MATH 230A/STATS 310A NOTES

ARUN DEBRAY
OCTOBER 1, 2014

These notes were taken in Stanford's Math 230A class (crosslisted as Stats 310A) in Fall 2014, taught by Persei Diaconis. I live-TEXed them using `vim`, and as such there may be typos; please send questions, comments, complaints, and corrections to adebray@stanford.edu.

CONTENTS

1. INTRODUCTION AND EXAMPLES OF PROBABILITY SPACES: 9/22/14

"I like inequalities. I hope you learn to like them too."

In this class, we're going to develop measure theory; probability is one of the subjects in which it comes alive. So there will be proofs and constructions of many standard measure-theory ideas, but the explicit goal is for probability.

The course will cover roughly the first twenty sections of the book; there will be weekly homework of about six problems, due in class on Mondays. However, about 25% of the class material is not in the textbook, so come to class and take notes!

Before discussing the structure of the course, let's do some probability. The nice thing about probability is that it's very easy to say what it's about. In probability, there is a (for today, finite or countable) set Ω, weights $P(\omega) \geq 0$ such that

$$\sum_{\omega \in \Omega} P(\omega) = 1,$$

and a subset $A \subseteq \Omega$. These things are all told to us; the question of probability is to compute or approximate the probability

$$P(A) = \sum_{\omega \in A} P(\omega).$$

We'll refine this, develop it, etc., but it looks really trivial from this viewpoint, doesn't it?

Example 1.1. The Birthday Problem is a well-known problem in probability. It asks, "how many people do we need in a group to have a 50% or greater chance that two or more have the same birthday?" The answer is about 23, which is viscerally surprising. Even in this class, it's likely that two people share a birthday.

There are many variations; this is one of the buildings blocks of combinatorial probability. For example, you could do the same thing with the second hands on a watch (which, for the sake of the argument, are i.i.d.); this would be known as a *birthday problem with sixty categories*.

For a birthday problem with c categories and n people, Ω is the set of n-tuples of numbers from 1 to c: $\Omega = \{(\omega_1, \dots, \omega_n) : \omega_i \in \{1, \dots, c\}\}$. Then, $P(\omega) = 1/c^n$, and for A, we'll calculate the probability of failure (since this is easier to deal with): $A = \{\omega : \omega_i \neq \omega_j \text{ for all } i, j\}$.

Then,

$$P(A) = \sum_{\omega \in A} P(\omega) = \frac{|A|}{c^n}.$$

Awesome. So what's the size of A? The first person has c choices, then the next person needs to choose something different, so there are $c-1$ choices, then $c-2$ (different from the first two), and so on, so $|A| = c(c-1)(c-2)\cdots(c=n+1)$. Thus,

$$P(A) = \frac{c(c-1)\cdots(c-n+1)}{c^n},$$

so the proabbility that two people share a birthday (the complement of A) is

$$\left(1-\frac{1}{c}\right)\left(1-\frac{2}{c}\right)\cdots\left(1-\frac{n-1}{c}\right).$$

This is a fine answer; it's 2014, so we can plug it into a computer and have fun. But it's not at all understandable to people. Recall that asymptotically, $\log(1-x)\sim -x$ (since it's the first term in the power series).[1] So rewrite

$$\begin{aligned} P(A) &= e^{\sum_{i=1}^{n}\log(1-1/c)} \\ &\sim e^{-\sum_{i=1}^{n} i/c} \\ &= e^{-\binom{n}{2}/c}. \end{aligned}$$

Thus, set this equal to 1/2, so

$$\frac{\binom{n}{2}}{c} = \log 2 \doteq 0.69$$

(in this class, all logs will be base e), and therefore $n^2/2c = 0.69$, or $n = 1.2\sqrt{c}$.

This is human-understandable: n should be about 1.2 times the square root of c. For example, if $c = 365$, then $1.2\sqrt{365} = 22.9$, so the goal is 23 students. If we wanted it with 95% probability, one can calculate that $n = 2.5\sqrt{c}$.

In probability, these kinds of heuristics are commonplace: reasoning intuitively, but then backing it up with mathematics.

Let's refine this approximation: $\log(1-x) = -x + O(x^2)$, since $-x - x^2 \le \log(1-x) \le -x$ when $0 \le x \le 1/2$. Thus,

$$\begin{aligned} P(A) &= e^{\sum_{i=1}^{n}\log(1-1/c)} \\ &= e^{-\sum_i i/c + O((1/c)^2)} \\ &= e^{-\binom{n}{2}/c + O(n^3/c^2)}. \end{aligned}$$

This is a rigorous mathematical result:

Theorem 1.2. *If $n, c \to \infty$ such that $n^3/c^2 \to 0$ and $\binom{n}{2}/c \to 1$, then $P(A) \to e^{-1}$.*

Exercise 1. Extend this to three people: how many people are needed in a room so there are even odds that three of them share a birthday?

This is one of a million versions of the birthday problem. For example, the birthday distribution is nonuniform (for example, people are more likely to be born on weekdays, because one birth in five is induced, and doctors don't like working on weekends). By the end of the course, we'll have tools (e.g. Poisson approximation and Stein's process) to analyze this where $P(\omega)$ is nonuniform.

Example 1.3. Put N points at random in the unit square $[0,1]^2$, and an ε-ball around each one. Then, what are the odds that the square is covered? What's the size of the largest disc covered? And so on. This is an example of a continuous distribution; it is common, for example, to put probabilies on $C([0,1])$, the space of continuous functions n the interval. One may even put probability distributions on manifolds, such as

$$\Omega = \left\{x_1, \dots, x_{35} \in \mathbb{R}_+^{35} : \sum x_n = s, \prod x_n = p\right\},$$

which is a 33-dimensional manifold that appears in the study of the gamma distribution; or on the set of probability measures $M_1(\mathbb{R})$ with certain properties.

Let's take a halfway house, $\Omega = [0,1]$, and define probabilities on a certain class of sets: left-open, right-closed sets $(a_i, b_i]$. Then, if $A = \bigcup_{i=1}^{n} I_n$, where the I_n are disjoint, then define

$$P(A) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (b_i - a_i).$$

This can be used as a model for lots of different things, e.g. fair coin tossing. Right now, we're only considering finite sums of disjoint sets. For fair coin tossing, write each number in $[0,1]$ as its nonterminal binary expansion, $\omega = 0.\omega_1\omega_2\omega_3\dots$, e.g. $1/2 = 0.011111\dots$. Let $d_n : [0,1] \to \{0,1\}$ denote the function which retrieves the n^{th} binary digit from a number, so that $\omega = \sum_{n=1}^{\infty} d_n(\omega)/2^n$.

Each d_i defines intervals where $d_i(\omega) = 1$; the set $A = \{\omega : d_i(\omega) = 1\}$ has probability 1/2 for all i, and similarly $P(d_1 = d_2 = 1) = 1/4$, and so forth. Then (as will be explicated on the homework), $P(d_1 = \varepsilon_1, \dots, d_n = \varepsilon_n) = 1/2^n$.

This is a nice model for flipping a coin, and is even the standard model, but... coins have physics, air resistance, coefficients of restitution, and so on. The difference is that it's a model, so we should calculate with it and then compare to the real world. But it does seem to model coins very well.[2]

[1] The notation $f(n) \sim g(n)$, said "$f(n)$ is asymptotic to $g(n)$," means that $\lim_{n\to\infty} f(n)/g(n) = 1$.

[2] "I am possibly the world's most accomplished expert on flipping coins... don't miss a chance to ask me about that."

Theorem 1.4 (Bernoulli's Law of Large Numbers (1713))**.** *For this model of coin tossing and for all $\varepsilon > 0$, the probability*

$$P\left(\left|\frac{d_1 + \cdots + d_n}{n} - \frac{1}{2}\right| > \varepsilon\right) \to 0$$

as $n \to \infty$.

This implies that the proportion of heads that happens as one flips more and more coins should be fair, tending to 1/2.

Proof of Theorem 1.4. To prove the theorem, we know that

$$A = \left\{\omega : \left|\frac{1}{n}\sum d_i(\omega) - \frac{1}{2}\right| > \varepsilon\right\}$$

has a probability; we need to find it.

Define

$$r_i(\omega) = 2d_i(\omega) - 1 = \begin{cases} 1, & d_i = 1 \\ -1, & d_i = 0. \end{cases}$$

This is a simple renormalization, but makes life a bit easier. In particular, it's enough to prove that for all $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum r_i\right| > \varepsilon\right) \to 0.$$

We can use the Riemann integral, and in particular know that

$$\int_0^1 r_i(\omega)\,\mathrm{d}\omega = 0,$$

and the different r_i are orthogonal:

$$\int_0^1 r_i(\omega) r_j(\omega)\,\mathrm{d}\omega = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

Thus, we can find the first and second moments of the sum.

$$\int_0^1 \left(\sum r_i\right)\mathrm{d}\omega = \int_0^1 \left(\sum r_i\right)^2 = n.$$

Thus,

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum r_i\right| > 2\varepsilon\right) &= P\left(\left(\sum r_i\right)^2 > ut^2n^2\right) \\ &\leq \frac{1}{4t^2n^2}\int \left(\sum r_i\right)^2 = \frac{1}{ut^2n}, \end{aligned} \tag{1}$$

which goes to 0 as $n \to \infty$. □

(1) coms from *Markov's inequality*:

Lemma 1.5 (Markov's inequality)**.** *If $f : [0,1] \to \mathbb{R}$ is a nonnegative step function, then for all $a > 0$,*

$$P(\omega : f(\omega) \geq a) \leq \frac{1}{a}\int_0^1 f(\omega)\,\mathrm{d}\omega.$$

Proof. Let $A = \{\omega : f(\omega) \geq a\}$; then,

$$\int_0^1 f(\omega)\,\mathrm{d}\omega \geq \int_A f(\omega)\,\mathrm{d}\omega \geq a \cdot P(A).$$

□

We'll also want another theorem. We want to say that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{n=1}^{\infty} r_n(\omega) = 0,$$

but this isn't true: for example, if $\omega = 0.111111\ldots$, then the limit is 1, and if $\omega = 0.011000011111111\ldots$ (each time, twice as many as before), the limit simply doesn't exist. So we need to make this theorem a little more precise.

Definition. Define a set $A \subset \Omega$ to be *negligible* if for any $\varepsilon > 0$, A can be covered by countably many intervals with total length less than ε.

For example, any one-point set is negligible, as is $\mathbb{Q}$.

Now, we can state the theorem.

Theorem 1.6 (Strong Law of Large Numbers (Borel))**.** *Except on a negligible $N \subseteq \Omega$, for all $\omega \in N^c$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} r_i(\omega) = 0.$$

Remark. Look at the set $A = N^c$, the set of places where this statement is true. It can be characterized as

$$A = \bigcap_{h=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \left\{ \omega : \left| \frac{1}{n} \sum_{i=1}^{n} r_i(\omega) \right| < \frac{1}{h} \right\}.$$

This is very complicated, but we'll develop general methods for dealing with these sets.

This class will intersperse measure theory and probability, since the latter is more interesting and the former is very useful.

2. The Strong Law of Coin Tossing: 9/24/14

"The chalk knows how to do this. I need to stop thinking."

Recall the setup we had yesterday: that $\Omega = [0,1]$, an interval I is $I = (a,b]$, with $P(I) = b - a$, and the base-2 expansion of an $\omega \in \Omega$ is $\omega = \sum_{n=1}^{\infty} d_n(\omega)/2^n$. Then, rescale these to the funtions $r_n(\omega) = 2d_n(\omega) - 1$. Finally, we define

$$S_n(\omega) = \sum_{n=1}^{\infty} r_n(\omega).$$

Theorem 2.1.

$$\lim_{n\to\infty} \frac{s_n(\omega)}{n} = 0,$$

except on a negligible set.

Recall that a negligible set us a set that can be covered by a countable collection of intervals with total measure less than ε for any $\varepsilon > 0$.

Proof. Let $B = \{\omega : \lim_n s_{n^2}/n^2 = 0\}$, since $s_n/n \to 0$ iff $s_{n^2}/n^2 \to 0$; we want to show that B^c is negligible.

Choose a sequence d_n approaching 0 from above; then,

$$\left\{ \omega : \frac{|s_{n^2}|}{n^2} < d_n \text{ for large } n \right\} \subseteq B.$$

Thu, B^C is the set of things such that $|s_{n^2}|/n^2 \geq d_n$ infinitely often, and thus is contained in $\bigcup_{n=1}^{\infty} B_n$, where

$$B_n = \left\{ \omega : \frac{|s_{n^2}|}{n^2} < d_n \right\}.$$

Since B_n is a disjoint union of integers, then by Markov's inequality, $P(B_n) \leq 1/(n^2\delta_n^2)$. Thus, we're done if we choose δ_n such that this converges, e.g. $\delta_n = 1/n^{1/4}$.

We still have to prove the result on subsequences: that $s_n/n \to 0$ iff $s_{n^2}/n^2 \to 0$. Unfortunately, I'm really far from the board and can't read the professor's handwriting, so this is going to remain an exercise... and full of tricky typos.

It looks like we are assuming that $y_{n^2}/n^2 \to 0$ and

$$\lim_{n\to\infty} \frac{(y_n - y_{k_n+1}}{k_n^2} \to 0,$$

where $k_n = \lfloor \sqrt{n} \rfloor$. Then, we want to show that $y_n/n \to 0$.

Since $k_n^2 \leq n \leq (k_n+1)^2$, then

$$\begin{aligned} \left| \frac{y_n}{n} \right| &\leq \frac{|y_n| k_n^2}{=} \frac{|y_{k_n} - (y_{k_n} - y_n)|}{k_n^2} \\ &\leq \frac{|y_{k_n}| k_n^2}{+} \frac{|y_{k_n} - y_n|}{k_n^2}. \end{aligned}$$

Thus, if y_n/n is bounded, then we can expand over the bound and get the second term to go to zero. □

For the strong law, we used something more general than just Chebyshev's inequality; eiither you need better bounds (e.g. the method of fourth moments, as in the book), or subsequences as we used. The two statements are different, and it's important to see how: if $P(j) = c/(j62\log|j|)$ for $|j| \geq 2$, then pick $X_1, \ldots, X_n$ independently from P_j. These are symmetric, so the mean tends to 0 as $n \to \infty$, but the strong law fails: we can't know how many to try until it's below a certain ε. It has no finite content.

So this probabilistic argument may be slightly confusing, but it stands as a motivating example of dealing with complicated sets probabilistically. The next topic will be the language that allows us to define probability spaces: algebras, σ-algebras, and so forth.

Definition. Let Ω be a set; then, a *field of subsets* of Ω is a collection $\mathcal{F} \subseteq \Omega$ such that

(1) $\emptyset \in \mathcal{F}$.

(2) If $F \in \mathscr{F}$, then $F^c \in \mathscr{F}$.
(3) If $A, B \in \mathscr{F}$, then $A \cup B \in \mathscr{F}$.

Example 2.2.

(1) The simplest example is $\mathscr{F} = \{\emptyset, \Omega\}$.
(2) If $\Omega = (0,1]$, then we can take $\mathscr{F}$ to be the set of all finite unions of disjoint intervals of half-open sets; the reason we use (and used above) half-open sets is so that complements work out correctly.

Definition. A σ-field is a field such that (3) is generalized to countable collections: if $F_n \in \mathscr{F}$ for $n = 1, 2, \dots$, then $\bigcup_{n=1}^{\infty} F_n \in \mathscr{F}$.

Thus, all subsets of Ω is a σ-field, and if $\{\mathscr{F}_n\}_{n \in I}$ is a collection of σ-fields, then their intersection $\bigcap_{n \in I} \mathscr{F}_n$ is also a σ-field.

Thus, we can take any collection $\mathscr{C}$ of subsets of Ω and define $\mathscr{F}(\mathscr{C})$ to be the smallest σ-field containing $\mathscr{C}$, which is the intersection of all σ-fields containing $\mathscr{C}$ (since that intersection itself is a σ-field containing $\mathscr{C}$, and is certainly the smallest).

This is right about where things get less trivial, but we also lost contact with phyisical intuition.

Definition. The *Borel sets* are the elements of the σ-field generated by $\mathscr{C} = \{(a,b] : 0 < a \le b \le 1\}$.

It's easy to work with Borel sets, but trying to describe them is very difficult; some are very complicated. This can be unsettling, but we'll see how it's possible to work with them anyways.

Definition. A *probability space* $(\Omega, \mathscr{F}, P)$ is a set Ω, σ-algebra $\mathscr{F}$ of subsets of Ω, and a function $P : \mathscr{F} \to [0,1]$ such that

(1) $P(\emptyset) = 0$,
(2) $P(A^c) = 1 - P(A)$ for all $A \in \mathscr{F}$, and
(3) If $A_1, A_2, \dots \in \mathscr{F}$ are all pairwise disjoint, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

Apparently one of Persei's friends has the license plate `F OMEGA P`.

We've learned some things since the Greeks; for example:

(1) It's not possible to define a probability on the σ-algebra of all subsets of some sets Ω. The professor's Halloween lecture will explain this further — this is when the monsters come out.
(2) In some sense, measure-theoretic probability is a wonderful approximate theory.

Suppose Ω is a set and $\mathscr{F}_0$ is a field (not necessarily a σ-field) and $P : \mathscr{F}_0 \to [0,1]$ is a probability on $\mathscr{F}_0$. Then, for *any* $A \subseteq \Omega$, define

$$P^*(A) = \inf \sum_{i=1}^{\infty} P(F_n),$$

where the infimum is taken over all coverings $F_1, \dots$ of A within $\mathscr{F}_0$. Basically, take any horrible set, and cover it by sets that we understand how to assign probabilities to. This is called the *outer measure* of A.

Here's a good idea in this field, given by Carathéodory: define a collection of *measurable sets* $\mathscr{M}$ whose probability we can understand. Specifically, we say that $M \in \mathscr{M}$ if for all $E \subseteq \Omega$, $P^*(E) = P^*(E \cap M) + P^*(E \cap M^c)$. That is, it splits every test set, no matter how horrible.

All right, but how did anyone think of *that*? It works and is useful... I guess sometimes people just have ideas sometimes.

Theorem 2.3. *Given an algebra $\mathscr{F}$ of subsets of Ω and a probability $P : \mathscr{F} \to [0,1]$,*

(1) $\mathscr{M}$ is a σ-algebra containing $\mathscr{F}$, and
(2) $P^*|_{\mathscr{M}}$ is a probability, and is equal to P.
(3) P^* is unique.

We'll do this in three steps: first showing that $P^*(\emptyset) = 0$, then that if $A \subseteq B$, then $P^*(A) \le P^*(B)$, and finally that if $\{A_n\}_{n=1}^{\infty}$ is any collection of sets, then

$$P^*\left(\bigcup_{n=1}^{\infty} A_n\right) \le \sum_{n=1}^{\infty} P^*(A_i).$$

Let's look at the last one in more detail. For all $\varepsilon > 0$, there are coverings $F_{i,k} \in \mathscr{F}$, where $\bigcup_k F_{i,k} \supseteq A_i$ and such that

$$\sum_{k=1}^{\infty} P(F_{i,k}) \le P^*(A_i) + \frac{\varepsilon}{2}.$$

To be continued...

3. The π-λ Theorem: 9/28/14

"By chalk, by math, by logic, it is defined!"

Recall that if we have a ground set Ω and a field $\mathscr{F}$ of subsets, with a probability P defined on $\mathscr{F}$, then for all $E \subseteq \Omega$, we can define the outer measure as the infinum over all countable covers $\{F_n\}$ with $F_n \in \mathscr{F}$:

$$P^*(E) = \inf \sum_1^\infty P(F_n).$$

Then, we defined

$$\mathscr{M} = \{A \subseteq \Omega : P^*(E) = P^*(A \cap E) + P^*(E \cap A^c) \text{ for all } E \subseteq \Omega\}.$$

Then, we stated Theorem 2.3, which provides that P^* is a probability on $\mathscr{M}$, but we have to prove it.

Proof of Theorem 2.3. First, we need to prove that $\mathscr{M}$ is a field. For any $A \subseteq \Omega$, $P^*(E) = P^*((A \cap E) \cup (A \cap E^c)) \leq P^*(A \cap E) + P^*(A^c \cap E)$, so $A \in \mathscr{M}$ iff $P^*(E) \geq P^*(A \cap E) + P^*(A^c \cap E)$. Thus, $\emptyset \in \mathscr{M}$ of course, and compliments are just as easy, since the expression is unchanged if A and A^c are switched. Thus, we need to show that if $A, B \in \mathscr{M}$, then $A \cap B \in \mathscr{M}$.

At this point, drawing a picture of a Venn diagram may be useful.

$$\begin{aligned} P^*(E) &= P^*(B \cap E) + P^*(B^c \cap E) \\ &= P^*(A \cap B \cap E) + P^*(A^c \cap B \cap E) + P^*(A \cap B^c \cap E) + P^*(A^c \cap B^c \cap E) \\ &= P^*(A \cap B \cap E) + P^*((A \cap B \cap E) \cup (A \cap B^c \cap E) \cup (B^c \cap A^c \cap E)) \\ &= P^*(A \cap B \cap E) + P^*((A \cap B)^c \cap E). \end{aligned}$$

The next thing we need to prove is countable additivity: if $\{A_n\}_{n=1}^\infty \in \mathscr{M}$ and are pairwise disjoint, then for any $E \subseteq \Omega$,

$$P\left(E \cap \bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty P^*(E \cap A_n).$$

We'll show it for the finite case by induction, and then use limits to generalize to the countable case. The induction is obvious for $n = 1$, and for $n = 2$, if $A_1 = A_2^c$, then it's obvious and otherwise,

$$\begin{aligned} L^*(E \cap (A_1 \cup A_2)) &= P^*(E \cap (A_1 \cup A_2) \cap A_1) + P^*(E \cap (A_1 \cup A_2) \cap A_1^c) \\ &= P^*(E \cap A_1) + P^*(E \cap A_2). \end{aligned}$$

This argument also works for finite n: pick out one, and then use the above reasoning to reduce to the $n-1$ case.

For the countable case, we have

$$P^*\left(E \cap \bigcup_1^\infty A_n\right) \geq P^*\left(E \cap \bigcup_1^n A_n\right) = \sum_{j=1}^n P^*(A_j \cap E);$$

then, let $n \to \infty$.

The next thing we need to show is that $\mathscr{M}$ is a σ-algebra, and P^* is countably additive on $\mathscr{M}$. Given any countable collection $A_1, A_2, \cdots \in \mathscr{M}$, disjointify[3] them, letting $A_1' = A_1$, $A_2' = A_2 \cap (A_1')^c$, and so forth, in general letting

$$A_n' = A_n \cap \left(\bigcup_{i=1}^n A_i'\right)^c.$$

Thus, the unions of the A_n and the A_n' are the same; call this union A, and let $F_n = \bigcup_{i=1}^n A_n'$, which we known is in $\mathscr{M}$ for each n. Thus,

$$\begin{aligned} P^*(E) &= P^*(E \cap F_n) + P^*(E \cap F_n^c) \geq P^*(E \cap F_n) + P^*(E \cap A^c) \\ &= \sum_{i=1}^n P^*(E \cap A_n') + P^*(E \cap A'). \end{aligned}$$

Thus,

$$\begin{aligned} P^*(E) &\geq \sum_1^\infty P^*(E \cap A_n') + P^*(E \cap A^c) \\ &= P^*\left(E \cap \bigcup_1^\infty A_n'\right) + P^*(E \cap A^c) \\ &= P^*(E \cap A) + P^*(E \cap A^c). \end{aligned}$$

Moreover, we have already shown that P^* is additive on M.

[3] Is this a word? I hope so.

Finally, we have to show that $\mathscr{F} \subseteq \mathscr{M}$. For any $\varepsilon > 0$ and $E \subseteq \Omega$, choose an $A \in \mathscr{F}$ and $A_1, A_2 \cdots \in \mathscr{F}$ such that E is covered by $\{A_n\}$ but

$$\sum_1^\infty P(A_n) \le P^*(E) + \varepsilon.$$

Set $B_n = A | cap A_n$ and $C_n = A^c \cap A_n$, so

$$E \cap A \subset \bigcup_1^\infty B_n \qquad \text{and} \qquad E \cap A^c \subset \bigcup_1^\infty C_n.$$

Then,

$$\begin{aligned} P^*(A \cap E) + P^*(A^c \cap E) &= \sum_n P(B_n) + P(C_n) \\ &= \sum_n P(A_n) \le P^*(E) + \varepsilon. \end{aligned}$$

This is about the only place where we used the actual definition of the outer measure.

Finally, we can prove that $P^*(F) = P(F)$; we already know it's at least as much, and we can cover F by $F_1, F_2, \dots$ in $\mathscr{F}$. Then,

$$P(F) \le \sum_n P(F \cap F_n) \le \sum_1^\infty P(F_n),$$

so $P^*(F) \le P(F)$. □

Though this is the right definition, in that this theorem is demonstrated, it's very abstract: nobody actually uses it to compute, and it can be specialized to Lesbegue measure, where it's a bit more tractable.

In this case, $\Omega = (0,1]$, and $\mathscr{F}$ is taken to be the set of finite unions of disjoint intervals, with

$$P\left(\sum_{i=1}^n (a_i, b_i]\right) = \sum_{i=1}^n b_i - a_i.$$

Unfortunately, it's a bit ☹ to show this,[4] but we can break it down into showing that if $\bigcup_1^\infty I_n \subseteq I$, then $\sum (b_n - a_n) \le b - a$ (where $I = (a,b]$ and $I_n = (a_n, b_n]$; and that if the inclusion is switched, then the inequality is also switched.

The first one follows from induction on n, at which point it's easy from disjointness, and the second point actually requires choosing an $\varepsilon > 0$, such that $\varepsilon < b - a$ and

$$[a + \varepsilon, b] \subseteq \bigcup_{n=1}^i nfty\left(a_n, b_n + \frac{\varepsilon}{2}\right).$$

Then, the Heine-Borel theorem implies there's a finite subcover of $(a_1, b_1], \dots, (a_N, b_N]$. This is really analysis (where the rest was definition-chasing), and

$$(b - a - \varepsilon) \le \sum_1^\infty (b_n - a_n) + \varepsilon,$$

so the result is shown when $\varepsilon \to 0$.

One of the most useful things to have happened in the last fifty years is something called the π-λ theorem, which is a bit technical. We'll prove it, but first we'll use it to avoid getting too technical too soon.

Definition. A collection of sets $\mathscr{L}$ in Ω is a π-system if it is closed under finite intersections. It is a λ-system if $\emptyset \in \mathscr{L}$, $\mathscr{L}$ is closed under complements, and countable unions of disjoint sets in $\mathscr{L}$ are in $\mathscr{L}$.

Example 3.1. For example, if $\Omega = \{1,2,3,4\}$, then $\mathscr{L} = \{\emptyset, \Omega, \{1,2\}, \{3,4\}, \{1,3\}, \{2,4\}\}$ is a λ-system but not a field.

Theorem 3.2 (π-λ). *Let $\mathscr{P}$ be a π-system on Ω and $\mathscr{L}$ be a λ-system on Ω; then, if $\mathscr{P} \subseteq \mathscr{L}$, then $\sigma(\mathscr{P}) = \mathscr{L}$.*

This theorem has plenty of applications:

(1) If two probability measures agree on a π-system $\mathscr{P}$ (e.g. on the set of intervals!), then they agree on $\sigma(\mathscr{P})$.

Proof. Call these two probability measures P and Q, and let $\mathscr{L} = \{A : P(A) = Q(A)\}$; this happens to be a λ-system (which is the bulk of the argument, but devolves to definition-checking), and it contains $\mathscr{P}$, so $\sigma(\mathscr{P}) \subseteq \mathscr{L}$. □

(2) As a corollary, we get the uniqueness in Carathéadory's theorem.

[4] The professor actually drew the frowny face on the chalkboard.

4. Independence: 10/1/14

"They're trying to make me not heard, BUT THEY WILL NOT SUCCEED!"

Recall that a class $\mathscr{P}$ is a π-system if whenever $A,B\in\mathscr{P}$, then $A\cap B\in\mathscr{P}$, and that a class $\mathscr{L}$ is a λ-system on Ω if $\Omega\in\mathscr{L}$, $\mathscr{L}$ is closed under complements, and $\mathscr{L}$ is closed under countable disjoint unions.

Claim. If $\mathscr{L}$ is a λ-system and $A,B\in\mathscr{L}$ are such that $A\subseteq B$, then $B\setminus A=B\cap A^c\in\mathscr{L}$.

Proof. Since $B\in\mathscr{L}$, then $B^c\in\mathscr{L}$ and it's disjoint from A, so $A\cup B^c\in\mathscr{L}$, and thus so is its complement $A^c\cap B$. □

Claim. If $\mathscr{L}$ is a σ-system and a π-system, then it is a σ-algebra.

Proof. We just need to show that non-disjoint unions are in $\mathscr{L}$, since the rest follow by definition. For any $A,B\in\mathscr{L}$, $A\cup B=(A\setminus(A\cap B))\cup(A\cap B)\cup(B\setminus(A\cap B))$ as a disjoint union, so it's in $\mathscr{L}$; similarly, one can "disjointify" a countable collection of sets into a disjoint countable collection, so its union is in $\mathscr{L}$. □

Recall also the π-λ theorem, Theorem 3.2.

Proof of Theorem 3.2. Let $\mathscr{L}_0$ be the smallest λ-system containing $\mathscr{P}$; we'll show that $\mathscr{L}_0$ is also a π-system, so $\sigma(\mathscr{P})\subseteq\mathscr{L}_0\subseteq\mathscr{L}$.

For any $A\in\mathscr{P}$, let $\mathscr{L}_A=\{B\subseteq\Omega : A\cap B\in\mathscr{L}_0\}$. Then, $\mathscr{L}_A$ is a λ-system: $\Omega\in\mathscr{L}_A$, certainly, and if $B\in\mathscr{L}_A$, then $A\cap B\in\mathscr{L}_0$, so $A\setminus(A\cap B)\in\mathscr{L}_0$ and $A\cap(A^c\cup B^c)-A\cap B^c\in\mathscr{L}_0$, so $B^c\in\mathscr{L}_A$. Finally, if $B_1,\dots$ is a disjoint collection in L_A, then each $B_i\cap A\in L_)$ are till disjoint, so

$$\left(\bigcup B_i\right)\cap A=\bigcup_1^\infty B_i\cap A\in\mathscr{L}_0.$$

Thus, the union of the B_i is in $\mathscr{L}_A$.

Moreover, if $A\in\mathscr{P}$, then $\mathscr{L}_0\subseteq\mathscr{L}_A$, because if $A,B\in\mathscr{P}$, then $\mathscr{P}\subseteq\mathscr{L}_A$, so $\mathscr{L}_0\subseteq\mathscr{L}_A$ (there's something to show here, but only a single line).

If $B\in\mathscr{L}_0$, then $\mathscr{L}_0\in\mathscr{L}_B$, because if $A\in\mathscr{P}$, then $B\in\mathscr{L}_A$, so $A\cap P\in\mathscr{L}_0$ and thus $A\in\mathscr{L}_B$. Thus, $\mathscr{P}\subseteq\mathscr{L}_B$, so $\mathscr{L}_0\in\mathscr{L}_B$.

Finally, if $B,C\in\mathscr{L}_0$ and $C\in\mathscr{L}_B$, then $\mathscr{L}_0$ is a π-system; thus, $\mathscr{L}_0$ is a σ-algebra in $\mathscr{L}$, and $\sigma(\mathscr{P})\subseteq\mathscr{L}_0\subseteq\mathscr{L}$. □

Independence. Kolmogorov, when asked to explain the difference between probability and analysis, responded that probability is analysis plus independence; that may have been truer in the 1930s than today, but independence is still important.

Definition. Let $(\Omega,\mathscr{F},P)$ be a probability space; then, $A,B\in\mathscr{F}$ are independent if $P(A\cap B)=P(A)P(B)$. A collection $\{A_n\}_{n\in I}$ is *independent* if for any finite subcollection $A_{n_1},\dots,A_{n_m}$ we have

$$P\left(\bigcap_{i=1}^m A_{n_i}\right)=\prod_{i=1}^n P(A_{n_i}).$$

Example 4.1. Remember $\Omega=(0,1]$ and $w=\sum_{n=1}^\infty d_n(\omega)/2^n$. If $A_n=\{d_n=0\}$, then the collection $\{A_n\}$ is independent.

Let $\mathscr{A}_i=\sigma(A_i)-\{\emptyset,\Omega,\{d_i=0\},\{d_i=1\}\}$ (in some sense, all you can know about the i^{th} digit). Then, let $\mathscr{A}_{\text{ODD}}=\sigma(\mathscr{A}_{2n+1},n\in\mathbb{N}\}$ and $\mathscr{A}_{\text{EVEN}}=\sigma(\mathscr{A}_{2n},n\in\mathbb{N}\}$. Intuitively, $\mathscr{A}_{\text{ODD}}$ and $\mathscr{A}_{\text{EVEN}}$ are independent, but how do we show this?

Proposition 4.2. *Let $(\Omega,\mathscr{F},P)$ be a probability space and $\mathscr{A}_n$, $1\le n\le I$ be independent π-systems. Then, $\sigma(\mathscr{A}_n)$ for $1\le n\le I$ are independent.*

Proof. Let $\mathscr{B}_i=A_i\cup\Omega$, which is still a π-system, and the $\mathscr{B}_i$ are still independent.

Let

$$\mathscr{L}_1=\left\{B_1\in\Omega:\text{ for all choices } B_2,\dots,B_I\in\Omega, P\left(\bigcup_{j=1}^I B_j\right)=\prod_{j=1}^n P(B_i)\right\}.$$

Then, $\mathscr{B}_1\subseteq\mathscr{L}_1$, and $\mathscr{L}_1$ is a λ-system (just check all of the axioms). Thus, $\sigma(\mathscr{B}_1)\subseteq\mathscr{L}_1$. Now, independence follows. □

We're about to do two extremely useful, trivial theorems, and one hard, useless theorem.

Definition. The statement that A_n *occurs infinitely often* denotes the set

$$\bigcap_{n=1}^\infty\bigcup_{m=n}^\infty A_n,$$

which is the set of ω in which A_n occurs infinitely often (e.g. in a decimal expansion).

Theorem 4.3 (Borel-Cantelli). *Let $(\Omega,\mathscr{F},P)$ be a probability space and $\{A_n\}_{n=1}^\infty\in\mathscr{F}$. Then,*

(1) If $\sum_{n=1}^\infty P(A_n)$ is finite, then $P\{A_n$ infinitely often$\}=0$.
(2) If the sum is infinite and the A_n are independent, then $P\{A_n$ infinitely often$\}=1$.

Proof. For the first part, given an $\varepsilon > 0$,

$$\begin{aligned} P(A_n \text{ infinitely often}) &\le P\left(\bigcup_{i=N}^{\infty} A_i\right) \\ &\le \sum_{N}^{\infty} P(A_i) < \varepsilon. \end{aligned}$$

For the second part, we'll use that $1 - x \le e^{-x}$ for all $x > 0$, and study

$$P\left((A_i \text{ infinitely often})^c\right) = P\left(\bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_n^c\right).$$

Now,

$$\begin{aligned} P\left(\bigcap_{i=n}^{\infty} A_i^c\right) &\le P\left(\bigcap_{i=n}^{N} A_i^c\right) \\ &= \prod_{i=n}^{N} P(A_i^c) \\ &= \prod_{i=n}^{N} (1 - P(A_i)) \\ &\le e^{-\sum_n^N P(A_n)}. \end{aligned}$$

Letting $n \to \infty$, the right-hand side goes to 0, so the probability of the complement of A_i happening infinitely often is 0, so $P(A \text{ infinitely often}) = 1$. □

Now for a useless, but beautiful, theorem.

Return to $\Omega = (0,1]$, $\omega = \sum d_n/2^n$, and $r_n = 2d_n - 1 \in \{\pm 1\}$. The strong law showed that $s_n = \sum_{i=1}^n r_n$, and $s_n/n \to 0$.

Recall the definitions of the $\underline{\lim}$ and $\overline{\lim}$: if $x_n \in \mathbb{R}$, then

$$\begin{aligned} \overline{\lim} x_n &= \lim_{n\to\infty} \sup_{i \ge n} x_i, \text{ and} \\ \underline{\lim} x_n &= \lim_{n\to\infty} \inf_{i \ge n} x_i. \end{aligned}$$

The sequence $\{x_n\}$ has a limit iff $\overline{\lim} x_n = \underline{\lim} x_n$.

Interestingly, $\overline{\lim} s_n = +\infty$ and $\underline{\lim} s_n = -\infty$, and this is also true for $s_n/\sqrt{n}$, and Hardy showed $\lim s_n/(\sqrt{n}\log n) = 0$. But we can prove a tighter bound on the fluctuation.

Theorem 4.4 (Law of the Iterated Logarithm)**.**

$$\overline{\lim} \frac{s_n}{\sqrt{2n \log\log n}} = 1 \qquad \textit{and} \qquad \underline{\lim} \frac{s_n}{\sqrt{2n \log\log n}} = -1.$$

Remark.

(1) In statistical testing (in the real world), people look at the event $\{s_n/n > .96\}$, which is usually taken as evidence for the hypothesis, but the theorem shows that if you wait long enough, this will eventually happen, even for a fair coin!

(2) Well, but why is it useless? The issue is that it takes a long time: $\log\log 10^{100} = 5.43$, so don't sit around waiting for it to happen.

The proof needs tail bounds for the binomial distribution, which come from the homework (and aren't too bad with hints).