q be a rational with $|(\frac{1}{2} - x) - q| < \frac{1}{2}$. Then $x + q \in C_x$ and $|(x + q) - \frac{1}{2}| < \frac{1}{2}$ so that $x + q \in [0, 1)$.) For each equivalence class C pick exactly one $h \in C \cap [0, 1)$ and let H be the set of h so chosen. We will show that $H \notin \mathcal{B}(\mathbb{R})$. To do this we will use the following translates of H, mod 1. For $r \in [0, 1)$ define

$$\begin{aligned} H_r &= \{y \in [0, 1) : \ y = h + r \text{ mod } 1, \text{ some } h \in H\} \\ &= (H \cap [0, 1 - r)) + r \quad \bigcup \quad (H \cap [1 - r, 1)) + (r - 1). \end{aligned}$$

Since ℓ is translation invariant (Problem 8) we see that if $H \in \mathcal{B}(\mathbb{R})$ the $\ell(H) = \ell(H_r)$, and $0 \leq \ell(H) \leq 1$.

Let $\{r_1, r_2, r_3, \dots\}$ be an enumeration of the rational numbers in $[0, 1)$. The key facts are that the H_{r_i} are disjoint and $[0, 1) = \cup_1^\infty H_{r_i}$. Then if H were in $\mathcal{B}(\mathbb{R})$ we would have

$$\ell([0, 1)) = 1 = \sum_1^\infty \ell(H_{r_i}) = \sum_1^\infty \ell(H),$$

which would be 0 if $\ell(H) = 0$ or ∞ if $\ell(H) > 0$. Either way we have a contradiction! To finish, we need to check the two key facts.

To see that the H_{r_i} are disjoint, suppose that $z \in H_{r_i} \cap H_{r_j}$ for some $r_i \neq r_j$. Then there would be $h_i, h_j \in H$ so that $z - h_i =$ either r_i or $r_i - 1$ and likewise $z - h_j = r_j$ or $r_j - 1$. Thus $h_i - h_j$ is rational, so that $h_i, h_j \in H$ come from the same equivalence class. By construction of H we deduce that $h_i = h_j$. This means that either $r_i = r_j$ (not possible since we assumed otherwise) or they differ by ± 1 (not possible since both are in $[0, 1)$). Thus the H_{r_i} are disjoint.

Lastly, given any $x \in [0, 1)$ there is $h \in H$ with $x \sim h$. Thus $x = h + q$ for some rational q, which in fact must be $-1 < q < 1$ since both $x, h \in [0, 1)$. If $q \geq 0$ then $x \in H_q$. If $-1 < q < 0$ then $0 < 1 + q < 1$ and $x \in H_{1+q}$. Either way, $x \in \cup_1^\infty H_{r_i}$. Thus $[0, 1) = \cup_1^\infty H_{r_i}$, as claimed.

Applications to Independence

Working with the concept of independence illustrates the usefulness of the π–λ Theorem. We assume that $(\Omega, \mathcal{F}, P)$ is a probability space and all sets referred to are $\mathcal{F}$-sets.

Definition. *Two sets A and B are independent if $P(A \cap B) = P(A)P(B)$. We say the sets of a an indexed list or family $\{A_\theta : \ \theta \in \Theta\}$ are independent (of each other) if for every selection of a finite number of them, $A_{\theta_1}, \dots, A_{\theta_n}$ with the θ_i distinct, we have*

$$P(\cap_1^n A_{\theta_i}) = P(A_{\theta_1}) \cdots P(A_{\theta_n}).$$

In this definition we allow duplicates among the A_θ. This is so that we can talk about sets being independent of themselves. For instance, the 3 sets $\Omega, \Omega, \emptyset$ are independent. In the notation of the definition we could list these as $A_1 = \Omega$, $A_2 = \Omega$, $A_3 = \emptyset$; $\Theta = \{1, 2, 3\}$ and say that the sets of $\{A_1, A_2, A_3\}$ are independent. On the other hand if $P(H) = \frac{1}{2}$ then H is not independent of itself. Thus if $B_1 = \Omega$, $B_2 = \emptyset$, $B_3 = H$, $B_4 = H$ it would be false to say that the sets of $\{B_\theta : \ \theta \in \{1, 2, 3, 4\}\}$ are independent because the definition would require $P(B_3 \cap B_4) = P(B_3)P(B_4)$, which is false. However $\{B_\theta, \theta \in \{1, 2, 3\}\}$ are independent because the definition's restriction to distinct θ_i prevents us from considering $H \cap H$. Thus the definition allows duplicates in the collection of sets A_θ provided they occur with different indices $\theta \in \Theta$.

This idea generalizes from individual sets to classes of sets.

Definition. *Let $(\Omega, \mathcal{F}, P)$ be a probability space.*

1) *Two classes $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ are called independent if A, B are independent for any choice of $A \in \mathcal{G}, B \in \mathcal{H}$.*
2) *Suppose for each index $\theta \in \Theta$ we have a class $\mathcal{A}_\theta$ of $\mathcal{F}$-sets. We say that the classes $\mathcal{A}_\theta, \theta \in \Theta$ are independent if for each selection of a finite number of distinct indices $\theta_1, \ldots, \theta_n \in \Theta$ and each choice of sets $A_{\theta_i} \in \mathcal{A}_{\theta_i}$, $i = 1, \ldots, n$ the sets of $\{A_{\theta_1}, \ldots, A_{\theta_1}\}$ are independent.*

Theorem F. *If the classes $\mathcal{A}_1, \cdots, \mathcal{A}_n$ are independent, and each is a π-system, then the generated σ-fields*

$$\sigma(\mathcal{A}_1) \cdots, \sigma(\mathcal{A}_n)$$

are independent.

Proof: Observe (for $n = 2$) that the collection $\mathcal{L}$ of those $A \in \sigma(\mathcal{A}_1)$ which are independent of $B \in \mathcal{A}_2$ is a λ-system. (Apply this idea repeatedly.) ▮

Example 12. For any individual set A, $\mathcal{A} = \{A\}$ is a π-system. Thus if $A_\theta, \theta \in \Theta$ are independent then $\sigma(A_\theta) = \{\emptyset, \Omega, A_\theta, A_\theta^c\}$ over $\theta \in \Theta$ are independent. Compare this with problem 12 below.

Corollary G. *Suppose $\mathcal{A}_\theta, \theta \in \Theta$ is a collection of independent π-systems. Suppose Θ_1 and Θ_2 are disjoint subsets of the index set Θ. Then $\sigma(\cup_{\theta \in \Theta_1} \mathcal{A}_\theta)$ and $\sigma(\cup_{\theta \in \Theta_1} \mathcal{A}_\theta)$ are independent. (This extends to any partition $\Theta = \cup_{\lambda \in \Lambda} \Theta_\lambda$ of Θ.)*

Proof: Just note that the class $\mathcal{A}_1^*$ of finite intersections of sets from $\cup_{\Theta_1} \mathcal{A}_\theta$ is a π-system and is independent of $\mathcal{A}_2^*$ (defined analogously). ▮

Example 13. Let $\Omega = (0, 1]$ and $d_n(\cdot)$ the digits of decimal expansion, as in Unit M. The Borel sets $\mathcal{B}$ can be described as the σ-field generated by the sets $\{w : d_n(w) = k\}$ using all $n = 1, 2, \ldots$ and $k = 0, 1, \ldots, 9$. Consider

- Let $\mathcal{F}_{\text{even}}$ be the σ-field generated by the sets $\{w : d_n(w) = k\}$ using just the even n.
- Let $\mathcal{F}_{\text{odd}}$ the σ-field generated by the same sets, but using only the odd n.

Then $\mathcal{F}_{\text{even}}$ and $\mathcal{F}_{\text{odd}}$ are independent. (Notice that $\mathcal{A}_n = \{\emptyset, \{d_n = 0\}, \ldots, \{d_n = 9\}\}$ is a π-system for each n.) ◇◇

Tail Events

Suppose $A_1, A_2, \cdots$ is a sequence of sets. Consider the following sets

$$\limsup A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = \{\omega : \ \omega \in A_k \text{ for infinitely many } k\} = \text{“}\{A_k \text{ i.o. }\}\text{”}$$
$$\liminf A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k = \{\omega : \ \omega \in A_k \text{ for all but a finite number of } k\}.$$

If $\limsup A_n = \liminf A_n$, then sometimes we say “$A_n \to A$”.

Example 14. Consider $\limsup A_n$ and $\liminf A_n$ for $A_n = \{\omega : d_n(\omega) = 6\}$. ◇◇

Notice, if $x \in \liminf A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$ then $x \in \cap_{k=n'}^{\infty} A_k$ for some n', i.e.

$$x \in A_k \text{ all } k \geq n'.$$

Therefore

$$x \in \cup_{k=n}^{\infty} A_k \text{ for } \underline{\text{all}}\ n.$$

I.e. $x \in \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = \limsup A_n$. Thus

$$\liminf A_n \subseteq \limsup A_n,$$

as is clear from the alternate descriptions.

Recall the notions of lim inf and lim sup for sequences $\{a_1, a_2, \cdots\}$ of real numbers:

$$\liminf a_n = \lim_{n \to \infty} (\inf\{a_n : \ k \geq n\}).$$

(The limit always exists if we allow $\pm\infty$.) Similarly,

$$\limsup a_n = \lim_{n \to \infty} (\sup\{a_k : \ k \geq n\}).$$

We always have $\liminf a_n \leq \limsup a_n$. Equality holds (with a finite value) if and only if $\lim_{n\to\infty} a_n$ exists (in which case $\lim a_n = \limsup a_n = \liminf a_n$).

Theorem H. *$P(\liminf A_n) \leq \liminf P(A_n) \leq \limsup P(A_n) \leq P(\limsup A_n)$. If $A_n \to A$, then $P(A_n) \to P(A)$*

Proof: $P(\liminf A_n) = \lim_{n\to\infty} P(\cap_n^{\infty} A_k) \leq \lim_n \inf_{k \geq n} P(A_k)$, and similarly for the other half. ∎

The Borel-Cantelli Lemmas are concerned with $P(\limsup A_n)$

First Borel-Cantelli Lemma (I). *If $\sum_1^{\infty} P(A_n)$ converges, then $P(\limsup A_n) = 0$.*

Proof: Since $\limsup A_n \subseteq \cup_{k=n} A_k$, for each n, we have

$$P(\limsup A_n) \leq \sum_{k=n}^{\infty} P(A_k) \to 0 \text{ as } n \to \infty.$$

∎

Second Borel-Cantelli Lemma (J). *Suppose $\{A_n\}_1^{\infty}$ is a sequence of <u>independent</u> sets. If $\sum_1^{\infty} P(A_n)$ diverges, then $P(\limsup \ A_n) = 1$.*

Proof: We will show $P((\limsup A_n)^c) = 0$. Since

$$(\cap_{k=n}^{\infty} A_k^c) \uparrow \cup_1^{\infty} \cap_{k=n}^{\infty} A_k^c = (\limsup A_n)^c,$$

we know that $P(\cap_{k=n}^{\infty} A_k^c) \uparrow P((\limsup A_n)^c)$. It will suffice therefore for us to show that $P((\cap_{k=n}^{\infty} A_k^c) = 0$ for each n. Since the A_n^c are independent and $1 - x \leq e^{-x}$,

$$\begin{aligned} P(\cap_{k=n}^{\infty} A_k^c) &= \lim_{N\to\infty} P(\cap_{k=n}^{N} A_k^c) \\ &= \lim_{N\to\infty} \Pi_{k=n}^{N} (1 - P(A_k)) \\ &\leq \lim_{N\to\infty} e^{-\sum_{k=n}^{N} P(A_k)}. \end{aligned}$$

Since the series diverges, $\sum_{k=n}^{N} P(A_k) \to +\infty$ as $N \to \infty$. Therefore, $P(\cap_{k=n}^{\infty} A_k^c) = 0$. ▮

Given a sequence $A_1, A_2, \cdots$ of $\mathcal{F}$-sets, $\limsup A_n$ and $\liminf A_n$ are examples of sets which depend on the A_n only as $n \to \infty$. We can define a special σ-field consisting of all such events, called the "tail σ-field" associated with the sequence $\{A_n\}_1^{\infty}$:

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

In particular, $\limsup A_n, \liminf A_n \in \mathcal{T}$.

The Kolmogorov 0−1 Law (K). *If $\{A_n\}_1^{\infty}$ is a sequence of independent events and A is in the associated tail σ-field, then $P(A) = 0$ or 1.*

Proof: Corollary G implies that

$$\sigma(A_1, A_2, \ldots, A_n) \text{ and } \sigma(A_{n+1}, \ldots)$$

are independent, for each n. Since $\mathcal{T}$ is contained in the second of these, the following are independent

$$\cup_1^{\infty} \sigma(A_1, A_2, \ldots, A_n) \text{ and } \mathcal{T}.$$

But since the left is a π-system,

$$\sigma(A_1, A_2, \ldots) \text{ and } \mathcal{T}$$

are also independent. Since $\mathcal{T}$ is contained in the first of these, we find that $\mathcal{T}$ must be independent of itself, and so for any $A \in \mathcal{T}$, A must be independent of itself: $P(A) = P(A)^2$. Hence $P(A)$ must be either 0 or 1. ▮

Problem **1** ....

a) Let $\Omega = \{1, 2, 3, \ldots\}$ and, for each n, let $\mathcal{F}_n$ consist those $A \subseteq \Omega$ with the property that $m' \in A$ for some $m' \geq n$ implies $m \in A$ for all $m \geq n$. Show that each $\mathcal{F}_n$ is a field and that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$. Is $\mathcal{F}_n$ a σ- field ?

b) For any Ω, if $\mathcal{F}_1 \subseteq \ldots \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \ldots$ are all fields, show that $\cup_1^{\infty} \mathcal{F}_n$ is also a field. Give an example to show that even if the $\mathcal{F}_n$ are all σ-fields $\cup_1^{\infty} \mathcal{F}_n$ may fail to be a σ-field.

c) If $\mathcal{F}_n$ are all σ-fields (no assumption of order), show that $\cap_1^{\infty} \mathcal{F}_n$ is also a σ-field.

Problem **2** ....

a) Let

$$\mathcal{F} = \{A \subseteq \Omega : \text{ either } A \text{ or } A^c \text{ is finite } \}.$$

Show that $\mathcal{F}$ is a field, but is a σ-field if and only if Ω is finite.

b) Let $\mathcal{C}$ be as in Example 3. Show $\mathcal{C}$ is a σ- field. If $\Omega = \mathbb{R}$ find an $A \notin \mathcal{C}$.

Problem **3** ....

Let $\Omega = [0, 1)$ and take $\mathcal{S}$ as defined in Example 2. Define ρ on $\mathcal{S}$ by

$$\rho(A) = \begin{cases} 1 & \text{if } [\frac{1}{2} - \epsilon, \frac{1}{2}) \subseteq A \text{ for some } \epsilon > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Show that ρ is finitely but not countably additive.

Problem **4** . . . .

Given $B \subseteq \Omega$ show that

$$\mathcal{G}_B = \{G \subseteq \Omega : \text{ either } B \subseteq G \text{ or } B \subseteq G^c\}$$

is a σ-field. As a consequence show that if $C \in \sigma(\mathcal{A})$, and $\omega, \omega' \in \Omega$ with

$$\omega \in C \text{ and } \omega' \in C^c$$

then there must exist $A \in \mathcal{A}$ with either

$$\omega \in A \text{ and } \omega' \in A^c \quad \text{or} \quad \omega \in A^c \text{ and } \omega' \in A.$$

Problem **5** . . . .

We have defined $\mathcal{B}([0,1))$ and $\mathcal{B}(\mathbb{R})$ separately, but show that

$$A \in \mathcal{B}([0,1)) \text{ if and only if } A \subseteq [0,1) \text{ and } A \in \mathcal{B}(\mathbb{R}).$$

Set up each half of the argument following the pattern of our typical problem, and show that (in both cases) the resulting $\mathcal{L}$ is a σ-field. [Hints: for the "only if" part you might take (X) to be the property that

$$A \subseteq [0,1) \text{ and } A \in \mathcal{B}(\mathbb{R}),$$

while (X) could be

$$A \cap [0,1) \in \mathcal{B}([0,1)).$$

for the "if" part. You need to be careful about what Ω is – does A^c mean $\mathbb{R} \setminus A$ or $[0,1) \setminus A$?]

Problem **6** . . . .

Give an example of Ω, a σ-field $\mathcal{F}$ on Ω, and two probability measures P and Q on $\mathcal{F}$ for which

$$\mathcal{L} = \{A \in \mathcal{F} : P(A) = Q(A)\}$$

fails to be a σ-field. [Hint: try $\Omega = \{a, b, c, d\}$ and $\mathcal{F}$ = all subsets.]

Problem **7** . . . .

Let N be a fixed positive integer. Suppose P is a probability measure on $(\mathbb{R}, \mathcal{B})$ with the property that for every interval $(a,b]$, $P((a,b])$ is a multiple of $1/N$. Show that $P(A)$ must be a multiple of $1/N$ for all $A \in \mathcal{B}$.

Problem **8** . . . .

Suppose λ is a measure on $(\mathbb{R}, \mathcal{B})$ which is *translation invariant*:

$$\lambda(A + x) = \lambda(A) \quad \text{for all } x \in \mathbb{R},\ A \in \mathcal{B}.$$

(Here $A + x = \{a + x : \ a \in A\}$.) If $\lambda((0,1]) < \infty$ show that $\lambda(\cdot) = \alpha\ell(\cdot)$ for some constant α. [Hints: What must the value of α be? First consider all intervals of length 1, then length 1/2, then length 1/4, ...]

Problem **9** . . . .

Suppose μ and ν are finite measures on $(\Omega, \mathcal{F})$ and $\mathcal{F} = \sigma(\mathcal{G})$ where $\mathcal{G}$ is a field. Suppose $\nu(G) \leq \mu(G)$ for all $G \in \mathcal{G}$. Show that $\nu(A) \leq \mu(A)$ for all $A \in \mathcal{F}$. [Hint: set up following our typical problem and apply the Monotone Class Theorem.]

Problem **10** ....

Let P be a probability measure on $(\mathbb{R}, \mathcal{B})$. Show that P must be *tight*; i.e. for every $A \in \mathcal{B}$ and $\epsilon > 0$ there is a compact set $K \subseteq A$ with $P(A) - P(K) \leq \epsilon$. [Hint: the class of A for which this holds can be shown (with some effort) to be a monotone class.]

Problem **11** ....

Consider $(\mathbb{R}, \mathcal{F}, \nu)$ where $\mathcal{F} = \mathcal{B}$ (the Borel sets) and ν is counting measure on the integers:

$$\nu(A) = \text{ the number of integers } k \in A.$$

Suppose $(\mathbb{R}, \mathcal{F}^+, \nu^+)$ is the completion. Describe the sets in $\mathcal{F}^+$.

Problem **12** ....

Show directly that if A,B,C are independent, then A^c,B,C are independent.

Problem **13** ....

Suppose that A, B and C are events such that A and B are independent, B and C are independent and A and C are independent. Are A, B and C necessarily independent? Prove or give a counterexample.

Problem **14** ....

Show that $P(A \cap B \cap C) = P(A)P(B)P(C)$ alone does not imply that A, B and C are independent. However show that $P(A \cap B \cap C) = P(A)P(B)P(C)$ for all $A \in \mathcal{A}$, $B \in \mathcal{B}$ and $C \in \mathcal{C}$ does imply that $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are independent provided $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are fields.

Problem **15** ....

Suppose that A_1, A_2, ... , A_N are independent sets. Show that A_1^c, A_2^c, ... , A_N^c are independent sets.

Problem **16** ....

Suppose $P(\limsup A_n) = 1$ and $P(\liminf B_n) = 1$. Show that $P(\limsup A_n \cap B_n) = 1$. However if the requirement on the B_n is weakened to $P(\limsup B_n) = 1$ then give an example to show that $P(\limsup A_n \cap B_n)$ can be < 1.

Problem **17** ....

Suppose that $A_1, A_2, \ldots$ is a sequence of independent sets, each with $P(A_n) = 1/2$. Define

$$Y_n(\omega) = \begin{cases} 1 & \text{if } \omega \in A_n \\ -1 & \text{if } \omega \notin A_n. \end{cases}$$

The Central Limit Theorem says that the distribution of

$$N^{-1/2} \sum_{n=1}^{N} Y_n(\omega)$$

converges (as $N \to \infty$) to the standard normal distribution, which gives probability 1/2 to the values < 0. However show that the Kolmogorov 0-1 Law implies

$$P\left(\{\omega : \lim_{N\to\infty} N^{-1/2} \sum_{n=1}^{N} Y_n(\omega) < 0\}\right) \neq 1/2.$$

(Does this surprise you? In fact the above probability is $= 0$, though you are not being asked to show that.)

Unit II .... **Random Variables and Measurable Functions**

Suppose $(\Omega, \mathcal{F}, P)$ is a a probability space. A random variable X is a quantity whose value is determined by the specification of an $\omega \subset \Omega$. In other words X is a function, $X : \Omega \to \mathbb{R}$ We want to be able to calculate quantities such as

$$P(X = a) = P(\{\omega : \ X(\omega) = a\})$$
$$P(-2 \le X < 7) = P(\{\omega : \ X(\omega) \in [-2, 7)\})$$

For these sets of ω to be in $\mathcal{F}$ there needs to be some compatibility between X and $\mathcal{F}$.

DEFINITION. *Given a measurable space $(\Omega, \mathcal{F})$, a function $X : \Omega \to \mathbb{R}$ is called a random variable if for every $x \in \mathbb{R}$, $\{\omega \in \Omega : \ X(\omega) \le x\} \in \mathcal{F}$.*

Example 1. Let $\Omega = [0, 1)$, $\mathcal{F} = \mathcal{B}$ the Borel sets and $P = \ell$ (Lebesgue measure). The digits of decimal expansion $d_n(\cdot)$ of Unit M are random variables, because $\{\omega : \ d_n(\omega) \le x\}$ is a finite union of intervals, i.e. a set in $\mathcal{S} \subseteq \mathcal{B}$. ◇◇

Notation. If $A \subseteq \Omega$ its *indicator function* is

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

(Other common notations are $I_A(\cdot)$ and $\chi_A(\cdot)$.) 1_A is a random variable if and only if $A \in \mathcal{F}$.

Measurable Mappings

Notation. If $T : \Omega \to \Gamma$ is a mapping and $B \subseteq \Gamma$, the notation $T^{-1}B$ refers to the following subset of Ω:

$$T^{-1}B = \{\omega \in \Omega : \ T(\omega) \in B\}.$$

(There is no presumption here that T^{-1} exists as an inverse function. For instance if $T(\omega) = 0$ for all ω then $T^{-1}\{0\} = \Omega$.)

The definition of random variable above says

$$X^{-1}(-\infty, x] \in \mathcal{F} \quad \text{for every } x \in \mathbb{R}.$$

It is simple to check that for any $X : \Omega \to \mathbb{R}$ the class

$$\mathcal{G} = \{B \subseteq \mathbb{R} : \ X^{-1}B \in \mathcal{F}\}$$

is a σ-field – see problem 1. Let $\mathcal{A} = \{(-\infty, x] : \ x \in \mathbb{R}\}$. The definition of X being a random variable says $\mathcal{A} \subseteq \mathcal{G}$. Therefore $\sigma(\mathcal{A}) \subseteq \mathcal{G}$. But notice that $\sigma(\mathcal{A}) = \mathcal{B}$, the Borel sets in $\mathbb{R}$. Thus if X is a random variable, then $X^{-1}B \in \mathcal{F}$ for all Borel sets $B \in \mathcal{B}$. Therefore $P(X \in B)$ is defined for every Borel set. This shows that our definition of random variable is a special case of the following more general concept.

DEFINITION. *If $(\Omega, \mathcal{F})$ and $(\Gamma, \mathcal{H})$ are two measurable spaces, then $T : \Omega \to \Gamma$ is called $\mathcal{F}/\mathcal{H}$ measurable if $T^{-1}A \in \mathcal{F}$ whenever $A \in \mathcal{H}$.*

Thus a random variable X is an $\mathcal{F}/\mathcal{B}$ measurable map $X : \Omega \to \mathbb{R}$.

Theorem A. *Suppose $(\Omega, \mathcal{F})$ and $(\Gamma, \mathcal{H})$ are measurable spaces, and $T : \Omega \to \Gamma$.*

1) If $\mathcal{H} = \sigma(\mathcal{A})$ and $T^{-1}A \in \mathcal{F}$ for every $A \in \mathcal{A}$, then T is $\mathcal{F}/\mathcal{H}$ measurable .

2) If T is $\mathcal{F}/\mathcal{H}$ measurable, $(\Theta, \mathcal{M})$ is another measurable space and $S : \Gamma \to \Theta$ is $\mathcal{H}/\mathcal{M}$ measurable, then $S \circ T$ is $\mathcal{F}/\mathcal{M}$ measurable.

Lebesgue vs. Borel. Recall that associated with Lebesgue measure ℓ on $\mathbb{R}$, there is a larger σ-field $\mathcal{B}^+$ called the Lebesgue measurable sets. So there is more than one type of measurability for $f : \mathbb{R} \to \mathbb{R}$ (or $f : \mathbb{R}^d \to \mathbb{R}^r$, though we assume $d = r = 1$ in this discussion). The two most common are $\mathcal{B}/\mathcal{B}$ (Borel) and $\mathcal{B}^+/\mathcal{B}$ (Lebesgue).

- f is Borel measurable if $f^{-1}(-\infty, x] \in \mathcal{B}$ for all x.
- f is Lebesgue measurable if $f^{-1}(-\infty, x] \in \mathcal{B}^+$ for all x.
- If f is Borel measurable, then f is Lebesgue measurable, since $\mathcal{B} \subseteq \mathcal{B}^+$. But not conversely.
- If f and g are both Borel measurable, then $f \circ g$ is also Borel measurable.
- If f and g are both Lebesgue measurable, $f \circ g$ may fail to be Lebesgue measurable.

Treatments of real analysis typically use Lebesgue measurability as the standard, but the last point above makes this choice inconvenient for other settings. Our convention will be that on $\mathbb{R}$ (or $\mathbb{R}^k$) we will always assume the Borel σ-field $\mathcal{B}$ is intended, unless otherwise specified.

Vectors vs. Components. Suppose we have several $f_i = \Omega \to \mathbb{R}$, $i = 1, \ldots, k$. We can consider them individually, and ask that they each be measurable $\mathcal{F}/\mathcal{B}(\mathbb{R})$: $f_i^{-1}(a, b] \in \mathcal{F}$ all $a < b$. Or we can view them as the coordinates of a single "vector-valued" $f = (f_1, \ldots, f_k)$, so that $f : \Omega \to \mathbb{R}^k$, and ask that f be $\mathcal{F}/\mathcal{B}(\mathbb{R}^k)$ measurable. How do these two notions compare?

Suppose first that each f_i is measurable. Consider any bounded rectangle $J \subseteq \mathbb{R}^k$ in $\mathcal{R}$, $J = \times_1^k (a_i, b_i]$. The measurability of each f_i implies that each $f_i^{-1}(a_i, b_i] \in \mathcal{F}$. Hence

$$f^{-1}J = \cap_1^k f_i^{-1}(a_i, b_i] \in \mathcal{F}.$$

Since $\mathcal{R}$ generates $\mathcal{B}(\mathbb{R}^k)$, it follows that f is $\mathcal{F}/\mathcal{B}(\mathbb{R}^k)$ measurable.

On the other hand, if f is measurable in the vector sense then for any i and $c \in \mathbb{R}$, let

$$A = \{(x_1, \ldots, x_k) \in \mathbb{R}^k : \; x_i \le c\}.$$

Since $A \in \mathcal{B}(\mathbb{R}^k)$ we know that $f^{-1}A \in \mathcal{F}$. But

$$f^{-1}A = \{\omega \in \Omega : \; f_i(\omega) \le c\} = f_i^{-1}(-\infty, c].$$

Since this is in $\mathcal{F}$ for all c, we conclude that f_i is measurable, for each i.

Thus a "vector-valued" function $f(\cdot) = (f_1(\cdot), \ldots, f_k(\cdot))$ is $\mathcal{B}(\mathbb{R}^k)$ measurable if and only if each of the component functions $f_i(\cdot)$ is $\mathcal{B}(\mathbb{R})$ measurable. If the underlying space on which f and the f_i are defined is a probability space $(\Omega, \mathcal{F}, P)$ we call f a *random vector*.

Sufficient Conditions for Measurability

Extended Real Numbers. The discussion of limits in Theorem D below is streamlined by using the extended real numbers $\mathbb{R}_\infty = [-\infty, \infty]$, as described in Unit S. The σ-field of Borel sets in $\mathbb{R}_\infty$ can be defined as $\mathcal{B}(\mathbb{R}_\infty) = \sigma(\mathcal{J}_\infty)$ where $\mathcal{J}_\infty$ is the class of all intervals $(a, b]$, $-\infty \le a \le b \le \infty$. Functions $f : \Omega \to \mathbb{R}_\infty$ are just like ordinary functions except that we allow the values $f(\omega) = \pm\infty$. Here are some facts relating $\mathcal{B}(\mathbb{R}_\infty)$ measurability to $\mathcal{B}(\mathbb{R})$ measurability. Of course $(\Omega, \mathcal{F})$ is assumed to be a measurable space.

- $A \in \mathcal{B}(\mathbb{R}_\infty)$ if and only if $A \cap \mathbb{R} \in \mathcal{B}(\mathbb{R})$.
- $f : \Omega \to \mathbb{R}_\infty$ is $\mathcal{F}/\mathcal{B}(\mathbb{R}_\infty)$ measurable if and only if the sets $\{\omega :\ f(\omega) = -\infty\}$ and $\{\omega : f(\omega) = \infty\}$ are in $\mathcal{F}$ and $f^{-1}B \in \mathcal{F}$ for all Borel subsets $B \subseteq \mathbb{R}$ of the usual real numbers, $B \in \mathcal{B}(\mathbb{R})$.

Note that $\mathcal{B}(\mathbb{R}_\infty) = \sigma(\mathcal{A})$ where $\mathcal{A}$ consists of all $[-\infty, x]$, $x \in \mathbb{R}$ (finite). Thus $f : \Omega \to \mathbb{R}_\infty$ is measurable if and only if $f^{-1}[-\infty, x] = \{\omega :\ f(\omega) \le x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.

The next several results make verifying measurability rather easy for most functions we want to work with.

Theorem B. *If $f : \mathbb{R}^d \to \mathbb{R}^r$ is continuous, then it is measurable (Borel).*

Corollary C. *If $f_i : \Omega \to \mathbb{R}$ is $\mathcal{F}/\mathcal{B}$ measurable for each $i = 1, \ldots, k$ and $g : \mathbb{R}^k \to \mathbb{R}$ is continuous, then*

$$g(f_1(\omega), \ldots, f_k(\omega))$$

is measurable.

Proofs: ... $\blacksquare$

Example 2. If f_i are measurable, then $\sin(f_1(\omega))e^{-2\frac{f_2(\omega)\wedge f_3(\omega)}{(f_4(\omega))^2+1}}$ is measurable. ◇◇

Theorem D. *Suppose $f_n : \Omega \to \mathbb{R}_\infty$ is a sequence of $\mathcal{F}$ measurable functions.*

1) Each of the functions

$$\sup_n f_n, \quad \inf_n f_n, \quad \liminf f_n, \quad \limsup f_n$$

are also $\mathcal{F}$ measurable.

2) $L = \{\omega :\ \lim f_n(\omega)$ converges $\} \in \mathcal{F}$ and $1_L \cdot \lim f_n$ is measurable.

3) If g is any other measurable function defined on Ω, then

$$\{\omega :\ g(\omega) = \lim f_n(\omega)\} \in \mathcal{F}.$$

In 2) we are interpreting "$\lim f_n$ converges" is the strict sense: finite values only.

Proof:

Part 1): Since $\sup_n a_n \le x$ if and only if $a_n \le x$ for all n, we can write

$$\{\omega :\ \sup_n f_n(\omega) \le x\} = \cap_n \{\omega :\ f_n(x) \le x\} \in \mathcal{F}.$$

Since $\inf_n a_n = -\sup_n(-a_n)$, $\inf f_n = -\sup(-f_n)$. If f_n is measurable then $-f_n$ is also measurable (because $g(x) = -x$ is continuous). Hence $-\sup(-f_n) = \inf f_n$ is measurable. Next notice that

$$\limsup a_n = \lim_{n\to\infty} [\sup_{k\ge n} a_k] = \inf_n [\sup_{k\ge n} a_k],$$

because $A_n = \sup_{k\ge n} a_k$ is nonincreasing: $A_{n+1} \le A_n$. So we can conclude that

$$\limsup f_n = \inf_n [\sup_{k\ge n} f_k]$$

is measurable, and similarly for $\liminf f_n = \sup_n [\inf_{k\ge n} f_k]$.

Part 2): First observe that if $g : \Omega \to \mathbb{R}_\infty$ is measurable and $B \in \mathcal{F}$, then

$$1_B(\omega)g(\omega) = \begin{cases} g(\omega) & \text{if } \omega \notin B \\ 0 & \text{if } \omega \in B \end{cases} \quad \text{is measurable.}$$

(Note that our convention of $0 \cdot \infty = 0$ is convenient here.) To check this, for any $A \in \mathcal{B}(\mathbb{R}_\infty)$ we have

$$(1_B g)^{-1}A = \big[(g^{-1}A) \cap B\big] \cup \begin{cases} B^c & \text{if } 0 \in A \\ \emptyset & \text{if } 0 \notin A \end{cases}.$$

Now we want to show that the following set is measurable:

$$L = \{\omega : \ \liminf f_n(\omega) = \limsup f_n(\omega) \text{ and are finite}\}.$$

It would be nice if we could simply let $h(\omega) = \liminf f_n(\omega) - \limsup f_n(\omega)$ and then say $L = h^{-1}\{0\}$. The difficulty with this is that $h(\omega)$ may be undefined ($\infty - \infty$). To get around this, let

$$A = \{\liminf f_n = \pm\infty\} \cup \{\limsup f_n = \pm\infty\}$$

and define

$$F(\omega) = \begin{cases} \liminf \ f_n(\omega) & \text{if } \omega \notin A \\ -1 & \text{if } \omega \in A \end{cases} \quad G(\omega) = \begin{cases} \limsup \ f_n(\omega) & \text{if } \omega \notin A \\ +1 & \text{if } \omega \in A \end{cases}.$$

These are measurable, and finite valued. Therefore $F - G$ is defined and measurable, and we can conclude that

$$L = \{\omega : \ F(\omega) = G(\omega)\} = \{\omega : \ F(\omega) - G(\omega) = 0\}$$

is indeed measurable. To finish the proof of 2) notice that $1_L \cdot \lim f_n = 1_L \cdot \liminf f_n$.
Part 3): We can show that $G = \{\liminf f_n = g \text{ both finite }\} \in \mathcal{F}$, by the same technique as in 2). Now observe that

$$\{\omega : \ \lim f_n(\omega) = g(\omega)\} = L \cap G.$$

∎

Example 3.. (Refer to Problem I.15 for notation.) Let $\mathcal{F} = \sigma(\{A_1, A_2, \dots\})$ and

$$C = \{\omega : \ \lim \frac{1}{\sqrt{N}} \sum_1^N Y_i(\omega) < 0\}.$$

Let $f_n = \frac{1}{\sqrt{n}} \sum_1^n Y_i$, which is measurable by Corollary C. Then by 2) of Theorem D,

$$L = \{\omega : \ \lim f_n(\omega) \text{ converges}\} \in \mathcal{F},$$

and so

$$C = L \cap (\liminf f_n)^{-1}(-\infty, 0) \in \mathcal{F}.$$

◇◇

Example 4. Going back to Unit M again, $\Omega = [0,1)$, $\mathcal{F} = \mathcal{B}$ and each d_n is measurable. Corollary C says that for each n, $\frac{1}{n}\sum_1^n d_k(\omega)$ is measurable function, and so Theorem D says

$$H = \{\omega : \ \lim \frac{1}{2} \sum_1^n d_k(\omega) = 4.5\}$$

is indeed a $\mathcal{B}$ measurable set. ◇◇

Simple Functions. A random variable, or function, $f : \Omega \to \mathrm{I\!R}$ is called simple if it has finite range. This means we can write

$$f(\omega) \; = \sum_1^n x_i 1_{A_i}(\omega) \tag{1}$$

where $A_i \in \mathcal{F}$ are disjoint, $\Omega = \overset{n}{\underset{1}{\cup}} A_i$ and $x_i \in \mathrm{I\!R}$ are the distinct values in the range. ($A_i = \{\omega : \; f(\omega) = x_i\} = f^{-1}\{x_i\}$.) Even if the x_i are not distinct and $A_i \in \mathcal{F}$ are not disjoint, the above formula still produces a simple function. See problem 2.

Lemma E. *If f is real-valued and $\mathcal{F}$ measurable, there exists a sequence $\{f_n\}$ of $\mathcal{F}$ measurable simple functions with*

$$\begin{aligned} 0 \le f_n(\omega) \uparrow f(\omega) \quad & \textit{if } f(\omega) \ge 0 \\ 0 \ge f_n(\omega) \downarrow f(\omega) \quad & \textit{if } f(\omega) < 0 \end{aligned}$$

Proof: Just check that $f_n = \phi_n \circ f$ works, where $\phi_n : \mathrm{I\!R} \to \mathrm{I\!R}$ is the simple function

$$\phi_n(x) = \begin{cases} n & \text{if } x > n \\ k2^{-n} & \text{if } \; k2^{-n} < x \le (k+1)2^{-n} \text{ for some } 0 \le k \le n2^n - 1 \\ -k2^{-n} & \text{if } -k2^{-n} \ge x > -(k+1)2^{-n} \\ -n & \text{if } x \le -n. \end{cases}$$

∎

Generated Sigma-Fields and Functional Dependence

Suppose $T : \Omega \to \Gamma$ and $\mathcal{H}$ is a σ-field on Γ. Any σ-field $\mathcal{F}$ on Ω with respect to which T is measurable must contain all the sets $T^{-1}B, B \in \mathcal{H}$. If we have several $X_i : \Omega \to \Gamma$, $i = 1, 2, 3, \dots$ then for them all to be $\mathcal{F}/\mathcal{H}$ measurable means $\mathcal{F}$ contains $X_i^{-1}B$ for all i, all $B \in \mathcal{H}$. Let

$$\mathcal{A} = \{X_i^{-1}B : B \in \mathcal{H}, \text{ some } i\}.$$

Thus all the X_i are measurable if and only if $\mathcal{A} \subseteq \mathcal{F}$. Therefore

$$\sigma(\mathcal{A}) = \sigma(X_i : \; i = 1, 2, \dots), \text{ or simply } \sigma(X_i)$$

is the smallest σ-field on Ω with respect to which the X_i are all measurable. $\sigma(X_i)$ is called the σ-field generated by the X_i. Note that in the case of a single mapping, $\mathcal{A}$ is already a σ-field:

$$\sigma(T) = \{T^{-1}B : \; B \in \mathcal{H}\}.$$

Intuitively the sets $A \in \sigma(X_i)$ are ones that can be distinguished by the values of X_i alone. I.e. to tell if $\omega \in A$ it ought to be enough to know the values of the $X_i(\omega)$; knowing ω itself should not be necessary. This is essentially correct.

Example 5. Let $\Omega = \mathrm{I\!R}^2$ and $T(x, y) = x + y$. The sets $A \in \sigma(T)$ are all made up of unions of lines of the form $x + y$ =constant. (But not all such unions are in $\sigma(T)$.) ◇◇

Example 6. Let $\Omega = [0,1)$ and $d_n : \Omega \to \{0,1,\dots 9\}$ as in Unit M. Consider $\sigma(d_1)$, $\sigma(d_1, d_2)$ and $\sigma(d_1, d_2, d_3)$ $\dots$ The sequence of sigma-fields $\mathcal{F}_n = \sigma(d_i : i = 1,\dots,n)$ (sometimes called a "filtration") represents the increasingly refined knowledge about $\omega \in \Omega$ that is available as we collect the information revealed by $d_1(\omega), \dots, d_n(\omega)$ for increasing n. $\diamond\diamond$

Theorem F. *Suppose $X_1, \dots, X_n$ are random variables, $X_i : \Omega \to \mathbb{R}$.*

1) $A \in \sigma(X_1, X_2, \dots, X_n)$ if and only if A can be written as

$$A = \{\omega \in \Omega : \ (X_1(\omega), \dots, X_n(\omega)) \in H\}$$

for some $H \in \mathcal{B}(\mathbb{R}^n)$.

2) Y is a $\sigma(X_1, \dots, X_n)$ measurable random variable if and only if

$$Y(\omega) = f(X_1(\omega), \dots, X_n(\omega))$$

for some measurable function $f : \mathbb{R}^n \to \mathbb{R}$.

Proof: For 1) consider the following classes of subsets of Ω:

$$\mathcal{M} = \{X^{-1}H : \ H \in \mathcal{B}(\mathbb{R}^n)\}, \quad \mathcal{F} = \sigma(X_1, \dots, X_n).$$

Since (see page II.2) $X = (X_1, \dots, X_n)$ is $\mathcal{F}/\mathcal{B}(\mathbb{R}^n)$ measurable, $\mathcal{M} \subseteq \mathcal{F}$. On the other hand you can check that $\mathcal{M}$ is a σ-field. X is $\mathcal{M}/\mathcal{B}(\mathbb{R}^n)$ measurable. This implies that each X_i is $\mathcal{M}/\mathcal{B}(\mathbb{R}^n)$ measurable, which implies that $\mathcal{F} \subseteq \mathcal{M}$. Therefore $\mathcal{M} = \mathcal{F}$.

For 2) Theorem A tells us that if $Y = f \circ X$ then X is $\mathcal{F}/\mathcal{B}(\mathbb{R}^n)$ measurable, then Y is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ measurable. Conversely, suppose Y is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ measurable. First consider $Y = 1_A$ for $A \in \mathcal{F}$. By 1) this implies $A = X^{-1}H$ for some $H \in \mathcal{B}(\mathbb{R}^n)$. Therefore

$$Y(\omega) = 1_A(\omega) = 1_H(X_1(\omega), \dots, X_n(\omega)),$$

proving $Y = f \circ X$ for $f = 1_H$. If $Y = \sum_1^M y_m 1_{A_m}$ with $A_m = X^{-1}H_m$, $H_m \in \mathcal{B}(\mathbb{R}^n)$ then we write

$$\begin{aligned} Y(\omega) &= \sum_1^M y_m 1_{A_m}(\omega) \\ &= \sum_1^M y_m 1_{H_m}(X(\omega)) \\ &= f(X(\omega)) \quad \text{where } f = \sum_1^M y_m 1_{H_m} \end{aligned}$$

For the general case take a sequence of simple Y_j with $|Y_j| \le |Y|$ and $Y_j \to Y$. We know that $Y_j = f_j \circ X$ for Borel measurable f_j. Define

$$A_0 = \{x \in \mathbb{R}^n : \ \lim_{j\to\infty} f_j(x) \text{ converges}\} \text{ and } f(x) = \lim 1_{A_0}(x) f_j(x).$$

Then f is $\mathcal{B}(\mathbb{R}^n)/\mathcal{B}(\mathbb{R})$ measurable. Since we know $Y_j(\omega) = f_j(X(\omega)) \to Y(\omega)$, we conclude that $X(\omega) \in A_0$ for each ω and therefore $Y(\omega) = f(X(\omega))$. ∎

Induced Measures and Distributions

Suppose $(\Omega, \mathcal{F})$ and $(\Gamma, \mathcal{H})$ are measurable spaces, $T : \Omega \to \Gamma$ is $\mathcal{F}/\mathcal{H}$ measurable and μ is a measure on $(\Omega, \mathcal{F})$. For any $B \in \mathcal{H}$ we know $T^{-1}B \in \mathcal{F}$, so $\mu(T^{-1}B)$ is defined. Thus

$$\nu(B) = \mu(T^{-1}B)$$

assigns numerical ($[0, \infty]$) values to those $B \subseteq \Gamma$ which are in $\mathcal{H}$. We can easily check that in fact ν is a measure on $(\Gamma, \mathcal{H})$:

$$T^{-1}\emptyset = \emptyset \text{ implies } \nu(\emptyset) = \mu(\emptyset) = 0.$$

If $B_1, B_2, \dots \in \mathcal{H}$ are disjoint, then $T^{-1}B_i \in \mathcal{F}$ are disjoint. (If $\omega \in T^{-1}B_n \cap T^{-1}B_m$ then $T(\omega) \in B_n \cap B_m$, contradicting the disjointness.) Since $T^{-1}(\cup B_n) = \cup(T^{-1}B_n)$,

$$\begin{aligned}\nu(\cup B_n) &= \mu(T^{-1}(\cup B_n)) \\ &= \mu(\cup T^{-1}B_n) = \sum \mu(T^{-1}B_n) \\ &= \sum \nu(B_n).\end{aligned}$$

Sometimes we use the notation

$$\nu = \mu T^{-1}.$$

Think of ν as the measure on Γ resulting from applying the map T to the measure μ on Ω; the measure *induced* on Γ by μ and T. If μ is a probability measure, then so is ν:

$$\nu(\Gamma) = \mu(T^{-1}\Gamma) = \mu(\Omega) = 1.$$

In the case of a random variable $X : \Omega \to \mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, P)$, the induced measure $\nu_X = PX^{-1}$ on $(\mathbb{R}, \mathcal{B})$ is called the *distribution* (or sometimes the *law*) of X: for any $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}\nu_X(B) &= P(X^{-1}B) = P(\{\omega \in \Omega : \ X(\omega) \in B\}) \\ &= \text{“}P(X \in B)\text{” for short.}\end{aligned}$$

Statements like “X is a standard normal random variable”, or “X has Poisson distribution with parameter λ” are describing the distribution of X, i.e. probabilities $P(X \in B)$. Such statements do not tell us what $(\Omega, \mathcal{F}, P)$ is, or the specific definition of $X(\omega)$ for $\omega \in \Omega$.

The distribution function of X is

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

Thus F_X completely determines the distribution ν_X, according to Theorem II.E, but it doesn't say much at all about P.

Example 7. Are all standard normal random variables the same? ◇◇

When we have several random variables X_i; $i = 1, \dots, n$ we can talk about their individual or “marginal” distributions, ν_{X_i}, but more information is contained in their *joint distribution* which is the measure induced on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ by the random vector $X = (X_1, \dots, X_n)$.

Finally, independence of random variables is defined to mean independence of their generated σ-fields. For instance X_1, X_2, $\dots$ are independent if the $\sigma(X_i)$ are independent.

Problem **1**

Suppose $(\Omega, \mathcal{F})$ and $(\Gamma, \mathcal{H})$ are measurable spaces and $T : \Omega \to \Gamma$ is a mapping.

a) Show that both of the following define σ-fields:

$$\mathcal{G} = \{B \subseteq \Gamma : \ T^{-1}B \in \mathcal{F}\}$$
$$\mathcal{K} = \{A \subseteq \Omega : \ A = T^{-1}B \text{ for some } B \in \mathcal{H}\}.$$

b) Show that both $\mathcal{H} \subseteq \mathcal{G}$ and $\mathcal{K} \subseteq \mathcal{F}$ are equivalent ways of saying that T is $\mathcal{F}/\mathcal{H}$ measurable.

c) If $\mathcal{F} = \{\emptyset, \Omega\}$ and $\mathcal{H}$ contains all singleton sets $\{\mathit{\Omega}\}$, show that T is measurable if and only if T is a constant mapping.

Problem **2**

Consider a simple function f expressed by (1). If the x_i are distinct show that f is measurable if and only if the A_i are measurable. However if the x_i are not distinct, show that f can be measurable even if some of the A_i are not.

Problem **3**

Suppose $f, g : \Omega \to \mathbb{R}$ and $\mathcal{F}$ is a σ-field on Ω. Show that $f(\omega) + g(\omega) < c$ if and only if there exist rational numbers r, s such that $r + s < c$, $f(\omega) < r$ and $g(\omega) < s$. Use this to give a direct proof that $f + g$ is measurable if both f and g are.

Problem **4**

Suppose that X is a random variable whose distribution function F is continuous and strictly increasing. Show that $F(X)$ is a random variable with uniform distribution on $[0, 1]$, i.e. $P(F(X) \leq x) = x$ for $0 \leq x \leq 1$. Can you do this assuming that F is continuous but only non-decreasing? What if F can be any probability distribution function (not necessarily continuous)?

Problem **5**

Let $\Omega = (-\pi, \pi]$ with the Borel sets and define $P(\cdot) = \frac{1}{2\pi}\ell(\cdot)$. Let $X(\omega) = \sin(\omega)$ and $Y = \cos(\omega)$. Compute the distribution functions of X and Y. Are X and Y independent?

Problem **6**

Suppose $X_1, X_2, \ldots$ is a sequence of independent random variables defined on a probability space $(\Omega, \mathcal{F}, P)$.

a) Define the tail σ-field $\mathcal{T}$ associated with the X_i.

b) Prove Kolmogorov's 0-1 Law in this context.

c) Show that any random variable Z which is measurable with respect to $\mathcal{T}$ must be constant almost surely; i.e. there must exist a constant c and a set $N \in \mathcal{T}$ with $P(N) = 0$ so that $Z(\omega) = c$ for all $\omega \notin N$.

d) Show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i$$

either diverges with probability 1 or converges with probability 1, and in the latter case there is a constant c so that the limit $= c$ with probability 1.

Unit III .... **Integration and Expectation**

The careful development of measure theory now begins to produce its benefits. The measure-theoretic integral (alias expectation in probability theory) is one of the most powerful and important tools of modern analysis, including probabilistic analysis.

Suppose $X : \Omega \to \mathbb{R}$ is a random variable, defined on a probability space $(\Omega, \mathcal{F}, P)$. We want to be able to discuss its expected value "$E[X]$", second moment "$E[X^2]$", or more generally

$$E[\phi(X)]$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is any measurable function. There are some settings in which you may already have an idea of how to do this.

Example 1. Suppose the distribution of X is given in terms of a density function $p(\cdot)$:

$$P(X \le a) = \int_{-\infty}^{a} p(x)\, dx.$$

Then you will probably agree with

$$E[\phi(X)] = \int_{-\infty}^{\infty} \phi(x) p(x)\, dx.$$

◇◇

Example 2. If Ω is countable $\Omega = \{\omega_1, \omega_2, \dots\}$ with $P(\{\omega_i\}) = p_i$, then

$$E[\phi(X)] = \sum \phi(X(\omega_i)) \cdot p_i.$$

(Such p_i are sometimes called a "frequency function".) ◇◇

In general, for a measurable space $(\Omega, \mathcal{F}, \mu)$ and a measurable function $f : \Omega \to \mathbb{R}$ we are going to define the integral of f with respect to μ:

$$\int f(\omega)\, \mu(d\omega), \text{ or } \int f\, d\mu \text{ for short.}$$

Expectations are just the case of a probability measure:

$$E[X] = \int X\, dP, \quad E[\phi(X)] = \int \phi(X(\omega))\, P(d\omega).$$

An important special case is when $\mu = \ell$, Lebesgue measure on $(\mathbb{R}, \mathcal{B})$. In that setting the integral $\int f(x)\, \ell(dx)$ that we define here is called the Lebesgue integral, a powerful extension of the Riemann integral $\int_{-\infty}^{\infty} f(x)\, dx$ that you studied in calculus. The Lebesgue integral exists for a broader collection of functions $f(\cdot)$ than the Riemann, and has a more complete set of properties for manipulations. But when both exist, they agree:

$$\int_{[a,b]} f(x)\, \ell(dx) = \int_a^b f(x)\, dx.$$

(See Problem 6 also.) This allows us to use the various techniques of integration learned in calculus to evaluate the more sophisticated integral with respect to Lebesgue measure.

The definition of $\int f\, d\mu$ is not hard to understand. Consider a measure space $(\Omega, \mathcal{F}, \mu)$ and a nonnegative function $f \ge 0$. If we believe $\int f\, d\mu$ should give the "area under the graph" of $y = f(\omega)$, using μ to measure the "size" of subsets of Ω, then for $f = 1_A$ the value of the integral should certainly be

$$\int 1_A\, d\mu = \mu(A).$$

If the integral is also to obey the usual rules,

$$\int c \cdot f \, d\mu = c \int f \, du, \qquad \int f + g \, d\mu = \int f \, d\mu + \int g \, d\mu,$$

($c =$ a constant) then for a simple function

$$f(\omega) = \sum_1^n x_i 1_{A_i}(\omega), \tag{1}$$

with $x_i \geq 0$ and $A_i \in \mathcal{F}$, its integral (over Ω with respect to μ) must be given by:

$$\int f \, d\mu = \sum_1^n x_i \mu(A_i). \tag{2}$$

Notice that if $\mu(\Omega) = \infty$ then $\int 0 \, d\mu = 0$ implies the convention $0 \cdot \infty = 0$, mentioned in the Mathematical Supplements.

The Definition and Elementary Properties

The definition below refers to "partitions" of Ω. We will call $\{A_i\}_1^n$ a *partition* of Ω if each $A_i \in \mathcal{F}$, the A_i are disjoint and $\cup_1^n A_i = \Omega$.

Formula (2) for nonnegative simple functions (1) is natural enough. The extension to measurable $f \geq 0$ in general is also reasonable. First consider $f \geq 0$. The idea is that $\int f \, d\mu$ should be the supremum of the values of $\int \psi \, d\mu$ over all (measurable) simple functions ψ with $0 \leq \psi \leq f$. For a given partition $\Omega = \cup_1^n A_i$, the largest such simple function is $\psi = \sum x_i 1_{A_i}$using $x_i = \inf_{A_i} f$. For this ψ, formula (2) says

$$\int \psi \, d\mu = \sum_1^n [\inf_{A_i} f] \mu(A_i).$$

This explains the first part of the definition below. Note that by allowing $+\infty$ as a value, $\int f \, d\mu$ is always defined for $f \geq 0$.

For f in general, we split f into its positive and negative parts, $f^\pm : \Omega \to [0, +\infty]$ defined by

$$f^+(\omega) = \begin{cases} f(\omega) & \text{if } f(\omega) \geq 0 \\ 0 & \text{if } f(\omega) < 0 \end{cases} = f \vee 0,$$

$$f^-(\omega) = \begin{cases} 0 & \text{if } f(\omega) > 0 \\ -f(\omega) & \text{if } f(\omega) \leq 0 \end{cases} = -(f \wedge 0).$$

Note that both $f^\pm \geq 0$ (so that both $\int f^\pm \, d\mu$ are defined) and that $f(\omega) = f^+(\omega) - f^-(\omega)$. The definition is $\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu$. The convention that $\infty - \infty$ is undefined means that some integrals must remain undefined. For instance the integral of

$$f(x) = 1_{[0,\infty)} - 1_{(-\infty,0)}$$

with respect to Lebesgue measure is undefined, because $\int f \, d\ell = 1 \cdot \infty - 1 \cdot \infty = \infty - \infty$.

Definition of Integral. *Suppose $(\Omega, \mathcal{F}, \mu)$ is a measure space and $f : \Omega \to [-\infty, \infty]$ is measurable. If $f \geq 0$ for all ω then we define*

$$\int f(\omega) \, \mu(d\omega) = \sup \left\{ \sum [\inf_{A_i} f] \mu(A_i) : \ \{A_i\} \textit{ is a finite partition of } \Omega \textit{ into } \mathcal{F} \textit{ sets} \right\}.$$

In general

$$\int f(\omega)\,\mu(d\omega) = \int f^+\,d\mu - \int f^-\,d\mu,$$

unless both $\int f^\pm\,d\mu = +\infty$ in which case $\int f\,d\mu$ is considered undefined. We say f is integrable with respect to μ (or μ-integrable) if both $\int f^\pm\,d\mu < \infty$. If $A \in \mathcal{F}$, the integral of f over A is defined by

$$\int_A f\,d\mu = \int 1_A f\,d\mu.$$

Notice that if one of $\int f^\pm\,d\mu$ is finite but the other is $+\infty$, then $\int f\,d\mu$ isdefined ($= \pm\infty$) although f is not integrable. You may occasionally see the notation $\mu(f)$ instead of $\int f\,d\mu$.

The next two theorems collect the important elementary properties of the integral which are consequences of the definition above. Theorem A concerns nonnegative functions; Theorem B is about integrable functions.

We say some property holds *almost everywhere* (a.e.) if there is $B \in \mathcal{F}$ with $\mu(B) = 0$ so that the property holds for all $\omega \notin B$. (If $\mu = P$ is a probability measure, we also say *almost surely* (a.s.)) For instance to say $f \geq g$ a.e. means that there exists B with $\mu(B) = 0$ so that $f(\omega) \geq g(\omega)$ for all ω except the $\omega \in B$.

Theorem A. *Suppose $f, g : \Omega \to I\!R_\infty$ are nonnegative measurable functions.*

1) If $f = \sum_1^m y_j 1_{B_j}\,(y_j \geq 0,\ B_j \in \mathcal{F})$ then

$$\int f\,d\mu = \sum_1^m y_j \mu(B_j).$$

2) If $f = g$ a.e., then $\int f\,d\mu = \int g\,d\mu$

3) If $f \leq g$, a.e., then $\int f\,d\mu \leq \int g\,d\mu$

4) If $\alpha, \beta \geq 0$ then $\int(\alpha f + \beta g)\,d\mu = \alpha \int f\,d\mu + \beta \int g\,d\mu$.

5) If $\mu(\{\omega : f(\omega) > 0\}) > 0$, then $\int f\,d\mu > 0$

6) If $\int f\,d\mu < \infty$ then $f < \infty$ a.e.

Theorem B. *Suppose f, g are integrable.*

1) If $f \leq g$ a.e., then $\int f\,d\mu \leq \int g\,d\mu$

2) If $\alpha, \beta \in I\!R$, then $(\alpha f + \beta g)$ is also integrable, and $\int(\alpha f + \beta g)\,d\mu = \alpha \int f\,d\mu + \beta \int g\,d\mu$

3) $|\int f\,d\mu| \leq \int |f|\,d\mu$.

Comments.

- A1) says that the definition produces what we expected for simple functions.
- f is integrable if and only if both $\int f^\pm\,d\mu < \infty$, which is equivalent to

$$\int f^+\,d\mu + \int f^-\,d\mu = \int f^+ + f^-\,d\mu = \int |f|\,d\mu < \infty.$$

- If g is integrable and $|f| \leq |g|$ then f is integrable.
- The integral $\int f\,d\mu$ is blind to what f does on any particular set of measure 0. This is reΩected in all the "a.e."s.
- $f = g$ a.e. implies $f^\pm = g^\pm$ a.e. which implies $\int f\,d\mu = \int g\,d\mu$.
- If $\int$is replaced by $\int_A$, then all of the above remain true with "a.e." replaced by "a.e. on A", with the obvious meaning.

- $A, B \in \mathcal{F}$ disjoint and f integrable (or nonnegative) implies $\int_{A\cup B} f\,d\mu = \int_A f\,d\mu + \int_B f\,d\mu$. This is simply because $1_{A\cup B}f = 1_A f + 1_B f$.

Proof of Theorem A: First we record a simple fact. Suppose $\{A_i\}_1^n$ and $\{B_j\}_1^m$ are both partitions of Ω and

$$\phi = \sum_1^n x_i 1_{A_i}, \qquad \psi = \sum_1^m y_j 1_{B_j}$$

are simple functions with $\phi(\omega) \le \psi(\omega)$ for all ω. Then

$$\sum_1^n x_i \mu(A_i) \le \sum_1^m y_j \mu(B_j). \tag{3}$$

(Until A1 is proven we have no right to call these expressions $\int \phi$ or $\int \psi$.) Notice that if $\mu(A_i \cap B_j) > 0$ then there exists $\omega \in A_i \cap B_j$, and so $x_i = \phi(\omega) \le \psi(\omega) = y_j$. This shows that $x_i\mu(A_i \cap B_j) \le y_j\mu(A_i \cap B_j)$. Clearly the same inequality is also true if $\mu(A_i \cap B_j) = 0$. Thus (3) follows from writing

$$\sum_1^n x_i\mu(A_i) = \sum_{i=1}^n\sum_{j=1}^m x_i\mu(A_i \cap B_j) \le \sum_{i=1}^n\sum_{j=1}^m y_j\mu(A_i \cap B_j) = \sum_1^m y_j\mu(B_j).$$

We can now prove A1) under the additional assumption that the B_j, $j = 1, \ldots, m$ are disjoint. By including one additional B_{m+1} and $y_{m+1} = 0$ we get a partition $\{B_j\}_1^{m+1}$ and $\sum y_j\mu(B_j)$ does not change since $y_{m+1}\mu(B_{m+1}) = 0$. Now consider any partition $\{A_i\}_1^n$ and let $\phi = \sum[\inf_{A_i} f]1_{A_i}$. Then since $\phi \le f$, (3) tells us that $\sum[\inf_{A_i} f]\mu(A_i) \le \sum y_j\mu(B_j)$, which according to the definition of $\int f\,d\mu$ means that

$$\int f\,d\mu \le \sum y_j\mu(B_j).$$

For $\{A_i\} = \{B_j\}$ in particular, $\phi = f$ in which case (3) implies $\sum[\inf_{A_i} f]\mu(A_i) = \sum y_j\mu(B_j)$. This means that $\int f\,d\mu \ge \sum y_j\mu(B_j)$. We conclude then that

$$\int f\,d\mu = \sum y_j\mu(B_j), \quad \text{if } 0 \le f = \sum y_j 1_{B_j} \text{with } \ B_j \text{ disjoint.} \tag{4}$$

Suppose $f(\omega) \le g(\omega)$ for *all* ω. Consider any partition $\{A_i\}$. Then $\inf_{A_i} f \le \inf_{A_i} g$ and so

$$\sum[\inf_{A_i} f]\mu(A_i) \le \sum[\inf_{A_i} g]\mu(A_i) \le \int g\,d\mu.$$

We conclude that $\int f\,d\mu \le \int g\,d\mu$, giving us a preliminary version of A3). ∎

We next establish the following fact, which is the precursor of the convergence theorems D, E and F below.

Approximation Lemma. *Suppose $0 \le \phi_n$ are measurable simple functions such that $\phi_n \uparrow f$, for every ω. Then $\int \phi_n\,d\mu \uparrow \int f\,d\mu$.*

Since $\phi_n \le f$ we know from above that $\int \phi_n\,d\mu \le \int f\,\mu$ for all n, so that $\limsup \int \phi_n\,d\mu \le \int f\,d\mu$. So the lemma will follow if we can show $\liminf \int \phi_n\,d\mu \ge \int f\,d\mu$. For this it suffices to show

$$\liminf \int \phi_n\,d\mu \ge \int \psi\,d\mu \tag{5}$$

for any simple $0 \le \psi \le f$. Consider such a ψ. Then

$$\psi = \sum_j y_j 1_{B_j} \quad \phi_n = \sum_i x_i^n 1_{A_i^n},$$

where $\{B_j\}_1^m$ is a partition and, for each n, $\{A_i^n\}_1^{k_n}$ is a partition. Pick an arbitrary $\epsilon > 0$ and define

$$B_j^n = \{\omega \in B_j : \ \phi_n(\omega) \ge y_j(1-\epsilon)\} = \cup_{x_i^n \ge y_j(1-\epsilon)} A_i^n \cap B_j.$$

Then $B_j^n \uparrow B_j$ as $n \to \infty$, so $\mu(B_j^n) \uparrow \mu(B_j)$. Based on this,

$$\begin{aligned}\int \phi_n \, d\mu &= \sum_j \sum_i x_i^n \mu(A_i^n \cap B_j) \\ &= \sum_j \left[\sum_{x_i^n < y_j(1-\epsilon)} x_i^n \mu(A_i^n \cap B_j) + \sum_{x_i^n \ge y_j(1-\epsilon)} x_i^n \mu(A_i^n \cap B_j) \right] \\ &\ge \sum_j [0 + y_j(1-\epsilon)\mu(B_j^n)] \to (1-\epsilon) \sum_j y_j \mu(B_j) = (1-\epsilon) \int \psi \, d\mu.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, (5) follows, proving the lemma.

Proof (Theorem A continued): We can now prove A4). First suppose $f = \sum x_i 1_{A_i}$ and $g = \sum y_j 1_{B_j}$ are simple, the $\{A_i\}$ and $\{B_j\}$ being partitions. Let $C_{ij} = A_i \cap B_j$. Then $\{C_{ij}\}$ is a partition and $\alpha f + \beta g = \sum (\alpha x_i + \beta y_j) 1_{C_{ij}}$ is also a simple function. By (3) we can write

$$\begin{aligned}\int \alpha f + \beta g \, d\mu &= \sum_i \sum_j (\alpha x_i + \beta y_j) \mu(A_i \cap B_j) \\ &= \alpha \sum_i x_i \mu(A_i) + \beta \sum_j y_j \mu(B_j) = \alpha \int f \, d\mu + \beta \int g \, d\mu.\end{aligned}$$

In general there exist simple $f_n, g_n \ge 0$ with $f_n \uparrow f$ and $g_n \uparrow g$. Then each $\alpha f_n + \beta g_n$ is simple and $\uparrow \alpha f + \beta g$. The lemma above can now be used to see that

$$\int \alpha f + \beta g \, d\mu = \lim \int \alpha f_n + \beta g_n \, d\mu = \alpha \lim \int f_n \, d\mu + \beta \lim \int g_n \, d\mu = \alpha \int f \, d\mu + \beta \int g \, d\mu.$$

A1) now follows from A4), even if the B_j are not disjoint.

Suppose $N \in \mathcal{F}$ with $\mu(N) = 0$, $\{A_i\}$ is any partition and $f \ge 0$ is measurable. If $\inf_{A_i}[f 1_N] > 0$ then $A_i \subseteq N$ so that $\mu(A_i) = 0$. Hence $\sum \inf_{A_i}[f 1_N] \mu(A_i) = 0$ for all partitions, which means

$$\int f 1_N \, d\mu = 0.$$

If $\{f \ne g\} \subseteq N$ then since $f = f 1_N + f 1_{N^c}$ and $f 1_{N^c} = g 1_{N^c}$,

$$\begin{aligned}\int f \, d\mu &= \int f 1_N \, d\mu + \int f 1_{N^c} \, d\mu \\ &= \int f 1_{N^c} \, d\mu \\ &= \int g 1_{N^c} \, d\mu \\ &= \int g 1_N \, d\mu + \int g 1_{N^c} \, d\mu = \int g \, d\mu.\end{aligned}$$

This establishes A2).

If $\mu(N) = 0$ and $\{f \not\leq g\} \subseteq N$, then $f 1_{N^c} \leq g 1_{N^c}$ so that we can use our preliminary version of A3) to conclude

$$\int f\, d\mu = \int f 1_{N^c}\, d\mu \leq \int g 1_{N^c}\, d\mu = \int g\, d\mu,$$

proving A3) in general.

Consider A5). Let $A_n = \{\omega : \ f(\omega) \geq 1/n\}$. Then $A_n \uparrow \{f > 0\}$ and so $\mu(A_n) \uparrow \mu(\{f > 0\})$ which is > 0 by assumption. Hence $\mu(A_n) > 0$ for some n. But then $f \geq \frac{1}{n} 1_{A_n}$, from which we conclude

$$\int f\, d\mu \geq \int \frac{1}{n} 1_{A_n} d\mu = \frac{1}{n} \mu(A_n) > 0.$$

Finally, for A6), let $A = \{f = +\infty\}$. Since $\infty 1_A \leq f$,

$$\infty \mu(A) = \infty \int 1_A\, d\mu \leq \int f\, d\mu < \infty,$$

which implies that $\mu(A) = 0$. ▮

Example 3. $\Omega = \{1, 2, 3, \ldots\}$ and $\mu =$ counting measure. Then $f : \Omega \to \mathbb{R}$ is just a sequence, $f(n) = f_n$:

$$f(\omega) = \sum_1^\infty f_n 1_{\{n\}}(\omega).$$

Let $\psi_n = \sum_1^n \ f_k 1_{\{k\}}$ The ψ_n are simple and $\psi_n \uparrow f$. Therefore, by (5),

$$\begin{aligned} \int f\, d\mu &= \lim_{n\to\infty} \int \psi_n\, d\mu \\ &= \lim_{n\to\infty} \sum_1^n f_k = \sum_1^\infty f_n. \end{aligned}$$

Thus the theory of infinite series is subsumed by our general integration theory. Summation is just one example of integration. ◇◇

Riemann and Lebesgue. The integral on $\mathbb{R}$ with respect to Lebesgue measure, or Lebesgue integral $\int_{[a,b]} f\, d\ell$, is defined differently than the Riemann integral $\int_a^b f(x)\, dx$ of calculus. The Lebesgue integral $\int f\, d\ell$ exists more generally and has more powerful theoretical properties, making it *by far* more appropriate conceptually. On the other hand, we have a more extensive set of computational techniques for the Riemann integral. (There is a tradeoff between theoretical generality and computational utility.) As we will see, both integrals produce the same value when the Riemann integral is defined, such as when $f : [a, b] \to \mathbb{R}$ is continuous. This allows us to appeal to the integration techniques of calculus for the evaluation of many Lebesgue integrals.

Suppose $f : [a, b] \to \mathbb{R}$ is measurable. (We can extend its definition to the rest of $\mathbb{R}$ by $f = 0$ on $[a, b]^c$). We want to understand the connection between the Lebesgue and Riemann integrals,

$$\int_{[a,b]} f\, d\ell, \quad \text{and} \quad \int_a^b f(x)\, dx.$$

(We assume $-\infty < a < b < \infty$ here. See problem 5 for unbounded intervals.) Using $\ell([c,\, d]) = d - c$, the definition of the Riemann integral can be stated as follows.

Definition of Riemann Integral. *To say f is Riemann integrable, with $\int_a^b f(x)\,d(x) = R$ means $|R| < \infty$ and given any $\epsilon > 0$ there exists $\delta > 0$ so that*

$$|R - \sum_1^n f(x_i)\ell(J_i)| < \epsilon$$

whenever $\{J_i\}_1^n$ is a partition of $[a,b]$ into intervals with $\ell(J_i) < \delta$ all i, and any choice of evaluation points $x_i \in J_i$.

Suppose f <u>is</u> Riemann integrable. Given $\epsilon > 0$ let $\delta > 0$ be as promised by the definition. Take any partition $\{J_i\}_1^n$ as specified. It follows that

$$|R - \sum_1^n [\inf_{J_i} f]\ell(J_i)| \le \epsilon \quad \text{and} \quad |R - \sum_1^n [\sup_{J_i} f]\ell(J_i)| \le \epsilon$$

Define the simple functions $g_* = \sum_1^n [\inf_{J_i} f]1_{J_i}$ and $g^* = \sum_1^n [\sup_{J_i} f]1_{J_i}$. Then $g_* \le f \le g^*$ on $[a,b]$, so

$$R - \epsilon \le \int_{[a,b]} g_*\,d\ell \le \int_{[a,b]} f\,d\ell \le \int_{[a,b]} g^*\,d\ell \le R + \epsilon$$

I.e. $|R - \int_{[a,b]} f\,d\ell| \le \epsilon$ for every $\epsilon > 0$. Therefore $\int_{[a,b]} f\,d\ell = R = \int_a^b f(x)dx$. This proves the following theorem.

Theorem C. *If the measurable function f is Riemann integrable on the bounded interval $[a,b]$ then f is ℓ-integrable on $[a,b]$ and*

$$\int_{[a,b]} f\,d\ell = \int_a^b f(x)\,dx.$$

Examples 4. If $Q \subseteq \mathbb{R}$ is the set of rational numbers, then for any Borel set A, $\int_A 1_Q\,d\ell = 0$ because $\ell(A \cap Q) \le \ell(Q) = 0$. However the Riemann integral $\int_a^b 1_Q(x)\,dx$ is undefined.

Consider the Lebesgue integral $\int_{[0,1]} \frac{1}{\sqrt{x}}\,d\ell$.

$$\begin{aligned} \int_{[0,1]} \frac{1}{\sqrt{x}}\,d\ell &= \lim_{n\to\infty} \int_{[\frac{1}{n},1]} x^{-1/2}\,d\ell \qquad \text{— see the convergence theorems below} \\ &= \lim \int_{\frac{1}{n}}^1 x^{-1/2}\,dx = \lim(2 - 2/\sqrt{n}) = 2. \end{aligned}$$

Thus $\int_{[0,1]} \frac{1}{\sqrt{x}}\,d\ell$ agrees with the value of $\int_0^1 \frac{1}{\sqrt{x}}\,dx$ as an *improper* Riemann integral. $\int_0^1 \frac{1}{\sqrt{x}}\,dx$ is not defined in the strict sense of the definition of Riemann integral. ◇◇

There are however some distinctions between the Riemann and Lebesgue integrals.

- On unbounded intervals, such as $[0,\infty)$ the improper Riemann integral

$$\int_0^\infty f(x)\,dx = \lim_{T\to\infty} \int_0^T f(x)\,dx$$

and Lebesgue integral

$$\int_{[0,\infty)} f\,d\ell = \int_{[0,\infty)} f^+\,d\ell - \int_{[0,\infty)} f^-\,d\ell$$

are defined differently. Either can exist without the other. For instance

$$\int_0^\infty \frac{\sin(x)}{x}\,dx = \pi/2,$$

but $\int_{[0,\infty)} \frac{\sin(x)}{x}\, d\ell$ is undefined. But when they both exist, they must agree.

- The Riemann integral incorporates a notion of orientation, reﬂected in the formula

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

I.e in addition to the set $[a,b]$ over which we integrate, we specify the direction of integration (from a to b, or from b to a). The Lebesgue integral has no such concept of orientation.

Instead of ℓ on $(\mathbb{R},\mathcal{B})$, we can consider a measure μ described in terms of a distribution function F: $\mu((a,b]) = F(b)-F(a)$. It is possible to define the Riemann-Stiltjes integral

$$\int_a^b f(x)\,dF(x) = R$$

by replacing ℓ with μ in the definition of Riemann integral above. This notion of integral is related to $\int_{[a,b]} f\,d\mu$ in the same way as described in Theorem C. Some authors write "$\int_{[a,b]} f\,dF$"to mean the measure-theoretic integral $\int_{[a,b]} f\,d\mu$. In general there is no standard notation to distinguish between Riemann and measure-theoretic integrals. In anything you read you will have to figure out what that author's individual conventions are. We will indicate Riemann integrals using limits of integration, $\int_a^b \cdot\, dF$, and measure-theoretic integrals with subscripted domains of integration, $\int_{[a,b]} \cdot\, d\mu$.

Expected Values. If X is a random variable defined on $(\Omega,\mathcal{F},P)$ then its expected value is just another name for its integral with respect to P:

$$E[X] = \int_\Omega X(\omega)\,dP(\omega),$$

provided this is defined. If $X = c1_\Omega$, a constant random variable, then since P is a probability measure

$$E[c] = \int c\,dP = cP(\Omega) = c\cdot 1 = c.$$

Theorem B 2) says $E[cX] = cE[X]$ in general.

$E[X^k]$, if it exists, is called the k-th *moment*. $E[|X|^k]$ always exists (possibly $+\infty$) and is called the k-th *absolute* moment. The first moment $m = E[X]$ is usually called the *mean*. If the mean is finite then we can also define the *variance*,

$$\mathrm{Var}[X] = E[(X-m)^2] = E[X^2 - 2mX + m^2] = E[X^2] - m^2.$$

We also write

$$E[X;A] = \int_A X\,dP = \int 1_A X\,dP = E[X\cdot 1_A].$$

Convergence Theorems

One of the features of the measure-theoretic integral which makes it more useful than the Riemann integral is the possibility of passing limits underneath integration:

$$\lim \int f_n\,d\mu \overset{?}{=} \int \lim f_n\,d\mu$$

I.e. if $f_n \to f$ (a.e.) then under what circumstances can we conclude that $\int f_n\,d\mu \to \int f\,d\mu$?

Example 5. Consider f_n defined on $\mathbb{R}$ by

$$f_n(x) = \begin{cases} n - n^2 x & \text{if } 0 < x \leq 1/n \\ 0 & \text{otherwise} \end{cases}$$

and

$$g_n(x) = \begin{cases} 1 & \text{for } n < x \leq 3n \\ -1 & \text{for } -2n < x < -n \\ 0 & \text{otherwise.} \end{cases}$$

Then $f_n(x) \to 0$ and $g_n(x) \to 0$ for all x, but $\int g_n \, d\ell = 1$ and $\int f_n \, d\ell = 1/2$ for all n. ◇◇

This shows that something beyond $f_n \to f$ is needed to imply $\int f_n \, d\mu \to \int f \, d\mu$. There are three famous results in this department. In all these, we assume $(\Omega, \mathcal{F}, \mu)$ is a measure space and the functions f_n, f, g are $\mathbb{R}_\infty$-valued and measurable.

The Monotone Convergence Theorem (D). *If* $0 \leq f_n \uparrow f$ *a.e., then* $\int f_n \, d\mu \uparrow \int f \, d\mu$

Fatou's Lemma (E). *If* $0 \leq f_n$*, then*

$$\int [\liminf_{n\to\infty} f_n] \, d\mu \leq \liminf_{n\to\infty} \int f_n \, d\mu.$$

The Dominated Convergence Theorem (F). *If* $|f_n| \leq g$ *a.e.,* g *is integrable and* $f_n \to f$ *a.e., then* f *is integrable and*

$$\int f_n \, d\mu \to \int f \, d\mu.$$

Proof (D): Notice that this is a generalization of our Approximation Lemma – we can use essentially the same proof. Consider any partition $\{A_i\}_1^\infty$ of Ω and define $v_i = \inf_{A_i} f$. Consider any $\epsilon > 0$. Let

$$A_i^m = \{\omega : \ f_m(\omega) > v_i(1 - \epsilon)\}.$$

Then $A_i^m \uparrow A_i$ as $m \to \infty$, so $\mu(A_i^m) \uparrow \mu(A_i)$. We can now justify the following sequence of assertions.

$$\begin{aligned} f_m &\geq (1-\epsilon) \sum_{i=1}^n v_i 1_{A_i^m} \\ \int f_m \, d\mu &\geq (1-\epsilon) \sum_{i=1}^n v_i \mu(A_i^m) \to (1-\epsilon) \sum_{i=1}^n v_i \mu(A_i) \\ \liminf \int f_n \, d\mu &\geq (1-\epsilon) \sum_{i=1}^n v_i \mu(A_i), \quad \text{for all } \epsilon > 0 \\ \liminf \int f_n \, d\mu &\geq \sum_1^n [\inf_{A_i}] \mu(A_i) \\ \liminf \int f_n \, d\mu &\geq \int f \, d\mu \end{aligned}$$

But $f_n \leq f$ implies $\liminf \int f_n \, d\mu \leq \int f \, d\mu$. We conclude that

$$\int f_n \, d\mu \uparrow \int f \, d\mu.$$

∎

Proof (E): Let $g = \liminf f_n$ and $g_n = \inf_{k\geq n} f_k$. Then $0 \leq g_n \uparrow g$. Therefore

$$\lim \int g_n \, d\mu = \int g \, d\mu.$$

Since $g_n \leq f_n$ we see that

$$\int \liminf f_n \, d\mu = \int g \, d\mu = \lim \int g_n \, d\mu \leq \liminf \int f_n \, d\mu.$$

∎

Proof (F): First, by modifying all the functions on a measurable set with $\mu(N) = 0$ we can assume that the convergence is for all ω. Next, $f_n \to f$ implies $f_n^\pm \to f^\pm$ and since $|f_n| \leq g$ we also have $0 \leq f_n^\pm \leq g$. It suffices therefore to assume $0 \leq f_n \leq g$.

Fatou's Lemma tells us that $\int f \, d\mu \leq \liminf \int f_n \, d\mu$. If we consider $h = g - f$ and $h_n = g - f_n$, then $h, h_n \geq 0$ and $h_n \to h$. Fatou's Lemma now tells us that

$$\begin{aligned} \int h \, d\mu &\leq \liminf \int h_n \, d\mu \\ \int g \, d\mu - \int f \, d\mu &\leq \liminf[\int g \, d\mu - \int f_n \, d\mu] \\ &= \int g \, d\mu - \limsup \int f_n \, d\mu \\ \limsup \int f_n \, d\mu &\leq \int f \, d\mu. \end{aligned}$$

We can now conclude that $\int f_n \, d\mu \to \int f, d\mu$. ∎

Example 4 (continued). Our assertion above that

$$\int_{[0,1]} \frac{1}{\sqrt{x}} \, d\ell = \lim_{n\to\infty} \int_{[\frac{1}{n},1]} x^{-1/2} \, d\ell$$

is the monotone convergence theorem, since $x^{-1/2} 1_{[\frac{1}{n},1]} \uparrow x^{-1/2} 1_{[0,1]}$ almost surely (the exception being $x = 0$). ◇◇

Example 6. Suppose μ is a probability measure on $\mathbb{R}$, perhaps the distribution of some random variable. The associated *moment generating function* is

$$M(s) = \int_{\mathbb{R}} e^{sx} \, \mu(dx),$$

defined for those s for which it is finite. This is closely related to the Laplace transform of μ, usually taken to be

$$\int e^{-sx} \, d\mu(x) = M(-s).$$

Notice that

- $M(0) = 1$ is always defined;
- If M is defined for s_0 and s_1, and $s_0 \leq s \leq s_1$, then $M(s)$ is also defined because $0 < e^{sx} \leq e^{s_0 x} + e^{s_1 x}$. Thus the domain of $M(s)$ will always be some kind of interval containing 0.
- If μ is "supported" on $[0, \infty)$ (i.e. $\mu(-\infty, 0)) = 0$) then $M(s)$ will be defined for all $s \leq 0$, at least.

Discussion of further properties of $M(s)$ provides a good illustration of the use of the convergence theorems.

Suppose $M(s)$ is defined on some interval $[s_0 - \epsilon, s_0 + \epsilon]$ around s_0. For any positive integer k there exists a constant c so that for all x

$$|x|^k \le c(e^{\epsilon x} + e^{-\epsilon x})$$
$$|x|^k e^{s_o x} \le c(e^{(s_0+\epsilon)x} + e^{(s_0-\epsilon)x}).$$

Thus $\int x^k e^{s_0 x}\, d\mu < \infty$. We want to write $e^{sx} = \sum_0^\infty \frac{(s-s_0)^k}{k!} x^k e^{s_0 x}$, and take the integral by integrating each term of the series individually. The Dominated Convergence Theorem justifies this, since

$$\begin{aligned}\left|\sum_0^N \frac{(s-s_0)^k}{k!} x^k\right| e^{s_0 x} &\le \sum_0^\infty \left|\frac{(s-s_0)^k}{k!} x^k\right| e^{s_0 x} \\ &= e^{|(s-s_0)x|} e^{s_0 x} \\ &\le e^{(s_0+\epsilon)x} + e^{(s_0-\epsilon)x},\end{aligned}$$

because $|(s-s_0)x| \le \epsilon|x| \le \pm\epsilon x$, depending on the sign of x. The right side above is μ-integrable, by assumption, providing the dominating function for the Dominated Convergence Theorem. Therefore

$$\begin{aligned}M(s) = \int e^{sx}\, d\mu &= \int \sum_0^\infty \frac{(s-s_0)^k}{k!} x^h e^{s_0 x}\, d\mu \\ &= \sum_0^\infty \int \frac{(s-s_0)^k}{k!} x^h e^{s_0 x}\, d\mu \\ &= \sum_0^\infty \frac{(s-s_0)^k}{k!} \int x^k e^{s_0 x}\, d\mu.\end{aligned}$$

In particular $\int x^k e^{s_0 x}\, d\mu = (\frac{d}{ds})^k M(s_0)$. If M is defined on $[-\epsilon, \epsilon]$ for some $\epsilon > 0$, then for $|s| < \epsilon$

$$M(s) = \sum_0^\infty s^k \frac{m_k}{k!},$$

where $m_k = \int x^k\, d\mu$ are the moments, hence the name "moment generating function". Its derivatives are the moments:

$$m_k = M^{(k)}(0).$$

See Example 12 below for a specific calculation. ◇◇

Densities and Changes of Variable

There are a couple situations in which the measure we need to integrate with respect to is related to another measure that we understand better. We would like to translate the original integral into one with respect to the better understood measure.

$$\begin{tikzcd}\Omega,\, P \arrow[r, "X"] & \mathbb{R},\, \mu_X \arrow[r, "\phi"] & \mathbb{R} \\ & \mathbb{R},\, \ell \arrow[u, "d\mu_X = p\, d\ell"'] & \end{tikzcd}$$

This issue comes up twice in the situation of Example 1. The expected value of $\phi(X)$ is defined to be an integral with respect to P on the Ω of the underlying probability space:

$$E[\phi(X)] = \int_\Omega \phi(X(\omega))\, P(d\omega).$$

The distribution μ of X is a different measure on a different space, $\mathbb{R}$ We expect to be able to calculate using

$$\int_\Omega \phi(X(\omega))\, P(d\omega) = \int_{\mathbb{R}} \phi(x)\, \mu(dx). \tag{6}$$

This is essentially a change of variables from $\omega \in \Omega$ to $x \in \mathbb{R}$. If μ has a "density" $p(x)$ we expect this in turn to be calculated as a Lebesgue integral,

$$\int_{\mathbb{R}} \phi(x)\, \mu(dx) = \int_{\mathbb{R}} \phi(x)p(x)\, \ell(dx), \tag{7}$$

which we hope to finally evaluate by connecting it to the Riemann integral

$$\int_{-\infty}^{\infty} \phi(x)p(x)dx.$$

Thus beyond the connection between the Riemann and Lebesgue integrals, we want to

- validate the change of variables $x = X(\omega)$ in (6), and
- understand what is meant by a "density" and why (7) is valid.

Densities. The distribution μ of a random variable X is just a probability measure on $(\mathbb{R}, \mathcal{B})$ constructed from X:

$$\mu(A) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

Many of the important distributions arising in practice can be described in terms of Lebesgue measure using a *density function* $p(x) \geq 0$: for all $-\infty < a < b < \infty$

$$\mu((a, b]) = \int_a^b p(x)\, dx = \int_{(a,b]} p\, d\ell. \tag{8}$$

Examples 7.

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} && \text{– standard normal} \\ p(x) &= 1_{[0,\infty)} \lambda e^{-\lambda x} && \text{– exponential } (\lambda > 0) \\ p(x) &= \frac{1}{\pi} \frac{u}{u^2 + x^2} && \text{– Cauchy } (u > 0). \end{aligned}$$

◇◇

Most, if not all, probability measures μ on $(\mathbb{R}, \mathcal{B})$ that you know either have such a density or are of the form

$$\mu(A) = \sum_{n \in A} p_n, \text{ for some } \sum_{-\infty}^{\infty} p_n = 1. \tag{9}$$

However there exist many probability measure that are of neither of these two types. For instance there exist distribution functions $F(x)$ which are continuous but not given by an integral integral of any density with respect to Lebesgue measure. (An example is the Cantor ternary function on $[0, 1]$.) Thus (8) and (9) by no means account for all μ on $\mathbb{R}$!

Problem 2 will show that (8) for intervals $(a, b]$ implies that the same formula holds for all Borel sets, $A \in \mathcal{B}$:

$$\mu(A) = \int_A p\, d\ell.$$

In general when μ and ν are two measures on a measurable space $(\Omega, \mathcal{F})$, we say ν has *density ρ with respect to μ* if $\rho : \Omega \to \mathbb{R}$ is a nonnegative measurable function such that

$$\nu(A) = \int_A \rho\, d\mu \quad \text{for all } A \in \mathcal{F}.$$

The following theorem tells us how ν-integrals are related to μ- integrals in such cases, justifying (7) above.

Theorem G. *Suppose ν has density ρ with respect to μ and $f : \Omega \to \mathbb{R}_\infty$ is measurable. Then*

1) $\int f\, d\nu = \int f \rho\, d\mu$, if $f \geq 0$;

2) f is ν-integrable if and only if $f \cdot \rho$ is μ-integrable, in which case

$$\int_A f\, d\nu = \int_A f \cdot \rho\, d\mu$$

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

Example 8. We calculate the mean of the exponential distribution, with parameter $\lambda > 0$, as follows.

$$\text{mean} = \int x\, d\mu = \int x^+\, d\mu - \int x^-\, d\mu,$$

where

$$x^+(x) = \max\{x, 0\} \quad \text{and} \quad x^-(x) = -\min\{x, 0\}.$$

Now

$$\int x^-\, d\mu = \int x^-(x) p(x)\, d\ell = 0,$$

because μ has density $p(x) = 1_{[0,\infty)} \lambda e^{-\lambda x}$ with respect to ℓ, and $x^- p = 0$ for all x.

$$\begin{aligned} \int x^+\, d\mu &= \int x^+(x) p(x)\, d\ell \\ &= \int_{[0,\infty)} x\lambda e^{-\lambda x}\, d\ell \\ &= \lim_{n\to\infty} \int_{[0,n]} x\lambda e^{-\lambda x}\, d\ell, \text{ by M.C.T.} \end{aligned}$$

But

$$\begin{aligned} \int_{[0,n]} x\lambda e^{-\lambda x}\, d\ell &= \int_0^n x\lambda e^{-\lambda x}\, dx \\ &= -e^{-\lambda x} \frac{(\lambda x + 1)}{\lambda} \Big|_{x=0}^{x=n} \\ &= \frac{1}{\lambda}[1 - e^{-n\lambda}(n\lambda + 1)] \to 1/\lambda \end{aligned}$$

as $n \to \infty$. We conclude that $\int x\, d\mu = 1/\lambda$. ◇◇

Change of Variable. The distribution μ of a random variable X is the measure induced by X on $\mathbb{R}$ using the probability measure P from the underlying space $(\Omega, \mathcal{F})$: $\mu = PX^{-1}$.

$$\begin{tikzcd} \mu,\, \Omega \arrow[r, "T", "\mu T^{-1} = \nu"'] & \nu,\, \Gamma \arrow[r, "f"] & \mathbb{R} \end{tikzcd}$$

Here is the general change of variable theorem.

Theorem H. *Suppose $(\Omega, \mathcal{F})$ and $(\Gamma, \mathcal{H})$ are measurable spaces, μ is a measure on $(\Omega, \mathcal{F})$, $T : \Omega \to \Gamma$ is measurable, and $\nu = \mu T^{-1}$ is the induced measure. Suppose $f : \Gamma \to \mathbb{R}_\infty$ is $\mathcal{F}/\mathcal{B}(\mathbb{R}_\infty)$ measurable. If $f \geq 0$ then*

$$\int_\Omega f(T(\omega))\, \mu(d\omega) = \int_\Gamma f(\mathit{\Omega})\, \nu(d\mathit{\Omega}).$$

f is ν-integrable if and only if $f \circ T$ is μ-integrable, in which case the preceding equation again holds.

Basically, this works in general because it works for indicator functions: let $f = 1_B$ where $B \in \mathcal{H}$. Then

$$f \circ T(\omega) = 1_B(T(\omega)) = 1_{T^{-1}B}(\omega),$$

so

$$\int_\Gamma f\, d\nu = \nu(B) = \mu(T^{-1}B) = \int_\Omega f \circ T\, d\mu.$$

Examples 9. $E[e^{sX}] = \int e^{sx}\, \mu(dx)$, $E[X^2] = \int x^2\, \mu(dx)$. ◇◇

Example 10. Suppose $T : \mathbb{R}^d \to \mathbb{R}^d$ is a one-to-one map having continuous derivatives and nonvanishing Jacobian, $J(x) = \det[\partial T_i/\partial x_j]$. In an advanced calculus course you might have learned that to make the change of variables $y = T(x)$ you should use $dy = |J(x)|\,dx$ in integrals. This is an instance of Theorem G, which we would write as

$$\int_{T^{-1}B} f(T(x))|J(x)|\,\ell(dx) = \int_B f(y)\,\ell(dy),$$

provided B is contained in the range of T. In the language of Theorem G this is saying that if μ is the measure on the domain having density $|J(x)|$ with respect to $\ell(dx)$, then the induced measure $\nu = \mu T^{-1}$ agrees with Lebesgue measure $\ell(dy)$ on the range. (To be precise if $R = \{T(x) : \ x \in \mathbb{R}^d\}$ is the subset of the range actually covered by T, ν would be Lebesgue measure restricted to R: $\nu(A) = \ell(A \cap R)$.) We might illustrate this with a diagram, such as

$$\begin{tikzcd} x \in T^{-1}B \subseteq \mathbb{R}^d \arrow[r, "T(x)=y", "|J(x)|\ell(dx)T^{-1}=1_R(y)\ell(dy)"'] & y \in B \subseteq \mathbb{R}^d \arrow[r, "f"] & \mathbb{R}. \end{tikzcd}$$

◇◇

Important Inequalities

Suppose X is a random variable. Consider any integer $k = 1, 2, \ldots$ and positive constant α. Since

$$\alpha^k 1_{\{|X| \geq \alpha\}} \leq |X|^k 1_{\{|X| \geq \alpha\}} \leq |X|^k$$

it follows that

$$\alpha^k P[|X| \geq \alpha] \leq E[X^k; \ |X| \geq \alpha] \leq E[|X|^k].$$

This gives

MARKOV'S INEQUALITY (I).

$$P[|X| \geq \alpha] \leq \frac{1}{\alpha^k} E[|X|^k; \ |X| \geq \alpha] \leq E[|X|^k]/\alpha^k.$$

Applying this to $X - m$, where the mean $m = E(X)$ is assumed finite, and using $k = 2$, we get

CHEBYSHEV'S INEQUALITY (J). *If $E[|X|] < \infty$ then for any $\alpha > 0$*

$$P[|X - m| \geq \alpha] \leq \frac{1}{\alpha^2} \mathit{Var}[X].$$

Another important inequality is Jensen's inequality for convex functions. Suppose $J \subseteq \mathbb{R}$ is an interval and

$$\phi : J \to \mathbb{R}$$

satisfies

$$\phi(px + qy) \leq p\phi(x) + q\phi(y)$$

for all $x, y \in J$ and $p, q \geq 0$ with $p + q = 1$. Then ϕ is called *convex*.

Examples 11. Any smooth function with $\phi'' \geq 0$ is convex, such as

$$\phi(x) = x^2, \quad \phi(x) = e^x, \quad \phi(x) = -\log(x).$$

◇◇

Jensen's Inequality (K). *If $\phi : J \to I\!R$ is convex, $X(\omega) \in J$ a.s., and $E[|X|] < \infty$, then*

$$\phi(E[X]) \leq E[\phi(X)].$$

In terms of the distribution μ of X Jensen's inequality could be written

$$\phi\left(\int x\,\mu(dx)\right) \leq \int \phi(x)\,\mu(dx).$$

This is <u>only</u> true when μ is a probability measure; it does not hold for all measures.

Hölder's Inequality (L). *: Suppose $(\Omega, \mathcal{F}, \mu)$ is any measure space, $f, g : \Omega \to I\!R_\infty$ are measurable and $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\int |fg|\,d\mu \leq \left[\int |f|^p\,d\mu\right]^{1/p} \left[\int |g|^q\,d\mu\right]^{1/q}.$$

Moment Generating and Characteristic Functions

If X is a random variable, and μ its distribution, the associated *moment generating function* is

$$M(s) = E[e^{sX}] = \int_{I\!R} e^{sx}\,\mu(dx),$$

defined for those s for which it is finite. We discussed some of its properties in Example 6 above. In particular if M is defined on $[-\epsilon, \epsilon]$ for some $\epsilon > 0$, then all the moments $m_k = E[X^k] = \int x^k\,d\mu$ are finite and for $|s| \leq \epsilon$

$$M(s) = \sum_0^\infty s^k \frac{m_k}{k!},$$

and $E[X^k] = M^{(k)}(0)$.

Example 12. Suppose X has normal distribution, mean 0, variance σ^2:

$$\begin{aligned} M(s) &= \int_{-\infty}^\infty e^{sx}(2\pi\sigma^2)^{-1/2} e^{-x^2/2\sigma^2}\,dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty e^{\frac{1}{2\sigma^2}[(x-\sigma^2 s)^2 - \sigma^4 s^2]}\,dx \\ &= e^{\sigma^2 s^2/2} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\sigma^2 s)^2/2\sigma^2}\,dx = e^{\sigma^2 s^2/2} \\ &= \sum_0^\infty \frac{\sigma^{2k}}{2^k k!} s^{2k}. \end{aligned}$$

Therefore

$$E[X^n] = \begin{cases} 0 & \text{for odd } n \\ \frac{(2k)!}{2^k k!}\sigma^{2k} & \text{for } n = 2k. \end{cases}$$

◇◇

The *characteristic function* (also called the Fourier transform) of μ,

$$\widehat{\mu}(t) = E[e^{itX}] = \int e^{itx}\,\mu(dx),$$

is a complex-valued function of $t \in I\!R$. Unlike the moment generating function, $\widehat{\mu}$is defined for <u>all</u> t. The Dominated Convergence Theorem tells us that it is continuous. We will discuss its significance more in the next unit.

Problem **1** ....

If $(\Omega, \mathcal{F}, \mu)$ is a measure space and f is a nonnegative real-valued function. Show that

$$\nu(A) = \int_A f \, d\mu, \quad A \in \mathcal{F},$$

defines a measure on $(\Omega, \mathcal{F})$.

Problem **2** ....

Suppose $X : \Omega \to \mathbb{R}$ is measurable and $p \geq 0$ is measurable with

$$P(a < X \leq b) = \int_{(a,b]} p(x) \, \ell(dx) \quad \text{all } -\infty < a \leq b < +\infty.$$

Show that

$$P(X \in A) = \int_A p(x) \, \ell(dx) \quad \text{for all } A \in \mathcal{B}.$$

Problem **3** ....

Prove Beppo Levi's Theorem: If f_n are integrable, $\sup \int f_n \, d\mu < \infty$ and $f_n \uparrow f$, then f is integrable and $\int f_n \, d\mu \to \int f \, d\mu$. [Hint: write $f_n = g_n + f_1$ and work with the g_n.]

Problem **4** ....

Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is one-to-one, onto and $\phi' \geq 0$ is continuous. State and prove a theorem, analogous to Theorems F and G above, concerning the validity of

$$\int f(y) \, \ell(dy) = \int f(\phi(x)) \phi'(x) \, \ell(dx).$$

(You may appeal to Theorems G and H in the course of your proof.)

Problem **5** ....

Let $T : \mathbb{R} \to \mathbb{R}$ be given by $T(x) = x^4$ and $\mu = \ell T^{-1}$ be the measure induced by Lebesgue measure. Identify μ by finding its density with respect to Lebesgue measure.

Problem **6** ....

Suppose that f is Lebesgue integrable on $[0, +\infty)$ and Riemann integrable on every bounded interval $\subseteq [0, \infty)$. Show that the improper Riemann integral

$$\int_0^\infty f(x) \; dx = \lim_{N \to +\infty} \int_0^N f(x) \; dx$$

converges to $\int_{[0,\infty)} f \; d\ell$. Prove the same thing if instead of assuming f is integrable we assume $f \geq 0$.

Problem **7** ....

If X is a positive random variable whose distribution has density p. Show that $1/X$ has distribution with density $p(1/x)/x^2$. [You may need to refine the statement of this to make it really true!]

Problem **8** ....

Show that the Cauchy distribution has no mean, not even an infinite one. (I.e. show that both $\int x^{\pm}(x) \, d\mu = +\infty$ where $x^{+}(x) = \max(x, 0)$ and $x^{-}(x) = -\min(x, 0)$.)

Problem **9** ....

The gamma density, with parameters $\alpha, u > 0$, is

$$p(x) = 1_{(0,\infty)} \frac{\alpha^u}{\Gamma(u)} x^{u-1} e^{-\alpha x}.$$

Show that the associated moment generating function is $(1 - s/\alpha)^{-u}$ for $s < \alpha$ and that the k-th moment is

$$u(u+1) \cdots (u+k-1)/\alpha^k.$$

Problem **10** ....

Show that $m = E[X]$ is the value which minimizes $E[(X - m)^2]$.

Problem **11** . . . .

a) Suppose $0 < \alpha < \beta$. Show that Lyapunov's inequality,

$$E[|X|^{\alpha}]^{1/\alpha} \leq E[|X|^{\beta}]^{1/\beta},$$

can be deduced both from Hölder's inequality (take $g \equiv 1$) and from Jensen's inequality.

b) Show that for any positive random variable X and $p > 0$,

$$E[1/X^p] \geq 1/E[X]^p.$$

Problem **12** . . . .

Suppose that N is a random variable with standard normal distribution. Find a density for $\log(|N|)$.

Problem **13** . . . .

Use the calculations of Example 12 to compute the characteristic function of a normal random variable X with mean 0 and variance σ^2. In other words, write

$$e^{itX} = \sum_{0}^{\infty} \frac{(it)^n}{n!} X^n$$

and take the expected value of both sides. (Use the Dominated Convergence Theorem to justify writing "$E \sum_0^\infty = \sum_0^\infty E$")

Unit IV .... **Convergence Concepts**

Suppose $(\Omega, \mathcal{F}, P)$ is a probability space on which are defined random variables X, X_n; $n \geq 1$. There are several different concepts of convergence "$X_n \to X$" in common usage. We will only summarize them and some of their properties.

Description	Notation	Meaning
Almost sure convergence	$X_n \to X$ a.s.	$X_n(\omega) \to X(\omega)$ all $\omega \notin N$, some $N \in \mathcal{F}$ with $P(N) = 0$
Convergence in probability (or "in measure")	$X_n \to_P X$	$P(\|X_n - X\| \geq \epsilon) \to 0$, all $\epsilon > 0$
Convergence in the mean (or L^1 convergence)	$X_n \to_{L^1} X$	$E[\|X_n - X\|] \to 0$
L^p convergence ($p \geq 1$)	$X_n \to_{L^p} X$	$E[\|X_n - X\|^p]^{1/p} \to 0$
Convergence in distribution	$X_n \Rightarrow X$	$E[\phi(X_n)] \to E[\phi(X)]$, all bounded continuous $\phi : \mathbb{R} \to \mathbb{R}$

These are all different. The only general implications are as follows.

1. If $X_n \to_{L^p} X$, then $X_n \to_{L^q} X$ for any $1 \leq q \leq p$.
2. If $X_n \to_{L^1} X$, then $X_n \to_P X$
3. If $X_n \to X$ a.s, then $X_n \to_P X$
4. If $X_n \to_P X$, then $X_n \Rightarrow X$.

The first of these follows from Lyapunov's inequality (problem III.11):

$$E[|X_n - X|^q]^{1/q} \leq E[|X_n - X|^p]^{1/p}.$$

The second is because of the inequality

$$P(|X_n - X| > \epsilon) \leq \frac{1}{\epsilon} E[|X_n - X|].$$

For 3, let $C = \{\omega : \; X_n(\omega) \to X(\omega)\}$. The assumption is that $P(C) = 1$. Consider any $\epsilon > 0$ and let $A_n = \{\omega : |X_k(\omega) - X(\omega)| < \epsilon \text{ all } k \geq n\}$. Then

$$A_n \uparrow A = \{\omega : \; |X_n(\omega) - X(\omega)| < \epsilon \text{ for all but finitely many } n\}.$$

Since $C \subseteq A$, we know that $P(A) = 1$. Hence $P(A_n) \uparrow 1$ and $P(A_n^c) \downarrow 0$. $\{|X_n - X| \geq \epsilon\} \subset A_n^c$ and thus $P[|X_n - X| \geq \epsilon] \to 0$. The last implication takes little more. (But notice that a.s. convergence implies $\phi(X_n) \to \phi(X)$ a.s., so that $X_n \Rightarrow X$ follows from the D.C.T.)

Counterexamples can be constructed to show that all other implications are false in general.

Examples 1. $\Omega = [0, 1)$, $P = \ell$. Let

$$X_n(x) = \begin{cases} 1 & \text{if } \frac{k}{2^m} \leq x < \frac{k+1}{2^m},\; 0 \leq k < 2^m \\ 0 & \text{otherwise} \end{cases}, \text{ where } n = 2^m + k < 2^{m+1}.$$

Then $X_n \to_P 0$, as well as in L^p for any $p > 1$, but $X_n(\omega)$ diverges for every ω! If $Y_n = 2^m X_n$, then $Y_n \to_P 0$, but not in any L^p.

Let $Z_n = \sin(2\pi n x)$, then $Z_n \Rightarrow Z_1$, because

$$\int_0^1 \phi(\sin(2\pi n x))\, dx = n \int_0^{1/n} \phi(\sin(2\pi n x))\, dx = \int_0^1 \phi(\sin(2\pi y))\, dy.$$

However $Z_n \not\to_P Z$. ◇◇

It is important to note that for convergence in distribution X_n and X are never compared directly! In fact $X_n \Rightarrow X$ is really a property of the respective distributions μ_n, μ.

$$E[\phi(X_n)] = \int_{\mathbb{R}} \phi \, d\mu_n \to \int_{\mathbb{R}} \phi \, d\mu = E[\phi(X)],$$

all bounded continuous ϕ. This is denoted $\mu_n \Rightarrow \mu$ and called *weak convergence* of the distributions. If F_n, F are the corresponding distribution functions we also write $F_n \Rightarrow F$. It turns out that this is also equivalent to

$$F_n(x) \to F(x) \quad \text{for all } x \text{ at which } F \text{ is continuous.}$$

(But this does not mean $\mu_n(A) \to \mu(A)$ all $A \in \mathcal{B}$!) Thus

$$\begin{aligned} X_n &\Rightarrow X \quad \text{(convergence in distribution)} \\ \mu_n &\Rightarrow \mu \quad \text{(weak convergence)} \\ F_n &\Rightarrow F \quad \text{(weak convergence)} \end{aligned}$$

all refer to the same thing.

The standard theory contains a number of additional results aimed at providing a more detailed understanding of what is required for $\mu_n \Rightarrow \mu$. In particular given a sequence $\{\mu_n\}$ of probability distributions on $(\mathbb{R}, \mathcal{B})$ it is important to know when there exists a subsequence which converges weakly to some other probability distribution: $\mu_{n_k} \Rightarrow \nu$. This is the issue of "tightness"; see Billingsley §25.

We will not be giving full proofs of the Laws of Large Numbers, or the Central Limit Theorem, but it is worthwhile to understand what type of convergence these results refer to. Suppose X_n is a sequence of independent, identically distributed random variables. (This means the σ-fields $\sigma(X_n)$ are independent and $\mu(\cdot) = PX_n^{-1}$ is the same for all n.) Suppose the mean $m = E[X_n]$ is finite. Then the standard limit laws are as follows.

Weak Law of Large Numbers (A). $\frac{1}{n}\sum_{i=1}^n X_i \to_P m.$

Strong Law of Large Numbers (B). $\frac{1}{n}\sum_{i=1}^n X_i \to m$ *a.s.*

Central Limit Theorem (C). *Assuming* $\sigma^2 = Var[X_n] < \infty$,

$$\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n (X_i - m) \Rightarrow N,$$

where N is a random variable with the standard normal distribution.

Problem 7 points out that the convergence in the Central Limit Theorem cannot be strengthened to a.s. convergence.

These are the standard limit laws that most students of probability theory have seen before. There are others that may not be as familiar.

Law of the Iterated Logarithm (D). *Assume* $m = 0$ *and* $E[X_n^2] = 1$. *Then*

$$P[\limsup_{n\to\infty} (2n \log(\log n))^{-1/2} \sum_1^n X_i = 1] = 1.$$

In other words, $\frac{1}{n}\sum_1^n X_i$ has the "asymptotic envelopes" $\pm\sqrt{2\log(\log n)/n}$ as $n \to \infty$. (See Breiman, Theorem 13.25 in particular, for more on the Law of the Iterated Logarithm.)

A more interesting result is Chernoff's Theorem. Of the limit laws stated here, it is the only one in which the distribution of the X_i (rather than just its mean) is involved in the final assertion of the theorem. We will need some definitions first. If $M(s) = E[e^{sX_i}]$ is the moment generating function of the X_i, the function

$$H(c) = \sup_{s \in \mathbb{R}} \{cs - \log(M(s))\}$$

is called the *Cramer transform* of the distribution of X_i.

Examples 2. Cramer transforms of some common distributions:
Binomial ($P[X = 0] = p,\ P[X = 1] = 1 - p$):

$$H(c) = \begin{cases} c\log(\frac{c}{1-p}) + (1-c)\log(\frac{1-c}{p}) & \text{if } c \in [0,1] \\ +\infty & \text{otherwise.} \end{cases}$$

Standard Normal:

$$H(c) = \frac{1}{2}c^2.$$

Exponential ($1_{[0,\infty)}e^{-x}$):

$$H(c) = \begin{cases} c - 1 - \log(c) & \text{if } c > 0 \\ +\infty & \text{if } c \le 0. \end{cases}$$

◇◇

Chernoff's Theorem (E). *For $c \ge m$,*

$$\lim \frac{1}{n} \log P[\frac{1}{n}\sum_1^n X_i \ge c] = -H(c).$$

For $c \le m$,

$$\lim \frac{1}{n} \log P[\frac{1}{n}\sum_1^n X_i \le c] = -H(c).$$

(You can find a treatment of Chernoff's Theorem in the little book by Bahadur. Our statement above appears more general than what you will find in Bahadur, but can be derived from it by considering $X_i - c$ in place of X_i.)

A Restricted Proof of the Strong Law. There is a short proof of the Strong Law of Large Numbers under the additional hypothesis that the X_i have finite fourth moments. Let

$$\xi^4 = E[X_i^4] < \infty$$
$$\sigma^2 = E[x_i^2] < \infty.$$

We can assume that $m = E[X_i] = 0$. Let

$$S_n = \sum_1^n X_i.$$

The key is to get an upper bound on $E[S_n^4]$.

$$E[S_n^4] = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n E[X_i X_j X_k X_l].$$

When the right side of this is multiplied out, we get 3 kinds of terms:

a) Terms with at least one index distinct from the others. (E.g. j different from i, k, or l.) By independence all such terms have expected value 0. (E.g. $E[X_j]E[X_iX_kX_l] = 0$).

b) Terms with indices in two distinct pairs (e.g. $i = k \neq j = l$.) For these we have $E[\cdot] = (\sigma^2)^2$. The number of such terms is $n \cdot 3 \cdot (n-1)$.

c) Terms with $i = j = k = l$. For these $E[\cdot] = \xi^4$, and there are n such terms.

We find that

$$E[S_n^4] = n\xi^4 + 3n(n-1)\sigma^4 \le Kn^2, \quad \text{where } K = 3\sigma^4 + \xi^4.$$

Now apply Markov's inequality:

$$P[|S_n| \ge n\epsilon] \le \frac{1}{n^4\epsilon^4}Kn^2 = n^{-2}\epsilon^{-4}K.$$

Since

$$\sum_1^\infty P[|S_n| \ge n\epsilon] \le \sum_1^\infty \frac{1}{n^2}\epsilon^{-4} < \infty,$$

we can apply the first Borel-Cantelli Lemma to obtain $P[L_\epsilon] = 0$ where

$$\begin{aligned} L_\epsilon &= \limsup\{\omega : \ |S_n| \ge n\epsilon\} \\ &= \{\omega : |\frac{1}{n}S_n(\omega)| \ge \epsilon \text{ for infinitely many } n\}. \end{aligned}$$

Let $N = \cup_{k=1}^\infty L_{1/k}$ We know $P[N] \le \sum_1^\infty P[L_{1/k}] = 0$. For any $\omega \neq N$ and any k it follows that $|\frac{1}{n}S_n(\omega)| \ge \frac{1}{k}$ for only a finite number of n, so that $|\frac{1}{n}S_n(\omega)| < \frac{1}{k}$ for all sufficiently large n. Hence $\frac{1}{n}S_n(\omega) \to 0$ for all $\omega \in N$. (The full proof is more complicated, since it must not assume the existence of higher moments.)

Characteristic Functions

There are several ways to identify a probability measure μ on $(\mathbb{R}, \mathcal{B})$ in terms of more conventional mathematical objects, specifically functions:

- a density with respect to Lebesgue measure,

$$\mu((a,b]) = \int_{(a,b]} p(x)\, d\ell,$$

  if one exists;
- the distribution function

$$F(x) = \mu((-\infty, x]);$$

- the moment generating function $M(s) = \int e^{sx}\mu(dx)$, provided it is defined on some open interval (method of moments);
- the characteristic function

$$\widehat{\mu}(t) = \int e^{itx}\mu(dx) = \int \cos(tx)\mu(dx) + i\int \sin\,(tx)\mu(dx).$$

Examples (2). Characteristic function of some common distributions:

Distribution	Description	$\hat{\mu}(t)$
Standard Normal	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	$e^{-t^2/2}$,
Uniform	$1_{[0,1]}$	$\frac{e^{it}-1}{it}$
Exponential	$e^{-x}1_{(0,\infty)}$	$\frac{1}{1-it}$
Cauchy	$\frac{1}{\pi}\frac{1}{1+x^2}$	$e^{-\|t\|}$
Poisson	$P(\{k\}) = \frac{1}{k!}e^{-1}$, $k = 0, 1, 2, \ldots$	$e^{e^{it}-1}$

◇◇

Moment generating functions may seem more appealing than characteristic functions since they do not involve complex numbers, but they have some drawbacks. One always has to worry about their domain, i.e. for what s is $M(s)$ defined? Because of that, knowing $M(s)$ does not determine μ in all cases. Specifically, there are many μ for which $M(0) = 1$ and $M(s)$ is undefined for all $s \neq 0$. (The Cauchy distributions, for any value of the parameter u, are examples.) Suppose however that $M_\mu(s)$ and $M_\nu(s)$ are the moment generating functions of two probability measures μ and ν. If there is some (nonempty) open interval (a, b) on which both M_μ and M_ν are finite and equal, then it is true that $\mu = \nu$.

Characteristic functions have a "cleaner" theory. $\widehat{\mu}(t)$ is always defined and continuous for all $t \in \mathbb{R}$, and we always have $|\widehat{\mu}(t)| \leq 1$. (This is the complex modulus.) The following facts explain some of the theoretical importance of characteristic functions.

- If μ and ν are two probability measures on $(\mathbb{R}, \mathcal{B})$ with characteristic functions $\widehat{\mu}$ and $\widehat{\nu}$, then $\mu = \nu$ if and only if $\widehat{\mu}(t) = \widehat{\nu}(t)$ for all t.
- If $\mu_n, n = 1, 2, \cdots$, and μ are probability measures on $(\mathbb{R}, \mathcal{B})$ then $\mu_n \Rightarrow \mu$ if and only if
$$\widehat{\mu_n}(t) \to \widehat{\mu}(t) \text{ for every } t.$$
- If X and Y are independent random variables with distributions μ_X and μ_Y, then the characteristic function of $X + Y$ is
$$\widehat{\mu}_{X+Y}(t) = \widehat{\mu}_X(t) \cdot \widehat{\mu}_Y(t).$$
(See Unit 5.)

Some further issues addressed in the standard theory of characteristic functions are

∗ inversion formulas, i.e. how $\mu((a, b])$ can be calculated from $\widehat{\mu}(\cdot)$;
∗ how properties of μ manifest themselves in $\widehat{\mu}$;
∗ given a function $\psi(t)$ how can we tell when it is the characteristic function of some μ: $\psi(t) = \widehat{\mu}(t)$?
∗ given a sequence μ_n, how can we tell from looking at $\widehat{\mu_n}$ if $\mu_n \Rightarrow \mu$ for some μ? (It is possible for $\widehat{\mu_n} \to \psi(t)$, but $\psi(t)$ fail to be $\psi = \widehat{\mu}$.)

Proving the Central Limit Theorem. We won't provide all the details, but the basic approach to proving the Central Limit Theorem can be quickly described. It illustrates the use of characteristic functions for convergence in distribution. We can assume the X_j have mean $m = E[X_j] = 0$. (Otherwise consider $X_j - m$.) Suppose that $\phi(t) = E[e^{itX_j}]$. Then $\frac{1}{\sigma\sqrt{n}} \sum_1^n X_j$ has characteristic function $\phi(\frac{t}{\sigma\sqrt{n}})^n$. Since $E[e^{itN}] = e^{-t^2/2}$, the goal is to show that for all $t \in \mathbb{R}$

$$\phi(\frac{t}{\sigma\sqrt{n}})^n \to e^{-t^2/2}, \quad \text{as } n \to \infty.$$

Formally,

$$\begin{aligned} \phi(0) &= 1 \\ \phi'(0) &= E[iX_j] = im = 0 \\ \phi''(0) &= E[-X_j^2] = -\sigma^2. \end{aligned}$$

So we expect (second order Taylor polynomial) $\phi(t) \approx 1 - \frac{\sigma^2}{2}t^2$ for $t \approx 0$. Using this it seems reasonable that

$$\begin{aligned} \phi(\frac{t}{\sigma\sqrt{n}})^n &\approx (1 - \frac{\sigma^2}{2}\frac{t^2}{\sigma^2 n})^n \\ &= (1 - \frac{1}{2}t^2 \cdot \frac{1}{n})^n \to e^{-\frac{1}{2}t^2}. \end{aligned}$$

These approximations can be justified to provide a rigorous proof.

Infinitely Divisible Distributions and Stable Laws

The search for all probability distributions that could play the role of the standard normal in CLT-like theorems was a major research topic in the early part of this century (1920 – 1940). (Problem 6 indicates that the Cauchy distribution is an example.) This led to the identification of two general classes of distributions with special properties. It was the use of characteristic functions that made a complete answer to this problem possible. (The text by Breiman is a good reference on these topics. See Chapter 9 in particular.)

Infinitely Divisible Distributions. A probability measure μ on $\mathbb{R},\mathcal{B}(\mathbb{R})$ is called *infinitely divisible* if for any n there exist i.i.d. $X_1, \ldots, X_n$ so that μ is the distribution of $\sum_1^n X_i$. (The distribution of the X_i may depend on n.) It turns out that μ is infinitely divisible if and only if $\hat{\mu}(t)$ has the following form:

$$\hat{\mu}(t) = \exp\left[i\beta t - \frac{1}{2}\sigma^2 t^2 + \int (e^{itx} - 1 - \frac{itx}{1+x^2})\frac{1+x^2}{x^2}\,\nu(dx)\right] ,$$

for some finite measure ν on $\mathbb{R}$ with $\nu(\{0\}) = 0$. (The role of $\nu(\{0\})$ is played by σ^2.) The normal, Poisson and Cauchy are all examples. (For the Cauchy, $\beta = \sigma^2 = 0$, $\nu(dx) = \frac{1}{\pi(1+x^2)}\,\ell(dx)$.)

Stable Distributions. The probability measure μ on $\mathbb{R}$ is called *stable* if when X_i are i.i.d. with distribution μ, then for every n there exist constants a_n, b_n so that $(\sum_1^n X_i - b_n)/a_n$ also has distribution μ. These are precisely the distributions which can occur in CLT-like results for X_i i.i.d.:

$$\frac{1}{a_n}(\sum_1^n X_i - b_n) \Rightarrow \mu.$$

A distribution μ is stable if and only if it is infinitely divisible and either normal or with $\hat{\mu}(t)$ as above using $\sigma^2 = 0$ and ν described by

$$\frac{1+x^2}{x^2}\nu(dx) = (m_- 1_{(-\infty)} + m_+ 1_{(0,+\infty)})\frac{1}{|x|^{1+\alpha}}\,\ell(dx)$$

for some "exponent" $0 < \alpha < 2$ and $m_\pm \geq 0$. For $\alpha < 0$ no moments of order $n < \alpha$ exist. Thus the normal is the <u>only</u> stable distribution with finite variance. (This is what makes it so important!) The Cauchy distribution has $\alpha = 1$. More explicit expressions for the characteristic functions of the non-normal stable distributions can be given. For $\alpha \neq 1$: for some $\beta, \mathit{\Omega}, \theta \in \mathbb{R}$; $\mathit{\Omega} > 0$ and $|\theta| < 1$,

$$\hat{\mu}(t) = \exp\left[i\beta t - \mathit{\Omega}|t|^\alpha(1 + i\theta\frac{t}{|t|}\tan(\frac{\pi}{2}\alpha))\right].$$

For $\alpha = 1$, $\beta, \mathit{\Omega}, \theta$ can be as above, but the formula changes to

$$\hat{\mu}(t) = \exp\left[i\beta t - \mathit{\Omega}|t|^\alpha(1 + i\theta\frac{t\,2}{|t|\pi}\,\log|t|)\right].$$

Problem **1** ....

a) Show that $\mu_n \Rightarrow \mu$ is possible for μ_n having densities but μ not.
b) Show that $\mu_n \Rightarrow \mu$ is possible for μ having a density but the μ_n not.
c) Show that it is possible for $\mu_n \Rightarrow \mu$ if all the μ_n and μ have densities but the densities do not converge.

Problem **2** . . . .

Suppose the distributions of random variables X_n and X have densities p_n and p. Show that if $p_n(x) \to p(x)$ for all x outside a set of Lebesgue measure 0, then $X_n \Rightarrow X$. Hint: Work in terms of the associated distribution functions $F_n(x)$ and $F(x)$. Use Fatou's lemma to show that $F(x) \leq \liminf F_n(x)$. Similarly, since $1 - F_n(x) = \int_{(x,\infty)} p_n \, d\lambda$, you can show that $1 - F(x) \leq \liminf[1 - F_n(x)]$. Now it is a basic property of lim inf and lim sup that

$$\liminf[1 - F_n(x)] = 1 - \limsup F_n(x).$$

(You should take this for granted.) Conclude that

$$F(x) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x).$$

This is equivalent to $F_n(x) \to F(x)$.

Problem **3** . . . .

If $X_n \to 0$ a.s. then show $\frac{1}{n}\sum_{k=1}^{n} X_k \to 0$ a.s. also. Give an example to show that this can fail if a.s. convergence is replaced by convergence in probability.

Problem **4** . . . .

Suppose X is a random variable with characteristic function

$$\phi(t) = E[e^{itX}].$$

Find a formula for the characteristic function of $\alpha(X - \beta)$ in terms of ϕ. Use this to generalize the table of characteristic functions above to the usual parameterized versions of the distributions cited.

Problem **5** . . . .

Suppose X_n is a sequence of independent, identically distributed random variables with the exponential distribution, parameter λ (i.e. density $1_{[0,\infty)}\lambda e^{-\lambda x}$). The characteristic function of this distribution is

$$\phi(t) = \frac{\lambda}{\lambda - it}.$$

a) Calculate the characteristic function of

$$\frac{1}{\sigma\sqrt{n}}\sum_{1}^{n}(X_k - m),$$

where m is the mean and σ^2 is the variance.

b) Using the formula $\log(1 + z) = z - \frac{1}{2}z^2 + \mathcal{O}(z^3)$ as $z \to 0$, show that

$$\frac{1}{\sigma\sqrt{n}}\sum_{1}^{n}(X_k - m) \Rightarrow N,$$

where, as in the Central Limit Theorem, N is a standard normal random variable. By $\mathcal{O}(z^3)$ is meant a term with $|\mathcal{O}(z^3)| \leq B|z^3|$ for some constant B and all z sufficiently close to 0. (Do this by showing the characteristic functions converge.)

Problem **6** . . . .

Suppose X_n is a sequence of independent, identically distributed random variables with the Cauchy distribution, parameter u (i.e. density $\frac{1}{\pi}\frac{u}{u^2+x^2}$). The characteristic function of this distribution is

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

(You can take this fact for granted.) By computing the characteristic function, find an exponent $\mathit{\Omega} > 0$ so that

$$\frac{1}{n^{\mathit{\Omega}}}\sum_{1}^{n} X_k$$

converges in distribution. Note that the value of $\mathit{\Omega}$ is not $\frac{1}{2}$! Compare what you find to the Strong Law and Central Limit Theorems. Isn't there a contradiction?

Problem **7** . . . .

Show that the convergence in the Central Limit Theorem is <u>not</u> almost sure convergence. Use the Law of the Iterated Logarithm to explain the remark in parentheses at the end of problem I.16.

Unit V .... Advanced Constructions

Suppose we have two random variables X, Y defined on $(\Omega, \mathcal{F}, P)$. Let μ_X and μ_Y be their individual (or marginal) distributions,

$$\mu_X(A) = P(X \in A) \quad \mu_Y(B) = P(Y \in B), \quad A, B \in \mathcal{B}(\mathbb{R}),$$

and π their joint distribution:

$$\pi(C) = P((X,Y) \in C), \quad C \in \mathcal{B}(\mathbb{R}^2).$$

μ_X and μ_Y are easily obtained from π:

$$\mu_X(A) = \pi(A \times \mathbb{R}), \quad \mu_Y(B) = \pi(\mathbb{R} \times B).$$

We can not determine the joint distribution π from the marginals μ_X and μ_Y in general. However if X and Y are independent (meaning that $\sigma(X)$ and $\sigma(Y)$ are independent) then, for any $A, B \in \mathcal{B}(\mathbb{R})$, $X^{-1}A \in \sigma(X)$ and $Y^{-1}B \in \sigma(Y)$ are independent as sets. Therefore

$$\begin{aligned} P(X^{-1}A \cap Y^{-1}B) &= P(X^{-1}A) \cdot P(Y^{-1}B) \\ P((X,Y) \in A \times B) &= P(X \in A) \cdot P(Y \in B) \\ \pi(A \times B) &= \mu_X(A) \cdot \mu_Y(B). \end{aligned}$$

The collection of all such $A \times B$ forms a π-system which generates $\mathcal{B}(\mathbb{R}^2)$. This means the joint distribution is determined by the marginals if X and Y are independent. Based on this we might expect that calculations with respect to π might be carried out in terms of μ_X and μ_Y:

$$\int_{\mathbb{R}^2} \phi(x,y)\, d\pi = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \phi(x,y)\, d\mu_Y \right] d\mu_X \quad \text{(or the other order).}$$

Here π is what we will call the product measure "$\pi = \mu_X \times \mu_Y$" and the reduction of the "double integral" $\int d\pi$ to an "iterated integral" is called Fubini's Theorem. The discussion of this is our next topic.

When X and Y are not independent we cannot reconstruct their joint distribution from just the marginals. The connection involves the idea of conditional probabilities, which will be our final topic.

Product Measure Spaces

Suppose $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ are measure spaces. The product set $X \times Y$ is just the set of all pairs (x, y) with $x \in X$ and $y \in Y$. We want to discuss the natural *product σ-field* $\mathcal{X} \times \mathcal{Y}$ on $X \times Y$ and the *product measure* $\pi = \mu \times \nu$ on $(X \times Y, \mathcal{X} \times \mathcal{Y})$.

The Product Sigma-Field. Let $\mathcal{M}$ consist of those subsets of $X \times Y$ which have the form $A \times B$, where $A \in \mathcal{X}, B \in \mathcal{Y}$. These are called the *measurable rectangles*. $\mathcal{M}$ is not a σ-field (it is not closed under complementation) but it isa π-system. The product σ-field is defined to be

$$\mathcal{X} \times \mathcal{Y} = \sigma(\mathcal{M}).$$

Note that $C \in \mathcal{X} \times \mathcal{Y}$ does not mean $C = A \times B$; see Example 1.

Lemma A. *Suppose $\mathcal{A}$ and $\mathcal{B}$ are classes of subsets of X and Y respectively, with the property that $X = \cup A_i$ and $Y = \cup B_j$ for some $A_i \in \mathcal{A}$ and $B_j \in \mathcal{B}$. If $\mathcal{X} = \sigma(\mathcal{A})$ and $\mathcal{Y} = \sigma(\mathcal{B})$ then $\mathcal{X} \times \mathcal{Y} = \sigma(\mathcal{C})$ where $\mathcal{C}$ consists of all $A \times B$ with $A \in \mathcal{A}$, $B \in \mathcal{B}$.*

Proof: If $\mathcal{M}$ is the class of measurable rectangles above, then $\mathcal{C} \subseteq \mathcal{M}$ implies $\sigma(\mathcal{C}) \subseteq \mathcal{M}$. Given $B \in \mathcal{B}$, the class of $A \subseteq X$ such that $A \times B \in \sigma(\mathcal{C})$ is a σ-field containing $\mathcal{A}$ and therefore contains $\mathcal{X}$. (For this we need to know $X \times B = \cup A_i \times B \in \sigma(\mathcal{C})$.) Now consider any $A \in \mathcal{X}$. The class of all $B \subseteq Y$ for which $A \times B \in \sigma(\mathcal{C})$ is a σ-field which contains $\mathcal{B}$ and so must also contain $\mathcal{X}$. This shows that $\mathcal{M} \subseteq \sigma(\mathcal{C})$, and so $\mathcal{X} \times \mathcal{Y} \subseteq \sigma(\mathcal{C})$. ▮

Example 1. With $X = Y = \mathbb{R}$ and $\mathcal{X} = \mathcal{Y} = \mathcal{B}(\mathbb{R})$, the lemma (with $\mathcal{A} = \mathcal{B} = \mathcal{J}_0$ consisting just of intervals with finite endpoints $(a, b]$) implies that

$$\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R}).$$

(Our original definition of $\mathcal{B}(\mathbb{R}^2)$ was given on page I.11. The $\mathcal{R}$ used there is the same as $\mathcal{C}$ is the lemma above.) In particular the diagonal $D = \{(x, y) : x = y\}$ is in the product σ-field, since it is a closed set. However D is clearly not in $\mathcal{M}$. ◇◇

Theorem B. *If $C \in \mathcal{X} \times \mathcal{Y}$ then its cross-sections are measurable:*

$$C_x = \{y \in Y : \ (x, y) \in C\} \in \mathcal{Y} \text{ for every } x \in X;$$
$$C_y = \{x \in X : \ (x, y) \in C\} \in \mathcal{X} \text{ for every } y \in Y.$$

If $(\Omega, \mathcal{F})$ is a measurable space and $f : X \times Y \to \Omega$ is $\mathcal{X} \times \mathcal{Y}/\mathcal{F}$ measurable, then for each x and y

$$f_x : Y \to \Omega \text{ given by } \ f_x(y) = f(x, y) \text{ is}\mathcal{Y}/\mathcal{F} \text{ measurable;}$$
$$f_y : X \to \Omega \text{ given by } \ f_y(x) = f(x, y) \text{ is}\mathcal{X}/\mathcal{F} \text{ measurable.}$$

The converse is false! $C_x \in \mathcal{Y}$ and $C_y \in \mathcal{X}$ for all x, y does not imply $C \in \mathcal{X} \times \mathcal{Y}$, nor do the measurability of f_x and f_y imply the joint measurability of f.

Proof: The first part of the theorem follows by applying the second to $f = 1_C$; e.g. $f_x(y) = 1_{C_x}(y)$. The second part is an application of Theorem II.A. Consider any $x \in X$. Define $T : Y \to X \times Y$ by $T(y) = (x, y)$. For any $A \times B \in \mathcal{M}$ (i.e. $A \in \mathcal{X}$ and $B \in \mathcal{Y}$) we have

$$T^{-1}A \times B = \begin{cases} B & \text{if } x \in A \\ \emptyset & \text{if } x \notin A \end{cases} \in \mathcal{Y}.$$

Since $\mathcal{M}$ generates $\mathcal{X} \times \mathcal{Y}$, this proves that T is $\mathcal{Y}/\mathcal{X} \times \mathcal{Y}$ measurable. Now notice that $f_x = f \circ T$, and is therefore $\mathcal{Y}/\mathcal{F}$ measurable. ▮

Example 2. Suppose $N \subset \mathbb{R}$ is a nonmeasurable set. Let $C = \{(x, x) : x \in N\} \subset \mathbb{R}^2$ Then C_x and C_y are either singletons or empty and thus are in $\mathcal{B}$ for every x, y. Let $T : \mathbb{R} \to \mathbb{R}^2$ be $T(x) = (x, x)$. T is measurable because $T^{-1}A \times B = A \cap B \in \mathcal{B}$ for all $A \times B \in \mathcal{M}$. But $T^{-1}C = N \notin \mathcal{B}$ so $C \notin \mathcal{B} \times \mathcal{B}$. ◇◇

Product Measure. Next, we want to define a measure π on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ with the property that

$$\pi(A \times B) = \mu(A)\nu(B) \text{ for all } A \in \mathcal{X}, B \in \mathcal{Y}$$

This specifies π on $\mathcal{M}$, but not on all of $\mathcal{X} \times \mathcal{Y} = \sigma(\mathcal{M})$.

Theorem C. *If $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ are σ-finite measure spaces, then there is a unique measure π on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ with the property that $\pi(A \times B) = \mu(A)\nu(B)$ for all $A \in \mathcal{X}, B \in \mathcal{Y}$.*

One approach to proving this is to apply the Carathéodory Extension Theorem, I.D. Another is to define π directly by

$$\pi(C) = \int_X \nu(C_x)\, \mu(dx).$$

We need to show several things to justify this.

1. $f(x) = \nu(C_x)$ is $\mathcal{X}/\mathcal{B}$ measurable for all $C \in \mathcal{X} \times \mathcal{Y}$. If ν is finite this follows from the facts that $\mathcal{M}$ is a π-system and
$$\mathcal{L} = \{C \in \mathcal{X} \times \mathcal{Y} : \ \nu(C_x) \text{ is measurable}\} \text{ is a } \lambda\text{-system.}$$
If ν is σ-finite, take $A_n \uparrow Y$ in $\mathcal{Y}$ with $\nu(A_n) < \infty$ and apply the preceding to $\nu_n(\cdot) = \nu(A_n \cap \cdot)$: $\nu(C_x) = \lim \nu_n(C_x)$ is therefore measurable.
2. $\pi(C)$ defines a measure. This follows from monotone convergence.
3. $\pi(A \times B) = \mu(A)\nu(B)$ for all $A \times B \in \mathcal{M}$. Checking this is just a calculation:
$$(A \times B)_x = \begin{cases} B & \text{if } x \in A \\ \emptyset & \text{if } x \notin A \end{cases}, \quad \nu((A \times B)_x) = \nu(B) \cdot 1_A(x),$$
so the definition gives
$$\pi(A \times B) = \int 1_A \cdot \nu(B)\, d\mu = \mu(A)\nu(B).$$

The uniqueness follows from Theorem I.B .

We have observed that the product measure $\mu_X \times \mu_Y$ is the joint distribution of a pair of independent random vairables X, Y with marginal distributions μ_X, μ_Y. If instead of considering X, Y as given, suppose we know μ_X, μ_Y and want to construct a probability space $(\Omega, \mathcal{F}, P)$ and a pair of random variables $X, Y : \Omega \to \mathbb{R}$ which are independent with the prescribed marginals. Theorem C provides one way to do so: take $\Omega = \mathbb{R}^2$ with $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$ and $P = \mu_X \times \mu_Y$. A typical $\omega \in \Omega$ is $\omega = (x, y)$. Define the two random variables to be the "coordinate maps":

$$X(\omega) = X((x, y)) = x$$
$$Y(\omega) = Y((x, y)) = y$$

However what if we want to build a probability space on which are defined a full sequence $X_1, x_2, \ldots$ of independent random variables, each with a specified marginal distribution μ_i — how might we do this? We did it in Unit M using $\Omega = [0, 1)$ and $d_i(\omega)$ given by the digits in the decimal expansion of ω. In that case μ_i was just the simple Bernoulli distribution. What if we want something more complicated for the μ_i? The natural analogue of the above is to take

$$\Omega = \mathbb{R}^{\mathbb{N}} = \{\omega = (x_1, x_2, \ldots) : \ x_i \in \mathbb{R}\},$$

the set of all sequences of real numbers. We would define

$$X_i(\omega) = x_i$$

and $\mathcal{F} = \sigma(X_i : \ i = 1, 2, \ldots)$. For P we would want a sort of infinite product measure

$$P = \mu_1 \times \mu_2 \times \ldots .$$

The existence of this P, an infinite dimensional version of Theorem C, is itself a special case of the even more general Kolmogorov Existence Theorem, which also allows correlation among the X_i. See Billingsley for more on this.

THEOREM D. *Suppose $(X,\mathcal{X},\mu)$ and $(Y,\mathcal{Y},\nu)$ are σ-finite measure spaces and $\pi=\mu\times\nu$ is the product measure on $(X\times Y,\mathcal{X}\times\mathcal{Y})$. Suppose $f : X\times Y\to \mathbb{R}$ is $\mathcal{X}\times\mathcal{Y}$ measurable.*

1. (TONELLI) *If $f\ge 0$, then*
   - *a) $\int_Y f(x,y)\,\nu(dy)$ is $\mathcal{X}$ measurable,*
   - *b) $\int_X f(x,y)\,\mu(dx)$ is $\mathcal{Y}$ measurable, and*
   - *c)*

$$\int_X\left[\int_Y f(x,y)\,\nu(dy)\right]\,\mu(dx)=\int_{X\times Y} f(x,y)\,\pi(d(x,y))=\int_Y\left[\int_X f(x,y)\,\mu(dx)\right]\,\nu(dy).$$

2. (FUBINI) *If f is π-integrable, then a) and b) still hold except that the functions may be undefined on sets measure 0. The functions in a) and b) are integrable, and c) holds.*

Example 3. Let $X=Y=\{1,2,3,4,\cdots\}$, $\mu=\nu=$ counting measure, and

$$f(x,y)=\begin{cases}1 & \text{if } x=y\\ -1 & \text{if } y=x+1\\ 0 & \text{otherwise.}\end{cases}$$

$$\int\int f(x,y)\nu(y)\mu(dx)=\sum_{x=1}^{\infty}\sum_{y=1}^{\infty}f(x,y)=\sum_{x=1}^{\infty}(1-1)=0$$

$$\int\int f(x,y)\mu(dx)\nu(dy)=\sum_{y=1}^{\infty}\sum_{x=1}^{\infty}f(x,y)=1+\sum_{y=2}^{\infty}(-1+1)=1$$

The problem is that Fubini's Theorem does not apply!

$$\int\int |f|\nu(dy)\mu(dx)=\sum_{x=1}^{\infty}(1+1)=\infty$$

◇◇

The usual way Theorem D is used is to first calculate $\int_X\int_Y |f(x,y)|\nu(dy)\mu(dx)$ to verify that f is π-integrable. If that works out, then you are justified to remove the absolute values and evaluate $\int f\,d\pi=\int_X\int_Y f\,d\nu\,d\mu$.

Applications to Independent Random Variables

Suppose again that $X,Y:\Omega\to\mathbb{R}$ are independent random variables on $(\Omega,\mathcal{F},P)$. Then their joint distribution π is the product of the marginals: $\pi=\mu_X\times\mu_Y$. Theorem D and III.H now allow us to calculate

$$\begin{aligned}E[|XY|]&=\int |xy|\,\pi(d(x,y))\\&=\int |x|\,\mu_X(dx)\cdot\int |y|\,\mu_Y(dy)\quad\text{(Tonelli)}\\&=E[|X|]\cdot E[|Y|].\end{aligned}$$

So $E[|X|],E[|Y|]<\infty$ implies $E[|XY|]<\infty$, and repeating the calculation, now with Fubini, we get $E[XY]=E[X]E[Y]$.

Example 4. Suppose X and Y are independent, exponentially distributed r.v.s with mean $\frac{1}{\lambda}$. Find the distribution of X/Y. For $z \geq 0$ we can calculate

$$\begin{aligned}
P(\frac{X}{Y} \leq z) &= P[X \leq zY] = \pi(\{(x,y) : x \leq zy\}) \\
&= \int_{\rm I\!R} \int_{\rm I\!R} 1_{\{x \leq zy\}}(x,y)\, \mu_X(dx)\mu_Y(dy) \\
&= \int_{(0,\infty)} \int_{(0,\infty)} I_{\{x \leq zy\}}(x,y) \lambda e^{-\lambda y} \ell(dx) \lambda e^{-\lambda y} \ell(dy) \\
&= \int_{(0,\infty)} \left[\int_0^{zy} \lambda e^{-\lambda x}\, dx \right] \lambda e^{-\lambda y}\, \ell(dy) \\
&= \int_0^\infty \lambda e^{-\lambda y} [1 - e^{-\lambda zy}]\, dy \\
&= \int_0^\infty \lambda [e^{-\lambda y} - e^{-\lambda y(z+1)}]\, dy \\
&= 1 - \frac{\lambda}{\lambda(z+1)} \quad = \frac{z}{z+1}
\end{aligned}$$

For $z < 0$ clearly $P(\frac{X}{Y} \leq z) = 0$. Thus the distribution function of X/Y is

$$P(\frac{X}{Y} \leq z) = \begin{cases} \frac{z}{z+1} & \text{for } \; z \geq 0 \\ 0 & \text{for } z < 0. \end{cases}$$

This distribution has density given by

$$p(z) = \begin{cases} \frac{1}{(z+1)^2} & z \geq 0 \\ 0 & z < 0 \end{cases}$$

◇◇

If $M_X(s) = E[e^{sX}]$ and $M_Y(s) = E[e^{sY}]$ are both defined, then

$$\begin{aligned}
M_{X+Y}(s) = E[e^{s(X+Y)}] &= \int e^{sx} \cdot e^{sy} \pi(d(x,y)) \\
&= \int e^{sx} \mu_X(dx) \cdot \int e^{sy} \mu_Y(dy) = M_X(s) M_T(s) < \infty.
\end{aligned}$$

The same holds regarding characteristic functions: for all t,

$$\widehat{\mu}_{X+Y}(t) = \widehat{\mu}_X(t) \cdot \widehat{\mu}_Y(t).$$

Suppose that μ_X and μ_Y have densities f_X, f_Y. Then for $A \in \mathcal{B}({\rm I\!R}^2)$ we have

$$\begin{aligned}
\pi(A) &= \int \int 1_A(x,y)\, \nu_Y(dy) \mu_X(dx) \\
&= \int_{\rm I\!R} \left[\int_{\rm I\!R} 1_A(x,y) f_Y(y)\, \ell(dy) \right] f_X(x)\, \ell(dx) \\
&= \int_{{\rm I\!R}^2} 1_A(x,y) f_X(x) f_Y(y)\, \ell(d(x,y)).
\end{aligned}$$

Thus π has density $f(x,y) = f_X(x) f_Y(y)$ with respect to ℓ on ${\rm I\!R}^2$. Moreover we can calculate a density for the distribution of $X+Y$:

$$F_{X+Y}(a) = \int_{\rm I\!R} \int_{\rm I\!R} 1_{\{x+y \leq a\}}(x,y) f_Y(y) f_X(x)\, \ell(dy)\, \ell(dx).$$

Now $\int_{(-\infty,a-x]} f_Y \, d\ell = \int_{(-\infty,a]} f_Y(z-x)\,\ell(dz)$. For Riemann integration this would just be the simple change of variables $y = z - x$, $dy = dz$. In our context we have to work harder to give a correct justification. Consider $T(z) = z - x$. By checking intervals $(a, b]$ you can confirm that then $\ell T^{-1} = \ell$. (This is problem I.8.) Now using III.H it follows that

$$\begin{aligned}\int_{(-\infty,a]} f_Y(z-x)\,\ell(dz) &= \int 1_{(-\infty,a-x]}(T(z)) f_Y(T(z))\,\ell(dz)\\ &= \int 1_{(-\infty,a-x]}(y) f_Y(y)\,\ell T^{-1}(dy)\\ &= \int 1_{(-\infty,a-x]}(y) f_Y(y)\,\ell(dy)\\ &= \int_{(-\infty,a-x]} f_Y \, d\ell,\end{aligned}$$

as claimed. Therefore

$$\begin{aligned}F_{X+Y}(a) &= \int_{\mathbb{R}} \int_{(-\infty,a]} f_Y(z-x) f_X(x)\,\ell(dz)\ell(dx)\\ &= \int_{(-\infty,a]} \left[\int_{\mathbb{R}} f_Y(z-x) f_X(x)\,\ell(dx)\right] \ell(dz).\end{aligned}$$

Thus μ_{X+Y} has density

$$f_X * f_Y(z) = \int_{\mathbb{R}} f_Y(z-x) f_X(x)\,\ell(dx).$$

This is called the *convolution* of the density functions f_X and f_Y.

The Radon-Nikodym Theorem

If μ, ν are two measures on the same $(\Omega, \mathcal{F})$ we have said that ν has density ρ with respect to μ if $\rho : \Omega \to \mathbb{R}$ is measurable, $\rho \geq 0$, and

$$\nu(A) = \int_A \rho \, d\mu, \quad \text{all } A \in \mathcal{F}.$$

If we are given μ and ν how might we check to see if such a ρ exists? (This will be the foundation of conditional probabilities!) Note that <u>if</u> a density exists then for any $A \in \mathcal{F}$ with $\mu(A) = 0$ we must also have $\nu(A) = \int 1_A \, \rho d\mu = 0$. In other words

$$\nu(A) = 0 \text{ whenever } A \in \mathcal{F} \text{ and } \mu(A) = 0. \tag{1}$$

When (1) holds we say that ν is *absolutely continuous* with respect to μ, written "$\nu \ll \mu$".

Example 5. $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, $\nu(A) = \sum_{-\infty}^{\infty} 1_A(n)$, counting measure on the integers, Z. $\nu \not\ll \ell$ because $\ell(Z) = 0$ but $\nu(Z) = \infty$. Also $\ell \not\ll \nu$ because $\ell((0,1)) = 1$ but $\nu((0,1)) = 0$. Let $\mu = \ell + \nu$. Then $\mu(A) = 0$ implies $\nu(A) = 0$ so $\nu \ll \mu$. In fact $\nu(A) = \int_A 1_Z d\mu$ so ν has density 1_Z with respect to μ. ◇◇

Example 6. Let P and Q be two probability measures on $(\Omega, \mathcal{F})$ and suppose X_i, $i = 1, 2, \ldots$ are i.i.d. with respect to both P and Q. Then $Q \ll P$ requires the distribution of X_i to be the same under P as under Q! (Absolute continuity is more difficult in infinite dimensional settings.) To see why, suppose for some a

$$p = P(X_i \leq a) \neq Q(X_i \leq a) = q.$$

Then the Strong Law of Large Numbers says that

$$\frac{1}{n}\sum_1^n 1_{(-\infty,a]}(X_i) \to \begin{cases} p & P \text{ a.s.} \\ q & Q \text{ a.s.} \end{cases}$$

◇◇

We have explained that if ν has density with respect to μ, then $\nu \ll \mu$. The next important theorem says that (for σ-finite measures) the converse is also true!

THE RADON-NIKODYM THEOREM (E). *If μ and ν are both σ-finite measures on $(\Omega, \mathcal{F})$ with $\nu \ll \mu$, then ν has a density with respect to μ. The density is unique up to sets of μ-measure* 0.

The density (lets call it ρ) is often called the *Radon-Nikodym derivative* of ν with respect to μ and is indicated by the notation

$$\rho = \frac{d\nu}{d\mu}.$$

(This notation agrees nicely with "$d\nu = \rho\, d\mu$" as in Theorem III.G.) Notice that if $g \geq 0$ is any other measurable function on Ω with $\rho = g$ μ-almost surely, then $\nu(A) = \int_A \rho\, d\mu = \int_A g\, d\mu$. Thus g deserves to be called the density just as much as ρ. In other words, densities are only determined "up to" almost sure equivalence.

Conditioning

Let $(\Omega, \mathcal{F}, P)$ be a probability space. What do we mean by a conditional probability? The next three examples review some "primitive" formulas.

Example 7. If $A, B \in \mathcal{F}$ and $P(B) > 0$ then

$$P[A|B] = \frac{P(A \cap B)}{P(B)}.$$

In other words $P[A|B] = \rho$ is the value that makes the formula

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

correct. ◇◇

Example 8. Suppose Y is a simple random variable.

$$P[A|Y = y_i] = P[A|B_i] \quad \text{where } B_i = \{Y = y_i\},$$

provided $P(Y = y_i) > 0$. ◇◇

Example 9. Suppose X, Y have joint density $f(x, y)$. Then

$$P(Y \in B) = \int_{\mathbb{R} \times B} f(x, y)\, \ell(d(x, y)).$$

$P(Y = y) = 0$ for any individual y, so $P[X \in A|Y = y]$ can't be defined as above. But most elementary texts will define the conditional density

$$f_{X|Y}(x|y) = \frac{f(x, y)}{\int f(x, y)\, \ell(dx)}.$$

Then if $A = \{X \in C\}$ where $C = (a, b]$, you would calculate

$$P[A|Y = y] = P[a < X \leq b|Y = y] = \int_a^b f_{X|Y}(x|y)\, dx.$$

Or,

$$P[X \in C|Y = y] = \int_C f_{X|Y}(x|y)\, \ell(dx).$$

◇◇

What is the unifying idea behind these definitions? The key is to focus on how they depend on what we condition with respect to. In the above examples define (respectively)

(Ex.6) $$\rho(\omega)=\begin{cases}P[A|B] & \omega\in B\\ P[A|B^c] & \omega\in B^c\end{cases}, \qquad \mathcal{G}=\{\emptyset,\Omega,B,B^c\}$$

(Ex.7) $$\rho(\omega)=\Sigma P[A|Y=y_i]1_{\{Y=y_i\}}(\omega), \qquad \mathcal{G}=\sigma(Y)$$

(Ex.8) $$\rho(\omega)=\int_C f_{X|Y}(x,Y(w))\ell(dx), \qquad \mathcal{G}=\sigma(Y) \quad A=\{X\in C\}$$

In each case a function $\rho(\omega)$ and σ-field $\mathcal{G}$ are defined (except possibly on a set of probability 0) and are characterized by the following properties:

(i) $\rho:\Omega\to\mathbb{R}$ is $\mathcal{G}$ measurable;

(ii) $P(A\cap B)=\int_B \rho\,dP$ for all $B\in\mathcal{G}$.

These properties provide the general definition.

Definition. *Suppose $(\Omega,\mathcal{F},P)$ is a probability space and $\mathcal{G}\subseteq\mathcal{F}$ is a sub-σ-field. If $A\in\mathcal{F}$, then $P[A|\mathcal{G}]$ is defined to be any $\mathcal{G}$ measurable random variable $\rho:\Omega\to\mathbb{R}$ with the property that*

$$\int_B \rho\,dP=P(A\cap B)$$

for all $B\in\mathcal{G}$.

Thus $P[A|\mathcal{G}](\omega)=\rho(\omega)$ is a measurable function – this usually takes some getting used to. The examples above describe its values in certain simple settings. In Example 7, to be $\mathcal{G}$ measurable means $\rho(\omega)=c_1 1_B+c_2 1_{B^c}$where $c_1=P[A|B]$ and $c_2=P[A|B^c]$. I.e. $P[A|B]$ is the constant value of $\rho=P[A|\mathcal{G}]$ over B. Similarly in Example 8, to be $\mathcal{G}$ measurable requires ρ be constant over each set $\{Y=y_i\}$. The value it takes on this set is what we denoted by $P[A|Y=y_i]$ in Example 8.

Example 9 (continued). Define

$$\psi(y)=\int_C f_{X|Y}(x|y)\,\ell(dx).$$

We want to show that $\rho(\omega)=\psi(Y(\omega))$ satisfies the definition we have given. With $\mathcal{G}=\sigma(Y)$ ρ will be $\mathcal{G}$ measurable if ψ is $\mathcal{B}$ measurable from $\mathbb{R}\to\mathbb{R}$. First, define

$$f_Y(y)=\int f(x,y)\,\ell(dx).$$

Tonelli's Theorem says that f_Y is $\mathcal{B}$ measurable. The careful definition of $f_{X|Y}(x,y)$ would be

$$f_{X|Y}(x,y)=\begin{cases}\frac{f(x,y)}{f_Y(y)} & \text{if } f_Y(y)\neq 0\\ 0 & \text{if } f_Y(y)=0\end{cases}.$$

(Note that $f(x,y)=f_{X|Y}(x,y)f_Y(y)$ a.e., because if $N=\{f_Y(y)=0\}$ then $\int_N\int_{\mathbb{R}} f(x,y)\,\ell(dx)\,\ell(dy)=\int_N f_Y(y)\,\ell(dy)=0$.) Another application of Tonelli's Theorem yields that

$$\psi(y)=\int_C f_{X|Y}(x,y)\,\ell(dx)$$

is $\mathcal{B}$ measurable as desired.

Next we need to show that

$$\int_B \rho\,dP=P(A\cap B)\quad\text{for all } B\in\mathcal{G}.$$

Recall that $A = \{X \in C\}$. First note that f_Y is the density of the distribution of Y; i.e. the marginal density:

$$\begin{aligned}\mu_Y(D) &= P(Y \in D)\\ &= \int 1_D(y) f(x,y)\,\ell(d(x,y))\\ &= \int\int 1_D(y) f(x,y)\,\ell(dx)\,\ell(dy)\\ &= \int_D f_Y(y)\,\ell(dy).\end{aligned}$$

Now, if $B \in \mathcal{G}$ then $B = Y^{-1}D$ for some $D \in \mathcal{B}$. We can now use III.G and F to verify the following sequence of equations.

$$\begin{aligned}\int_B \rho(\omega)\,P(d\omega) &= \int_{Y^{-1}D} \psi(Y(\omega))\,P(d\omega)\\ &= \int_D \psi(y)\,\mu_Y(dy)\\ &= \int_D \left[\int_C f_{X|Y}(x|y)\,\ell(dx)\right] f_Y(y)\,\ell(dy)\\ &= \int_D\int_C f(x,y)\,\ell(dx)\ell(dy)\\ &= \int_{C\times D} f(x,y)\,\ell(d(x,y))\\ &= P(X \in C, Y \in D)\\ &= P(A \cap B)\end{aligned}$$

Therefore, $\psi(Y(\omega)) = P[A|\sigma(Y)](\omega)$. ◇◇

Definition. *If X is an integrable random variable, we define $E[X|\mathcal{G}](\omega) = \rho(\omega)$ if $\rho : \Omega \to I\!R$ is $\mathcal{G}$ measurable and*

$$\int_B \rho\,dP = \int_B X\,dP \quad \textit{for all } B \in \mathcal{G}.$$

In other words, $\rho = E[X|\mathcal{G}]$ is a partially averaged/smoothed version of X making it $\mathcal{G}$ measurable but maintaining it integrated values over $\mathcal{G}$-sets. Observe that

$$P[A|\mathcal{G}] = E[1_A|\mathcal{G}].$$

Also note that $P[A|\mathcal{G}]$ and $E[X|\mathcal{G}]$ are only defined up to $\mathcal{G}$ measurable sets of $P = 0$. We can have different *versions* $\rho_1(\omega)$ and $\rho_2(\omega)$ of $E[X|\mathcal{G}](\omega)$ (but both must be $\mathcal{G}$ measurable).

Example 8 (continued). If $P(Y = y_i) = 0$ in Example 8, then $E[A|\sigma(Y)]$ may be given any value on $\{Y = y_i\}$. ◇◇

Example 10. If $\mathcal{G} = \{\emptyset, \Omega\}$, then $E[X|\mathcal{G}] \equiv E[X]$ and $P[A|\mathcal{G}] \equiv P(A)$ are constant functions. ◇◇

One might ask, "does $E[X|\mathcal{G}]$ always exist?" Lets assume X is integrable. If $X \geq 0$ then $\nu(B) = \int_B X\,dP$ defines a finite measure on $(\Omega, \mathcal{G})$ and $\nu \ll \mu$ where $\mu = P|_{\mathcal{G}}$. The Radon-Nikodym says there does exist a $\mathcal{G}$ measurable ρ with

$$E[X; B] = \nu(B) = \int_B \rho\,d\mu = \int_B \rho\,dP.$$

So, using the R.N. derivative notation, we can say

$$E[X|\mathcal{G}] = \frac{d\nu}{dP|_{\mathcal{G}}}.$$

In general $E[X|\mathcal{G}] = E[X^+|\mathcal{G}] - E[X^-|\mathcal{G}]$ works.

Elementary Properties. Conditional probabilities and expectations obey many of the same properties as ordinary probabilities and expectations (integrals). We assume all random variables mentioned here are integrable.

- $0 \le P[A|\mathcal{G}] \le 1$ almost surely.
- If A_n are disjoint, then $P[\cup_1^\infty A_n|\mathcal{G}] = \sum_1^\infty P[A_n|\mathcal{G}]$ a.s.
- If $X = c$ a.s. then $E[X|\mathcal{G}] = c$ a.s.
- $E[\alpha X + \beta Y|\mathcal{G}] = \alpha E[X|\mathcal{G}] + \beta E[Y|\mathcal{G}]$ a.s. (any $\alpha, \beta \in \mathbb{R}$)
- If $X \le Y$ a.s., then $E[X|\mathcal{G}] \le E[Y|\mathcal{G}]$ a.s.
- $|E[X|\mathcal{G}]| \le E[|X||\mathcal{G}]$ a.s.
- If $|X_n| \le Y$ a.s. and $X_n \to X$ a.s. then $E[X_n|\mathcal{G}] \to E[X|\mathcal{G}]$ a.s.
- If $\phi(\cdot)$ is a convex function, with $\phi(X)$ and X both integrable, then

$$\phi(E[X|\mathcal{G}]) \le E[\phi(X)|\mathcal{G}] \text{ a.s.}$$

The next two properties are very important tools for manipulating conditional expectations.

- If X is $\mathcal{G}$ measurable and Y, XY are both integrable, then

$$E[XY|\mathcal{G}] = X \cdot E[Y|\mathcal{G}]. \tag{2}$$

- If $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$ are nested sub-σ-fields, then

$$E[X|\mathcal{G}_1] = E[E[X|\mathcal{G}_2]|\mathcal{G}_1]. \tag{3}$$

Why are these true? (2) means that if $\rho(\cdot) = E[Y|\mathcal{G}]$ then $\int_A X\rho\, dP = \int_A XY\, dP$ all $A \in \mathcal{G}$. Note if $X = \sum_1^n c_i 1_{B_i}$with $B_i \in \mathcal{G}$, then

$$\int_A X\rho\, dP = \sum_1^n c_i \int_{A\cap B_i} \rho\, dP = \sum_1^n c_i \int_{A\cap B_i} Y\, dP = \int_A XY\, dP.$$

Now pass to limit from simple approximants: $|X_n| \le |X|$ with $X_n \to X$ and $\mathcal{G}$-measurable. Actually the mechanics of this are somewhat tricky. We want to apply the Dominated Convergence Theorem on each side, using $|X_n||Y| \le |X||Y|$ on the right and $|X_n||\rho| \le |X||\rho|$ on the left. We know that $|XY|$ is integrable; the tricky part is to show that $|X||\rho|$ is also integrable. Since $|X_n|$ is simple, what we have already argued justifies

$$E[|X_n||\rho|] \le E[|X_n|E[|Y|\,|\mathcal{G}]] = E[|X_n||Y|] \le E[|XY|] < \infty.$$

Now applying Fatou's Lemma to this tells us that

$$E[|X||\rho|] \le E[|XY|] < \infty.$$

Justifying (3) is easier. first notice that $E[E[X|\mathcal{G}_2]|\mathcal{G}_1]$ is $\mathcal{G}_1$-measurable. Now we check that for any $A \in \mathcal{G}_1$

$$\begin{aligned}\int_A E[E[X|\mathcal{G}_2]|\mathcal{G}_1]\, dP &= \int_A E[X|\mathcal{G}_2]\, dP \\ &= \int_A X\, dP \quad \text{since } A \in \mathcal{G}_2 \text{ also.}\end{aligned}$$

If $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ are independent and $B \in \mathcal{H}$, then

$$P(A \cap B) = P(A)P(B) = \int_A P(B)\, dP$$

for all $A \in \mathcal{G}$. I.e. $P[B|\mathcal{G}] \equiv P(B)$. If X is independent of $\mathcal{G}$ (i.e. $\sigma(X)$ and $\mathcal{G}$ are independent) then $E[X|\mathcal{G}] \equiv E[X]$. (See problem 10.)

Sufficient Statistics

In statistics we often observe various events $\subseteq \Omega$ and attempt to draw conclusions re. P. Suppose we are trying to identify the true P from among a collection $\{P_\theta : \ \theta \in \Theta\}$ of possibilities. A *sufficient statistic* is a random variable $T(\omega)$ so although $\theta \in \Theta$ affects the distribution of T, θ does $\underline{\text{not}}$ affect the conditional probabilities given T: for each $A \in \mathcal{F}$, i.e.

$$P_\theta[A|T] = \rho(\omega) \tag{4}$$

should be $\sigma(T)$ measurable but not depend on θ. Thus for T to be sufficient should mean that for each $A \in \mathcal{F}$ there exists a measurable f giving a θ-independent version probabilities of A given T:

$$P_\theta[A|T] = f(T(\omega)) \quad \text{for all } \theta \in \Theta,$$

Factorization Theorem(F). *Suppose μ is a σ-finite measure on $(\Omega, \mathcal{F})$ and $P_\theta \ll \mu$ for every θ. A necessary and sufficient condition that a random variable (or vector) T be sufficient for $\{P_\theta\}$ is that there exist a measurable $h : \Omega \to [0, \infty)$ and for each θ a (Borel) measurable $g_\theta : \mathbb{R} \to [0, \infty)$ so that*

$$\frac{dP_\theta}{d\mu} = h(\omega) g_\theta(T(\omega)).$$

.

Example 11. The "exponential families" of Bickel and Doksum describe commonly occuring parameterized families P_θ for which a sufficient statistic exists, by appeal to the Factorization Theorem. Here $T(x)$ and $c(\theta)$ are vector valued and we write $c(\theta) \cdot T(x)$ in place of $\sum_i \ c_i(\theta) T_i(x)$:

$$\begin{aligned}
p(x, \theta) &= \exp\left[c(\theta) \cdot T(x) + d(\theta) + S(x)\right] 1_A(x) \\
&= \exp\left[c(\theta) \cdot T(x) + d(\theta)\right] \cdot \exp\left[S(x)\right] 1_A(x) \\
&= g_\theta(T(x)) \cdot h(x)
\end{aligned}$$

When interpreted as a density the reference mesaures is $\mu = \ell$ (on $\mathbb{R}$ or $\mathbb{R}^d$). When interpreted as a "frequency function" the reference measure μ is counting measure on the appropriate set e.g. $A = \{1, 2, 3, \dots\}$.) $\diamond\diamond$

Example 12. Let X_1, X_2 be a pair of independent random variables each with uniform distribution on $[0, B]$. We view $B\,(= \theta) \in (0, \infty)\,(= \Theta)$ as a parameter for wihich we want to consider a sufficient statistic. P_B will be the distribution of $X = (X_1, X_2)$, a probability measure defined on $\Omega = \mathbb{R}^2$ with $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$. A typical $\omega \in \Omega$ is $\omega = (x_1, x_2)$ and $X_i(\omega) = x_i$.

Take ℓ as the common reference measure on Ω. Then $P_B \ll \ell$ with

$$\frac{dP_B}{d\ell} = \frac{1}{B^2} 1_{[0,B]}(X_1(\omega)) 1_{[0,B]}(X_2(\omega)).$$

Define

$$M(\omega) = \max(X_1(\omega), X_2(\omega)), \quad m(\omega) = \min(X_1(\omega), X_2(\omega)).$$

Then we can rewrite the density

$$\begin{aligned}
\frac{dP_B}{d\ell} &= \frac{1}{B^2} 1_{[0,\infty)}(m(\omega)) 1_{[0,B]}(M(\omega)) \\
&= h(\omega) g_B(M(\omega)),
\end{aligned}$$

as in the Factorization Theorem. Thus $M = \max(X_1, X_2)$ is a sufficient statistic. However $m = \min(X_1, X_2)$ is not. In other words

1) $P_B(A|\sigma(M))$ should have a B-independent version for any $A \in \mathcal{F}$;
2) $P_B(A|\sigma(m))$ should fail to have a B-independent version for some $A \in \mathcal{F}$;

We would like to see these features in terms of some more explicit calculations. For this purpose we will use the joint density for (m, M). By checking various cases you can convince yourself that for $0 \le \alpha, \beta \le B$

$$P_B[m \le \alpha;\ M \le \beta] = \int_{[0,\alpha]} \int_{[0,\beta]} \phi(u,v)\, d\ell(u,v),$$

where

$$\phi(u,v) = \begin{cases} \frac{2}{B^2} & 0 \le u \le v \le B \\ 0 & \text{otherwise.} \end{cases}$$

This is therefore the density.

For an example of 1), take $A = \{\omega :\ m(\omega) \le 1\}$. Following Example 9 we calculate

$$\phi_{m|M}(u|v) = \frac{1}{v} 1_{[0,v]}(u) \text{ for } 0 \le u, v \le B \text{ (0 otherwise).}$$

From here we find

$$P_B(A|\sigma(M)) = \begin{cases} \frac{1}{M(\omega)} & \text{if } 1 \le M(\omega) \\ 1 & \text{if } M(\omega) \le 1 \end{cases}$$

We see no dependence on B, as expected.

For an example of 2), take $A = \{\omega :\ M(\omega) \le 1\}$. Now we start with

$$\phi_{M|m}(v|u) = \frac{1}{B-u} 1_{[u,B]}(v) \text{ for } 0 \le u, v \le B \text{ (0 otherwise)}$$

and find

$$P_B(A|\sigma(m)) = \frac{1}{B - m(\omega)} \begin{cases} 1 - m(\omega) & \text{if } m(\omega) \le 1 \\ 0 & \text{if } m(\omega) > 1. \end{cases}$$

This obviously does depend on B. ◇◇

We will present a proof of the Factorization Theorem, since it exercises our understanding of conditional constructions and their manipluations. We start by assuming that $dP_\theta/d\mu = h(\omega)g_\theta(T)$ and prove that T is sufficient. For this we start with a special case: μ is a probability measure and h is integrable with respect to μ. For a given $A \in \mathcal{F}$ we need to exhibit a θ-independent version of $P_\theta(A|\mathcal{F})$. We will show that the following works. Let $N = \{E^\mu[h|T] = 0\}$ and define

$$\rho = 1_N \frac{E^\mu[1_A h|T]}{E^\mu[h|T]}.$$

Since $N \in \sigma(T)$ ρ is $\sigma(T)$ measurable. Notice that ρ is defined independent of any θ. It will follow that T is sufficient if we can verify that ρ satisfies the integral identity that defines $\rho = P_\theta(A|\mathcal{F})$. Notice that for any θ we have

$$\begin{aligned} P_\theta(N) &= E^\mu[hg_\theta(T); N] \\ &= E^\mu[E^\mu[hg_\theta(T)|T]; N] \\ &= E^\mu[g_\theta(T)E^\mu[h|T]; N] \\ &= 0. \end{aligned}$$

Consider any $B \in \sigma(T)$. Since $P_\theta(B \cap N) = 0$ we have

$$\begin{aligned}\int_B \rho\, dP_\theta &= \int_{B\cap N^c} \frac{E^\mu[1_A h|T]}{E^\mu[h|T]} h g_\theta(T)\, d\mu + 0 \\ &= \int_{B\cap N^c} E^\mu[1_A h|T] g_\theta(T)\, d\mu \\ &= \int_{B\cap N^c} 1_A h g_\theta(T)\, d\mu = P_\theta(A \cap B \cap N^c) = P_\theta(A\cap B).\end{aligned}$$

We will prove the general case by reducing it to the special case of μ a probability measure. The following lemma is the key to that reduction.

Lemma. *If $P_\theta \ll \mu$ for all θ where μ is σ-finite, there exists a countable collection $\theta_1, \theta_2, \ldots$ so that $P_{\theta_n}(A) = 0$ for all n is equivalent to $P_\theta(A) = 0$ for all θ.*

We will prove this lemma after using it to complete the proof of the theorem.

Let $\theta_1, \theta_2, \ldots$ be as in the lemma and define a new probability mesaure by

$$Q(A) = \sum_1^\infty 2^{-n} P_{\theta_n}(A).$$

Notice that $Q(A) = 0$ implies that $P_{\theta_n}(A) = 0$ for all n and therefore $P_\theta(A) = 0$ for all θ. I.e. $P_\theta \ll Q$ for all θ. In addition, if $dP_\theta/d\mu = h(\omega) g_\theta(T(\omega))$, define

$$f = \sum_1^\infty 2^{-n} g_{\theta_n}.$$

Then $dQ/d\mu = h \cdot f(T)$, so that f plays the role of g_Q. Next for each θ define

$$r_\theta = \begin{cases} g_\theta/f & \text{if } f > 0 \\ 0 & \text{if } f = 0. \end{cases}$$

We will show that $dP_\theta/dQ = r_\theta(T(\omega)) \cdot 1$ for each θ. This means that the argument for the special case above applies (with Q in the pace in μ and $1 = h$) to see that T is sufficient. Let

$$C = \{\omega : \ f(T(\omega)) = 0\}.$$

Then $Q(C) = 0$ so that $P_\theta(C) = 0$ for all θ. For $A \subseteq C^c$ we have

$$\begin{aligned}\int_A r_\theta(T(\omega))\, dQ &= \int_A \frac{g_\theta(T)}{f(T)} f(T) h\, d\mu \\ &= \int_A g_\theta(T) h\, d\mu = P_\theta(A).\end{aligned}$$

This completes the proof of the sufficiency half of the theorem.

We now turn to the necessity assertion. Suppose that T is sufficient. We must prove the existence of an appropriate factorization of $dP_\theta/d\mu$. Let θ_n be as in the lemma again and Q the same as we constructed above. Let $A \in \mathcal{F}$ and

$$\rho(T(\omega)) = P_\theta(A|\sigma(T))$$

be a θ-independent version, which des exist by the sufficiencey assumption. Consider any $B \in \sigma(T)$.

$$\begin{aligned}\int_B \rho(T)\, dQ &= \sum_1^\infty 2^{-n} \int_B \rho(T)\, dP_\theta \\ &= \sum_1^\infty 2^{-n} P_{\theta_n}(A \cap B) = Q(A \cap B).\end{aligned}$$

Thus $\rho(T) = Q(A|\sigma(T))$ as well.

Let

$$d_\theta = \frac{dP_\theta}{dQ} \quad \text{and} \quad g_\theta(T) = E^Q[d_\theta|\sigma(T)].$$

We will show that $g_\theta(T) = dP_\theta/dQ$. Consider any $A \in \mathcal{F}$.

$$\begin{aligned}
\int_A g_\theta(T)\,dQ &= \int 1_A g_\theta(T)\,dQ \\
&= \int E^Q[1_A g_\theta(T)|\sigma(T)]\,dQ \\
&= \int E^Q[1_A|\sigma(T)]E^Q[d_\theta|\sigma(T)]\,dQ \\
&= \int E^Q[E^Q[1_A|\sigma(T)]d_\theta|\sigma(T)]\,dQ \\
&= \int E^Q[1_A|\sigma(T)]d_\theta\,dQ \\
&= \int \rho(T)d_\theta\,dQ \\
&= \int P_\theta(A|\sigma(T))\,dP_\theta = P_\theta(A).
\end{aligned}$$

This confirms that $g_\theta(T) = dP_\theta/dQ$. Therefore we have

$$\begin{aligned}
\frac{dP_\theta}{d\mu} &= g_\theta(T)\frac{dQ}{d\mu} \\
&= g_\theta(T(\omega))h(\omega),
\end{aligned}$$

where $h = dQ/d\mu$. This establishes the desired factorization and completes the proof of the theorem.

Proof of the Lemma. Because μ is σ-finite there exist $A_n \in \mathcal{F}$ with $\mu(A_n) < \infty$ and $\Omega = \cup A_n$. We can assume the A_n are disjoint. Define a new measure ν by $d\nu = k\,d\mu$ where

$$k(\omega) = \begin{cases} 2^{-n}/\mu(A_n) & \text{if } \omega \in A_n,\ \mu(A_n) > 0 \\ 0 & \text{if } \omega \in A_n,\ \mu(A_n) = 0. \end{cases}$$

Since $\nu(\Omega) \le \sum_1^\infty 2^{-n} = 1$, ν is a finite measure. Now suppose $\nu(A) = 0$. For any A_n with $\mu(A_n) > 0$, $k(\omega) > 0$ for $\omega \in A_n$. Since $\nu(A_n \cap A) = 0$ we conclude that $\mu(A_n \cap A) = 0$ as well. Thus $\nu(A) = 0$ implies $\mu(A_n \cap A) = 0$ for all n, so that $\mu(A) = 0$ and therefore $P_\theta(A) = 0$ for all θ. In other words, we can replace μ by the finite measure ν.

Let

$$f_\theta = \frac{dP_\theta}{d\nu} \quad \text{and} \quad S_\theta = \{\omega :\ f_\theta(\omega) > 0\}.$$

We want to pick a countable $C \subseteq \Theta$ so that $\nu(\cup_{\theta \in C} S_\theta)$ is as large as possible. Let

$$\alpha = \sup\{\nu(\cup_C S_\theta) :\ \text{countable } C \subseteq \Theta\}.$$

Then $\alpha \le \nu(\Omega) < \infty$ is finite. For each n there is a countable C_n with

$$\nu(\cup_{C_n} S_\theta) > \alpha - \frac{1}{n}.$$

Then $C_* = \cup_1^\infty C_n$ is also countable and for each n we have

$$\alpha - \frac{1}{n} < \nu(\cup_{C_n} S_\theta) \le \nu(\cup_{C_*} S_\theta) \le \alpha.$$

Therefore

$$\nu(\cup_{C_*} S_\theta) = \alpha.$$

We will now show that C_* does what we want, i.e. that $P_\theta(A) = 0$ for all $\theta \in C_*$ implies $P_\theta(A) = 0$ for all θ. Let

$$K = \cup_{C_*} S_\theta.$$

If $P_{\theta'}(K^c) > 0$ for some θ', then $0 < \int_{K^c} f_{\theta'}\, d\nu$ implies $\nu(S_{\theta'} \cap K) > 0$ so that $\theta' \notin C_*$. By adding θ' to C_* we would increase the measure of $\nu(\cup_{C_*} S_\theta)$, which is not possible. We conclude that $P_\theta(K^c) = 0$ for all θ.

Suppose then that $P_\theta(A) = 0$ for all $\theta \in C_*$. Write

$$A \cap K = \cup_{\theta \in C_*} (A \cap S_\theta)$$

and consider a $\theta \in C_*$. Since $P_\theta(A) = 0$ we know

$$0 = P_\theta(A \cap S_\theta) = \int_{S_\theta} 1_A f_\theta \, d\nu.$$

Since $f_\theta > 0$ on S_θ, it follows that $\nu(A \cap S_\theta) = 0$ for all $\theta \in C_*$. Therefore $\nu(A \cap K) = 0$ which implies that $P_\theta(A \cap K) = 0$ for all θ. Since we already know $P_\theta(K^c) = 0$ for all θ, we conclude that $P_\theta(A) = 0$ for all θ, as desired.

Basics of Stochastic Processes

Suppose $(\Omega, \mathcal{F}, P)$ be a probability space. A stochastic process is intended to model a situation in which a random quantity as well as the information available to us both evolve in time. Time can be considered discrete or continuous.

Discrete Time. $n = 0, 1, 2, 3, \ldots$. The information available to us at the various times is described by an increasing sequence of σ-fields (called the "filtration"):

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \cdots \subseteq \mathcal{F}.$$

The evolving random quantity is described by a sequence of random variables,

$$X_0,\ X_1\, X_2,\ X_3,\ \ldots$$

such that X_n is $\mathcal{F}_n$ measurable. (I.e. we know X_n at time n.)

Continuous Time. $t \geq 0$. Now the filtration consists of a time-indexed family of σ-fields: $\mathcal{F}_t$, $t \geq 0$. We assume

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}, \quad \text{when } s \leq t.$$

The process itself should consist of random variables X_t, each being $\mathcal{F}_t$ measurable.

We will introduce two general types of stochastic processes here, considering primarily the discrete time case. Each of these two types, defined below, captures a certain type of dependency or evolutionary relationship between the successive X_n.

Definintion. *X_n is a martingale with respect to the $\mathcal{F}_n$ if for each n $E[|X_n|] < \infty$ and*

$$E[X_{n+1} | \mathcal{F}_n] = X_n.$$

Definition. *X_n is a Markov process with respect to the $\mathcal{F}_n$ if for each n and evey $A \in \mathcal{B}(I\!R)$ it is possible to express $E[X_{n+1} \in A|\mathcal{F}_n]$ as a function of X_n alone.*

There are additional technical hyopotheses in the continuous time case, but the essential properties are

$$E[X_t|\mathcal{F}_s] = X_s, \quad s < t \quad \text{(for martingales)},$$

and

$$E[X_t \in A|\mathcal{F}_s] \text{ depends only on } X_s, \quad s < t \quad \text{(for Markov processes)}.$$

There are also important connections – Markov process problems can often be formulated as martingale problems. We will just present a few examples and applications to illustrate the use of these types of processes.

Martingales

Example 13. Suppose $\xi_1, \xi_2, \ldots$ are i.i.d. random variables defined on $(\Omega, \mathcal{F}, P)$ with $E[\xi_i] = 0$. Let

$$\mathcal{F}_1 = \sigma(\xi_1), \quad \mathcal{F}_2 = \sigma(\xi_1, \xi_2), \quad \ldots \mathcal{F}_n = \sigma(\xi_1, \xi_2, \ldots, \xi_n),$$

and

$$X_n = \sum_1^n \xi_i = X_{n-1} + \xi_n.$$

Then X_n in $\mathcal{F}_n$ measurable, and

$$\begin{aligned} E[X_{n+1}|\mathcal{F}_n] &= E[X_n + \xi_{n+1}|\mathcal{F}_n] \\ &= E[X_n|\mathcal{F}_n] + E[\xi_{n+1}|\mathcal{F}_n] \end{aligned}$$

We know $E[X_n|\mathcal{F}_n] = X_n$ since X_n is $\mathcal{F}_n$ measurable, and $E[\xi_{n+1}|\mathcal{F}_n] = E[\xi_{n+1}] = 0$ since ξ_{n+1} is independent of $\mathcal{F}_n$. Therefore X_n is a martingale. ◇◇

Example 14. Imagine in the preceeding that the ξ_i are the successive outcomes of a "game of chance". The X_n is the "fortune" of the gambler after n plays. We could enhance this by allowing the gambler to choose a wager w_n for play $\#n$, determined in some way from the observed values of $\xi_1, \ldots, \xi_{n-1}$. I.e. w_n is an $\mathcal{F}_{n-1}$ measurable random variable. The gambler wins $w_n\xi_n$ as a result of play $\#n$. Then

$$X_n = \sum_1^n w_i\xi_i$$

is still a martingale:

$$\begin{aligned} E[X_{n+1}|\mathcal{F}_n] &= E[X_n + w_{n+1}\xi_{n+1}|\mathcal{F}_n] \\ &= E[X_n|\mathcal{F}_n] + w_{n+1}E[\xi_n|\mathcal{F}_n] \\ &= X_n. \end{aligned}$$

The martingale property is interpreted as meaning this gambling system constitutes a fair game. ◇◇

Application to Likelihood Ratios. Suppose $Y_1, Y_2, \dots$ is a sequence of random variables on $(\Omega, \mathcal{F})$ and there are two possible probability measures P and Q on Ω. The Y_i are i.i.d with respect to both P and Q, but their actual distribution under each of them is different. We observe the Y_i successively and based on these observations want to decide which of P or Q is the true measure. One way to do this is to look at the likelihood ratios:

$$X_n = \prod_1^n \frac{q(Y_i)}{p(Y_i)}.$$

Here p and q are the densities (with respect to ℓ on $\mathbb{R}$) of the distribution of Y_i with respect to P and Q respectively. (We asssume q, p exist and $p > 0$.) A large value of X_n indicates Q is more likely; small values indicate P.

Lets suppose P is the correct measure and consider X_n as random variables on $(\Omega, \mathcal{F}, P)$. Take $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$. Then the X_n form a martingale. To see this we will check that for each n

$$Q(A) = \int_A X_n \, dP \tag{5}$$

for all $A \in \mathcal{F}_n$. If (5) is true, then since any $A \in \mathcal{F}_n$ is also $A \in \mathcal{F}_{n+1}$ it will follow that

$$\int_A X_n \, dP = Q(A) = \int_A X_{n+1} \, dP.$$

Since X_n is clearly $\mathcal{F}_n$ measurable, this will show that $E[X_{n+1}|\mathcal{F}_n] = X_n$, verifying the martingale property.

Any $A \in \mathcal{F}_n$ can be written as

$$A = \{\omega : \ (Y_1, \dots, Y_n) \in H\},$$

for some $H \in \mathcal{B}(\mathbb{R}^n)$. (See Theorem II.F.) Let μ_n^Q and μ_n^P be the distributions of $(Y_1, \dots Y_n)$ with respect to Q and P respectively, and ℓ Lebesgue measure on $\mathbb{R}^n$.

$$\frac{d\mu_n^Q}{d\ell} = \prod_1^n q(y_i); \quad \frac{d\mu_n^P}{d\ell} = \prod_1^n p(y_i).$$

Now we can verify (21) as follows ($y = (y_1, \dots, y_n)$ here):

$$\begin{aligned}
\int_A X_n \, dP &= \int 1_H(Y_1, \dots Y_n) \prod_1^n \frac{q(Y_i)}{p(Y_i)} \, dP \\
&= \int 1_H(y) \prod_1^n \frac{q(y_i)}{p(y_i)} \, \mu_n^P(d(y_1, \dots, y_n)) \\
&= \int 1_H(y) \prod_1^n q(y_i) \, d\ell \\
&= \int 1_H(y) \, d\mu_n^Q \\
&= \int 1_H(Y_1, \dots, Y_n) \, dQ \\
&= Q(A)
\end{aligned}$$

This application and the examples above illustrate that there are a number of naturally occuring situations which have the structure that we have labeled "martingale". What make recognizing this common structure valuable is that there are a number of theorems that can be proven for martingales in general. Among the most important are convergence theorems, such as the following.

Theorem H. *Suppose X_n is a martingale with respect to $\mathcal{F}_n$ and the $E[|X_n|)$ are bounded. (I.e. there is K with $E[|X_n|] \leq K$ for all n.) Then the X_n converge almost surely as $n \to \infty$. (I.e. $\lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)$ exists almost surely.)*

We can use this in our likelihood ratio application, because (using (21))

$$E[|X_n|] = E[X_n] = \int_\Omega X_n \, dP = Q(\Omega) = 1.$$

Therefore the limit of the likelihood ratios $\lim X_n = X_\infty$ exists a.s. The important issue in considering the likelihood rations as statistical indicators is what this limit actually is. If we assume $p \neq q$, i.e. the distribution of an individual Y_i under P is different than under Q, we will show that

$$X_n = \prod_1^n \frac{q(Y_i)}{p(Y_i)} \to 0 = X_\infty.$$

Fatou's Lemma tells us that

$$\int_A X_\infty \, dP = \int_A \lim X_n \, dP \leq \liminf \int_A X_n \, dP.$$

Now if $\in \mathcal{F}_k$ then for any $n \geq k$, $\int_A X_n \, dP = Q(A)$. Therefore we have

$$\int_A X_\infty \, dP \leq Q(A)$$

for $A \in \cup_1^\infty \mathcal{F}_k$. The Monotome Class Theorem (I.C) extends this to all $A \in \mathcal{F}_\infty = \sigma(\cup_1^\infty \mathcal{F}_k) = \sigma(Y_1, Y_2, \dots)$. This would mean that $X_\infty \leq dP/dQ$ <u>if</u> the latter exists on $\mathcal{F}_\infty$. In fact quite the oppostie is true!

Since $p(\cdot) \neq q(\cdot)$ there is some bounded function $\phi(\cdot)$ for which

$$m_P = E^P[\phi(Y_i)] = \int \phi(y)p(y) \, d\ell \neq \int \phi(y)q(y) \, d\ell = E^Q[\phi(Y_i)] = m_Q.$$

The Strong Law of Large Numbers tells us that

$$\frac{1}{n}\sum_1^n \phi(Y_i(\omega)) \to \begin{cases} m_P \text{ a.s. under } P \\ m_Q \text{ a.s. under } Q \end{cases}.$$

If we define

$$A_P = \{\omega \in \Omega : \; \frac{1}{n}\sum_1^n \phi(Y_i(\omega)) \to m_P\}$$
$$A_Q = \{\omega \in \Omega : \; \frac{1}{n}\sum_1^n \phi(Y_i(\omega)) \to m_Q\},$$

then A_P and A_Q are disjoint, both in $\mathcal{F}_\infty$, and

$$P(A_P) = 1 \text{ while } Q(A_P) = 0$$
$$P(A_Q) = 0 \text{ while } Q(A_Q) = 1.$$

This says that P and Q are about as far from having densities with respect to each other as possible. With regard to X_∞ we find that

$$\int_\Omega X_\infty \, dP = \int_{A_P} X_\infty \, dP \leq Q(A_P) = 0.$$

This implies that $X_\infty = 0$ a.s.

What we see from all this is that

$$X_n \to 0 \text{ almost surely with respect to } P.$$

Interchanging the roles of P and Q above would imply that $1/X_n \to 0$ almost surely with respcet to Q, or

$$X_n \to \infty \text{ almost surely with respect to } Q.$$

This explains (rather convincingly) why the likelihood ratio (for large n) is a indicator of whether P or Q is the true probability measure.

Markov Processes

The defining property of a Markov Process X_n (with respect to $\mathcal{F}_n$) is that for each $A \in \mathcal{B}$, $P[X_{n+1} \in A|\mathcal{F}_n]$ should depend only on X_n. In other words,

$$P[X_{n+1} \in A|\mathcal{F}_n] = P_n(X_n(\omega), A),$$

for some function $P_n(x, A)$ called the *transition probability.* This should be a measure with respect to $A \in \mathcal{B}$ (for each x, n) and a measureable function of x (for each n and $A \in \mathcal{B}$). Often P_n does not depend on n. (In the simple examples we look at the transition probability can be given explicitly, but in many applications it can not be – the process would have to be identified in another way.)

Symmetric Random Walk. Here $X_n \in Z^d$, the "integer lattice" in d dimensions. The transition probability is

$$P(x, \{y\}) = \begin{cases} \frac{1}{2^d} & \text{if } |x-y| = 1 \\ & \text{otherwise} \end{cases}.$$

To complete the specification of the process we need to prescribe X_0; lets say $X_0 = 0$.

An interesting question concerning the random walk is that of its recurrence: does X_n eventually return to 0 (a.s) or is there some positive probability that $X_n \neq 0$ for all $n \geq 1$? To formulate this precisely, define the random variable

$$\eta = \begin{cases} \inf\ \{n \geq 1 :\ X_n = 0\} & \text{if } X_n \text{ does return to } 0 \\ +\infty & \text{if } X_n \text{ never returns.} \end{cases}$$

This is a "time-valued" random variable of the type called a *stopping time*:

$$\{\omega :\ \eta \leq n\} \in \mathcal{F}_n, \quad \text{for each } n.$$

The question then is whether $P(\eta < \infty) = 1$ or < 1. The answer depends on the dimension d:

	$d = 1$	$d = 2$	$d \geq 3$
$P(\eta < \infty)$	$= 1$	$= 1$	< 1
$E[\eta]$	$< \infty$	$= \infty$	$= \infty.$

In fact for $d \geq 3$ $\lim |X_n| = \infty$ a.s.

We will not justify all these assertions, but simply want to point out how the analysis of these recurrence questions involes a key feature of Markov processes, namely that probabilities involving them can often be described in terms of functions on "state spce" and appropriate difference or differential equations. To see this for the recurrence issue, consider the function $\phi(x)$ defined by

$$P[X_m = 0 \text{ some } m \geq n|\mathcal{F}_n] = \phi(X_n).$$

Here $0 \leq \phi(\cdot) \leq 1$ is a function on Z^d which satisfies

1) $\phi(0) = 1$;
2) $\phi(x) = \int \phi(y)\, P(x, dy)$, $x \neq 0$.

In the case of $d = 1$, 2) above translates into a simple difference equation:

$$\phi(x) = \frac{1}{2}\phi(x-1) + \frac{1}{2}\phi(x+1).$$

So we can answer the recurrence question by looking for solutions of this equation with $0 \leq \phi \leq 1$ and $\phi(0) = 1$. The equation is equivalent to

$$\begin{aligned} \phi(x) - \phi(x-1) &= \phi(x+1) - \phi(x) \\ \phi(x) &= \phi(0) + x \cdot m, \quad x \geq 1. \end{aligned}$$

This makes it clear that $\phi(x) \equiv 1$ is the only solution. The same conclusion works out for $d = 2$ though the reasoning is much harder. However in $d \geq 3$ there is a solution with $0 < \phi(x) < 1$ for $x \neq 0$ – this accounts for the non-recurrence in those dimensions.

There are numerous other discrete time examples of Markov processes (general Markov chains, branching processes) but we want to mention continuous time examples as well. Here the Markov property takes the form

$$P[X_t \in A|\mathcal{F}_s] = P(X_s, t-s, A) \quad \text{for} \quad s \leq t.$$

We will mention the example of Brownian motion brieΩy. Some other examples of continuous-time Markov processes are the Poission, contact and Cauchy processes.

Brownian Motion. (Wiener Process)

$$P(x, h, dy)/\ell(dy) = (2\pi h)^{-d/2} e^{-|y-x|^2/2h}.$$

This means that $X_t - X_s$ is Gaussian (normally) distrbuted with mean 0, variance= $\sqrt{t-s}$ (in $d = 1$). (In $d > 1$, the covariance matrix is $\sqrt{t-s} \cdot I$.) Given $t_0 < t_1 < \cdots < t_n$, it follows that

$$X_{t_0}, X_{t_1} - X_{t_0}, \ldots X_{t_n} - X_{t_{n-1}}$$

are independent with the Gaussian distributions indicated.

The remarkable fact (proven by N. Wiener in 1923) is that the $X_t(\omega)$ can be constructed so that, for each $\omega \in \Omega$, $X_t(\omega)$ is continuous in t. The X_t form a random continuous function. Here, brieΩy, are some properties:

$$P(\frac{d}{dt}X_t \text{ exists for some } t > 0) = 0$$

For any $f(x)$ with f' and f'' bounded,

$$M_t = f(X_t) - \int_0^t \frac{1}{2} f''(X_s)\, ds \quad \text{is a martingale.}$$

This is for $d = 1$. For $d > 1$ the f'' in the above integral is replaced wit;h the *Laplace operator*:

$$\Delta f = \sum_1^d \frac{\partial^2}{\partial x_i^2} f.$$

The expression Δf is central to a number of partial differential equations important in the sciences. As a result, Brownian motion can be used to “solve” such equations by taking expected values of appropriate expressions involving integrals of X_t.

Problem **1** ....

Suppose f is a nonnegative function on a σ-finite measure space $(\Omega, \mathcal{F}, P)$. Let

$$A = \{(\omega, y) : \; 0 \leq y \leq f(\omega)\}$$

be the region “under the graph” of $y = f(\omega)$. Show that the integral gives the area under the curve in the sense that

$$\int_\Omega f\, d\mu = \mu \times \ell(A).$$

Problem **2** ....

Prove for any probability distribution function F and $c \in \mathbb{R}$ that

$$\int_{\mathbb{R}} [F(x+c) - F(x)]\; \ell(dx) = c.$$

Problem **3** ....

If $X \geq 0$ is a random variable, show that for any positive integer n

$$E[X^n] = \int_{[0,\infty)} nx^{n-1} P[X > x] \, d\ell.$$

Problem **4** ....

Let $\Omega = [0, \infty) \times [0, 2\pi)$, with the typical point being $\omega = (r, \theta)$. On Ω consider the measure μ having density r with respect to Lebesgue measure. Let $\Phi : \Omega \to \mathbb{R}^2$ be given by $\Phi(r, \theta) = (r \cos\theta, r \sin\theta)$. Let $\mathcal{R}$ be the class of subsets of Ω of the form $[r_1, r_2) \times [\theta_1, \theta_2)$. $\sigma(\mathcal{R}) = \mathcal{B}(\Omega)$ are the Borel sets in Ω. Let

$$\mathcal{P} = \{S \subseteq \mathbb{R}^2 : \ \Phi^{-1}S \in \mathcal{R}\}.$$

a) Show that $\mathcal{P}$ is a π-system and that $\sigma(\mathcal{P}) = \mathcal{B}(\mathbb{R}^2)$.
b) If $\Phi^{-1}S = [r_1, r_2) \times [\theta_1, \theta_2)$, what should $\ell(S)$ be? (Just tell me what the correct formula is; don't actually verify it by careful calculation.)
c) Using a) and b), prove that $\ell = \mu\Phi^{-1}$.
d) Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is (Borel) measurable. Show that

$$\int_{\mathbb{R}^2} f(x, y) \, \ell(d(x, y)) = \int_{[0,2\pi)} \int_{[0,\infty)} f(r \cos\theta, r \sin\theta) r \, \ell(dr)\ell(d\theta)$$

if $f \geq 0$. Show that f is ℓ-integrable on $\mathbb{R}^2$ if and only if $r \cdot f \circ \Phi$ is μ-integrable on Ω, in which case the above formula also holds. (See III.F and G, and Example III.6)

Problem **5** ....

Show that if X and Y are independent random variables, both with the standard normal distribution, then X/Y has the Cauchy distribution with parameter $u = 1$.

Problem **6** ....

Let μ, ν and λ be σ-finite measures on the common space $(\Omega, \mathcal{F})$.

a) Show that $\nu \ll \mu$ and $\mu \ll \lambda$ imply that $\nu \ll \lambda$ and

$$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu} \cdot \frac{d\mu}{d\lambda}.$$

b) Suppose $\mu \ll \lambda$ and $\nu \ll \lambda$. Let A be the set where $d\nu/d\lambda > 0 = d\mu/d\lambda$. Show that $\nu \ll \mu$ if and only if $\lambda(A) = 0$, in which case

$$\frac{d\nu}{d\mu} = \begin{cases} \frac{d\nu/d\lambda}{d\mu/d\lambda} & \text{if } d\mu/d\lambda > 0 \\ 0 & \text{if } d\mu/d\lambda = 0 \end{cases}.$$

Problem **7** ....

Prove Bayes' Theorem in the form

$$P[G|A] = \frac{\int_G P[A|\mathcal{G}] \, dP}{\int_\Omega P[A|\mathcal{G}] \, dP},$$

for $G \in \mathcal{G}$.

Problem **8** ....

Suppose X is a random variable on $(\Omega, \mathcal{F}, P)$ with finite second moment and $\mathcal{G} \subseteq \mathcal{F}$ is a sub-σ-field. Show that for any $\mathcal{G}$-measurable g

$$E[(X - g)^2] = E[(X - E[X|\mathcal{G}])^2] + E[(E[X|\mathcal{G}] - g)^2].$$

In particular, $g = E[X|\mathcal{G}]$ minimizes $E[(X - g)^2]$ over all such g.

Problem **9** . . . .

Suppose P and Q are two probability measures on $(\Omega, \mathcal{F})$, with $Q \ll P$ and $f = dQ/dP$. If X is a random variable having $E^Q[|X|] < \infty$, show that for any sub-σ-field $\mathcal{G} \subseteq \mathcal{F}$,

$$E^Q[X|\mathcal{G}] = 1_{A^c} \frac{E^P[Xf|\mathcal{G}]}{E^P[f|\mathcal{G}]} \quad \text{a.s. w.r.t. } Q,$$

where $A = \{E^P[f|\mathcal{G}] = 0\}$.

Problem **10** . . . .

Show that the independence of X and Y implies that

$$E[Y|X] = E[Y] \tag{6}$$

and that (6) implies

$$E[XY] = E[X] \cdot E[Y]. \tag{7}$$

Find simple examples to show that both reverse implications are false.

Unit S . . . . **Mathematical Supplements**

We summarize here a number of mathematical details and facts which are used in our discussion. This is only a quick summary, not a thorough treatment. Please talk to me if you want more on these or other background topics.

Elements, Sets and Classes

Our discussions with sets involve three distinct types of objects:

- elements (denoted by lower case letters like ω, a or x);
- sets (denoted by upper case letters like Ω, A or X);
- classes (denoted using script letters such as $\mathcal{A}$, $\mathcal{F}$, or $\mathcal{B}$).

Elements are the most basic objects. A set is a collection of elements. Classes are collections of sets.

The statement $x \in A$ means that x is one of the elements which belongs to the set of elements called A. A subset $B \subseteq A$ is another set with the property that every element of B is also an element in A, in other words

$$x \in A \text{ whenever } x \in B.$$

The *empty set* is the set containing no elements at all: $\emptyset = \{\ \}$. It is considered to be a subset of every set: $\emptyset \subseteq A$, regardless of what the set A is. We typically use Ω for the total collection of all elements under consideration, the *master set*. All sets are then subsets of Ω.

The operations of intersection, union and difference of sets should be familiar:

$$A \cap B = \{x : \ x \in A \text{ and } x \in B\}$$

$$A \cup B = \{x : \ x \in A \text{ or } x \in B\}.$$

$$A \setminus B = \{x : \ x \in A \text{ but } x \notin B\}.$$

If we have a sequence A_1, A_2, . . . , A_n, . . . of sets, we write their intersection and union as

$$\cap_{n=1}^{\infty} A_n = \cap A_n = \{x : \ x \in A_n \text{ for every } n = 1, 2 \ldots\},$$

$$\cup_{n=1}^{\infty} A_n = \cup A_n \{x : \ x \in A_n \text{ for some } n = 1, 2 \ldots\}.$$

Set compliments only make sense with reference to the master set Ω:

$$A^c = \{\omega \in \Omega : \ \omega \notin A\} = \Omega \setminus A.$$

We can write $A \setminus B = A \cap B^c$ provided A and B are subsets of the same master set.

A *class* is a collection of subsets of the master set, Ω. The σ-fields and π-systems of our discussion are important examples of classes. It is tempting to talk about classes as "sets of sets" but this would be using the word "set" in two different ways and leads to some logical paradoxes. We insist therefore on reserving the word "set" for collections of elements, and "class" for collection of sets. It is important to keep the three types of objects (element, set and class) distinct in our thinking.

Example 1. Let the master set consist of all real numbers: $\Omega = \mathbb{R}$ Thus elements are individual real numbers x. Some examples of sets are

$$A = \{x \in \mathbb{R} : \ x > 0\} = (0, \infty) \quad \text{and} \qquad I = [-1, 2] = \{x \in \mathbb{R} : \ -1 \leq s \leq 2\}.$$

A is called the set of positive real numbers. I is an example of what we call a (bounded) closed interval. $I \subseteq \mathbb{R}$ but neither $I \subseteq A$ nor $A \subseteq I$. It should be obvious that

$$A \cap I = (0, 2] \quad A \setminus I = (2, +\infty) \quad I \setminus A = [-1, 0].$$

◇◇

A set such as $B = \{-3\}$ containing exactly one element is called a *singleton* set. Note that -3 and $\{-3\}$ are different mathematical objects; the first is an element , the second is a set (which in this case contains exactly one element). Again, do not mix the concepts of element and set.

Now lets look at some classes of subsets of $\Omega = \mathbb{R}$ Let $\mathcal{C}$ be the collection of all (bounded) closed intervals $[a, b]$ with $a \leq b$. In other words to say $J \in \mathcal{C}$ means that J is a subset of $\mathbb{R}$ of the particular form $J = \{x \in \mathbb{R} :\ a \leq x \leq b\}$ for some $a \leq b$. With A, B and I as already defined, $A \notin \mathcal{C}$ while $B \in \mathcal{C}$ and $I \in \mathcal{C}$. (Note that $B = [a, b]$ using $a = b = -3$, which falls within the scope of sets allowed in $\mathcal{C}$.)

Do not confuse the statement that $B = \{-3\} \in \mathcal{C}$, which is true, with $-3 \in \mathcal{C}$, which is entirely incorrect! Anything in $\mathcal{C}$ must be a subset of Ω, not an element. Classes contain sets; sets contain elements. It would be legitimate to ask whether $\Omega \in \mathcal{C}$, since Ω is a set and hence could conceivably be in the class $\mathcal{C}$. The answer however is no since $\Omega = \mathbb{R}$ cannot be written as a bounded closed interval $[a, b]$. Our requirement $a \leq b$ implies that $\emptyset \notin \mathcal{C}$.

Consider also the class $\mathcal{U}$ of all singleton sets. Then we would say $\mathcal{U} \subseteq \mathcal{C}$, because every singleton set is a closed interval: $\{x\} = [x, x]$. But $\mathcal{C} \subseteq \mathcal{U}$ is false.

The collection of all $A \subseteq \mathbb{R}$ with $\{0, 1\} \subseteq A$ forms another class; lets call it $\mathcal{P}$. We can intersect classes; $\mathcal{P} \cap \mathcal{C}$ would consist of all intervals $J = [a, b]$ which contain both elements 0 and 1, in other words $J = [a, b]$ with $a \leq 0$ and $1 \leq b$. Likewise we can form unions of classes, such as $\mathcal{U} \cup \mathcal{P}$. Note that $\mathcal{U} \cap \mathcal{C} = \mathcal{U}$, while $\mathcal{U} \cap \mathcal{P}$ contains no sets at all (not even $\emptyset$). ◇◇

Notice that we use the same symbols ($\in$, $\cap$, $\cup$, $\subseteq$) to discuss classes and the sets which they contain as we do to discuss sets and the elements they contain. If $\mathcal{A}_n$ is a class for each n then we can form new classes by forming the intersection or union of all of them:

$$\cap \mathcal{A}_n = \{F \subseteq \Omega :\ F \in \mathcal{A}_n \text{ for every } n\},$$

$$\cup \mathcal{A}_n = \{F \subseteq \Omega :\ F \in \mathcal{A}_n \text{ for some } n\}.$$

Product Sets. If X and Y are any two sets we can form a new set $X \times Y$ from them, called their *cartesian product*. $X \times Y$ is the set of all ordered pairs (x, y) where $x \in X$ and $y \in Y$. X^2 is shorthand for $X \times X$. Thus $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ is the familiar x, y plane from calculus. Likewise we can form the product of more than two sets:

$$X \times Y \times Z = \{(x, y, z) :\ x \in X, y \in Y, z \in Z\};$$

$$X^n = \{(x_1, \ldots, x_n) :\ x_i \in X \text{ for each } i = 1, \ldots, n\}.$$

If $A \subseteq X$ and $B \subseteq Y$ then $A \times B \subseteq X \times Y$, but $A \not\subseteq X \times Y$.

DeMorgan's Laws and Logical Negations

The compliment of an intersection (union) is the union (intersection) of the compliments:

$$(A \cap B)^c = (A^c) \cup (B^c), \quad (\cap A_n)^c = \cup A_n^c;$$

$$(A \cup B)^c = (A^c) \cap (B^c), \quad (\cup A_n)^c = \cap A_n^c.$$

These are called DeMorgan's laws in set theory. They are essentially the same as the rules for negating statements involving the logical quantifiers "for every" and "for some". For instance consider the statement $x \in \cap A_n$; we can write it using logical quantifiers as

$$\text{for every } n,\ x \in A_n. \tag{1}$$

The negation of this is the statement that $x \in (\cap A_n)^c$. DeMorgan's law says that $(\cap A_n)^c = \cup A_n^c$, so that the negated statement can be expressed as

(2) $$\text{for some } n,\ x \notin A_n.$$

Observe that the "for every" in (1) changed to "for some" in (2) and the phrase "$x \in A_n$" changed to its negation, "$x \notin A_n$". Complicated logical expressions are negated by reversing the quantifiers from the outside in, negating the inner statements as you proceed.

Example 2. Consider the statement that $\lim_{x\to a} f(x)$ exists. Using the definition of limit this can be expressed using quantifiers as

(3) There exists ℓ so that for every $\epsilon > 0$ there is some $\delta > 0$ so that for every x with $|x - a| < \delta$ the inequality
$$|f(x) - \ell| < \epsilon \text{ holds.}$$

The negation of this, i.e. the statement that $\lim_{x\to a} f(x)$ does not exist, becomes

(4) For every ℓ there is some $\epsilon > 0$ so that for every $\delta > 0$ there is some x with $|x - a| < \delta$ for which the inequality
$$|f(x) - \ell| < \epsilon \text{ fails.}$$

Let $L = \{f(\cdot) :\ \lim_{x\to a} f(x) \text{ exists }\}$. The statement (3) says

$$L = \cup_\ell \cap_{\epsilon>0} \cup_{\delta>0} \cap_{x\in B_\delta} A_{\epsilon,x},$$

where

$$B_\delta = \{x :\ |x - a| < \delta\} \quad \text{and} \quad A_{\epsilon,x} = \{f(\cdot) :\ |f(x) - \ell| < \epsilon\}.$$

Now (4) is the expression for L^c obtained by DeMorgan's laws:

$$L^c = [\cup_\ell \cap_{\epsilon>0} \cup_{\delta>0} \cap_{x\in B_\delta} A_{\epsilon,x}]^c = \cap_\ell \cup_{\epsilon>0} \cap_{\delta>0} \cup_{x\in B_\delta} A_{\epsilon,x}^c.$$

◇◇

Countability

Sets which contain an infinite number of distinct elements are called infinite sets, naturally. However some infinite sets must be considered as "bigger" than others. The "smallest" kind are those we call countable. An infinite set A is called *countably infinite* if its elements can be listed (in their entirety) as a sequence,

$$A = \{a_1,\ a_2,\ \ldots,\ a_n,\ \ldots\} = \{a_n :\ n = 1, 2, \ldots\}.$$

Another way to say this is that there is a way to put the elements of A in one-to-one correspondence with the positive integers $\mathbb{N} = \{1,\ 2, \ldots\}$; $\mathbb{N}$ and A have the same "number of elements". This defines what it means for an infinite set to be countable; every finite set is also considered countable. Thus a countable set is one which is either finite or countably infinite. Here are some properties and examples.

- If A and B are each countable then $A \cup B$ is also countable. In fact a countable union of countable sets is countable: $\cup_{n=1}^\infty A_n$ is countable if each A_n is countable.
- If A is countable and $B \subseteq A$ then B is also countable.
- Every infinite set has a countably infinite subset.
- The set of all integers $\{\cdots - 3, -2, -2, 0, 1, 2, 3, \ldots\}$ is countable.
- The rational numbers $Q = \{x \in \mathbb{R} :\ x = n/m \text{ where } n, m \text{ are integers}\}$ is countable.
- The reals numbers $\mathbb{R}$ is <u>un</u>countable.
- The set Ω of all infinite sequences ω of 0's and 1's, such as
$$\omega = 0101101111000101001010101101101110001010001110100001\ldots$$
is <u>un</u>countable.

Infimums and Supremums in $\mathbb{R}$

If $A \subseteq \mathbb{R}$ is a nonempty, finite set then it has a maximum element and a minimum element: $a = \min A$ is that $a \in A$ with the property that

$$a \leq x \text{ for all } x \in A; \tag{5}$$

similarly for $\max A$. However infinite subsets of $\mathbb{R}$ may fail to have maxima and/or minima. For instance $A = (0, 1]$ has a maximum ($\max A = 1$) but no minimum! We can't call 0 the minimum element because 0 is not an element of A. But otherwise 0 is what we would identify as the "bottom value" of A. The proper statement is that 0 is the *infimum* of A: $0 = \inf A$. The infimum of a set is not required to be one of its elements. The definition of $\alpha = \inf A$ is that (5) holds for $a = \alpha$ but for no value of $a > \alpha$; i.e. $\alpha = \inf A$ is the greatest lower bound for A. Similarly the *supremum* $\sup A$ is defined to be the smallest upper bound for A. When A has a minimum element then $\min A = \inf A$, but the infimum makes sense even when the idea of a minimum element does not. In fact it is a fundamental property of $\mathbb{R}$ that $\inf A$ does exist for any set $A \subseteq \mathbb{R}$ which is nonempty and bounded below (i.e. (5) holds for some $a \in \mathbb{R}$) . Likewise every nonempty set which is bounded above has a supremum.

The infimum and supremum have the following monotonicity property, as you can convince yourself: if $A \subseteq B$ then

$$\sup A \leq \sup B \quad \text{and} \quad \inf A \geq \inf B.$$

Example 3. Let $A = \{1/n : n = 1, 2, \dots\}$. A is bounded (both above and below) and is nonempty so $\inf A$ and $\sup A$ are guaranteed to exist. $\sup A = 1 = \max A$. $\inf A = 0$, but $\min A$ does not exist. ◇◇

Extended Real Numbers

Sometimes it is convenient to add two additional elements, $\pm\infty$, to $\mathbb{R}$. This "enlarged" version of the real numbers is called the *extended real numbers*, denoted by either $\mathbb{R}_\infty$ or $[-\infty, +\infty]$. Most rules of arithmetic and inequalities extend to the new elements $\pm\infty$ in a natural way. For instance $x/\pm\infty = 0$ for all finite x. If $x > 0$ then $x \cdot \pm\infty = \pm\infty$ (the signs would reverse if $x < 0$). The convention

$$\pm\infty \cdot 0 = 0$$

may not seem obvious, but it is the right thing for our discussion of measure theory. We do however leave $\infty - \infty$ and ∞/∞ undefined.

The inequalities

$$-\infty \leq x \leq +\infty$$

hold for all $x \in \mathbb{R}_\infty$. This is obvious, but it has the consequence that every set $A \subseteq \mathbb{R}_\infty$ has both an infimum and supremum in $\mathbb{R}_\infty$. For instance $\sup \mathbb{N} = +\infty$, even though $\mathbb{N}$ is not bounded in the usual sense. (In fact $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$. However $A = \emptyset$ is the only set for which $\sup A < \inf A$!)

Limits Superior and Inferior

You are familiar with the notion of the limit of a sequence of real numbers $\{a_n\}$, expressed by

$$\lim_{n\to\infty} a_n = \ell \quad \text{or just} \quad \lim a_n = \ell.$$

Of course not all sequences have limits; $a_n = (-1)^n$ diverges for instance. The more general notions of limit superior ($\limsup a_n$) and limit inferior ($\liminf a_n$) are useful because they often exist even when the conventional limit does not. For instance, if we are discussing a sequence $\{a_n\}$, but don't know yet whether or not $\lim a_n$ exists, we can still talk about $\liminf a_n$ and $\limsup a_n$.

The formal definitions are

$$\liminf a_n = \lim_{k\to\infty} (\inf\{a_n : \ n \geq k\}); \quad \limsup a_n = \lim_{k\to\infty} (\sup\{a_n : \ n \geq k\}).$$

The idea is that for any value $b < \liminf a_n$ there will be only a finite number of the a_n below b, while for any value $b > \liminf a_n$ there will be infinitely many a_n with $a_n < b$. An equivalent way to say it is that $\ell = \liminf a_n$ is the smallest value to which a subsequence $\{a_{n_k}\}$ can converge, $\ell = \lim_k a_{n_k}$. Likewise $\limsup a_n$ is the largest value to which any subsequence can converge.

Example 4. $\liminf(-1)^n = -1$ and $\limsup(-1)^n = +1$. ◇◇

We still need to assume something about the sequence $\{a_n\}$ to insure that $\liminf a_n$ and $\limsup a_n$ exist as real numbers. For instance if $\{a_n\}$ is bounded they will both exist. However if we accept $\pm\infty$ as legitimate values (i.e. work in $\mathbb{R}_\infty$) then $\liminf a_n$ and $\limsup a_n$ will always exist, regardless of the sequence $\{a_n\}$. Some general properties are as follows.

- $\liminf a_n \leq \limsup a_n$ for any sequence $\{a_n\}$.
- If $a_n \leq b_n$ with only a finite number of exceptions, then $\liminf a_n \leq \liminf b_n$ and $\limsup a_n \leq \limsup b_n$.
- $\limsup -a_n = -\liminf a_n$.
- $\lim a_n = \ell$ if and only if $\liminf a_n = \ell = \limsup a_n$. (This holds in particular for $\ell = \pm\infty$.)

Note that the first inequality means that simply showing $\limsup a_n \leq \liminf a_n$ is enough to imply that $\lim a_n$ exists!

Convergence and Topology in $\mathbb{R}^d$

We are used to calling an interval $J \subseteq \mathbb{R}$ open if it does not contain its endpoints; $J = (a, b)$, or $(-\infty, b)$ for instance. In general a set $A \subseteq \mathbb{R}$ is called *open* if for every $a \in A$ there is some $\epsilon > 0$ so that $(a-\epsilon, a+\epsilon) \subseteq A$. A is called *closed* if its compliment A^c is open. While it is true that every open set is a union of a countable number of open intervals, $A = \cup(a_n, b_n)$, it is not true that every closed set is a countable union of closed intervals!

In $\mathbb{R}^d$ the role of an interval $(a-\epsilon, a+\epsilon)$ is taken over by the ball of radius ϵ centered at a:

$$B_\epsilon(a) = \{x \in \mathbb{R}^d : \ d(x, a) < \epsilon\},$$

where $d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}$ is the usual Euclidean distance between two points. Thus $A \subseteq \mathbb{R}^d$ is called *open* if for every $a \in A$, there is some ball centered at a entirely contained in A: $B_\epsilon(a) \subseteq A$ for some $\epsilon > 0$. A is *closed* if A^c is open. Of course many sets are neither closed nor open.

We should also mention compact sets. $K \subseteq \mathbb{R}^d$ is compact if it is both closed and bounded. (This is not the definition but is equivalent to it by the Heine-Borel Theorem).

A sequence $\{a_n\}$ in $\mathbb{R}^d$ converges to b when each of its coordinate sequences converges to the respective coordinate of a. In other words if

$$a_n = (a_{n,1}, \ldots, a_{n,d}) \quad b = (b_1, \ldots, b_d)$$

then $\lim a_n = b$ means

$$\lim_{n\to\infty} a_{n,i} = b_i \quad \text{for each} \quad i = 1, \ldots, d.$$

Compact sets K have the property that any sequence of points in it, $a_n \in K$, will have a convergent subsequence $\{a_{n_m}\}$, with $\lim_m a_{n_m}$ also in K.

Finally we mention continuous functions. Several different descriptions can be given of what it means for $f : \mathbb{R}^d \to \mathbb{R}^r$ to be continuous. The two most useful for us are

- whenever $A \subseteq \mathbb{R}^r$ is open, then $f^{-1}A = \{x \in \mathbb{R}^d : \ f(x) \in A\}$ is also open;

or

- whenever $\{x_n\}$ is a convergent sequence in $\mathbb{R}^d$ then $\{f(x_n)\}$ converges in $\mathbb{R}^r$ to the value

$$\lim_n f(x_n) = f(\lim_n x_n).$$