STA 711: Probability & Measure Theory
Robert L. Wolpert

3 Random Variables & Distributions

Let Ω be any set, $\mathcal{F}$ any σ-field on Ω, and P any probability measure defined for each element of $\mathcal{F}$; such a triple $(\Omega, \mathcal{F}, \mathsf{P})$ is called a *probability space*. Let $\mathbb{R}$ denote the real numbers $(-\infty, \infty)$ and $\mathcal{B}$ the Borel sets on $\mathbb{R}$ generated by (for example) the half-open sets $(a, b]$.

Definition 1 *A real-valued* Random Variable *is a function* $X : \Omega \to \mathbb{R}$ *that is "*$\mathcal{F}\backslash\mathcal{B}$*-measurable", i.e., that satisfies* $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}$ *for each Borel set* $B \in \mathcal{B}$ *(or, equivalently, simply for each set* B *of the form* $(-\infty, b]$ *for some rational* $-\infty < b < \infty$*).*

This is sometimes denoted simply "$X^{-1}(\mathcal{B}) \subset \mathcal{F}$." Since the probability measure P is only defined on sets $F \in \mathcal{F}$, a random variable *must* satisfy this condition if we are to be able to find the probability $\mathsf{P}[X \in B]$ for each Borel set B, or even if we want to find the distribution function (DF) $F_X(b) := \mathsf{P}[X \le b]$ for each rational number b. Note that set-inverses are rather well-behaved functions from one class of sets to another; specifically, for any collection $\{A_\alpha\} \subset \mathcal{B}$ (countable or not),

$$[X^{-1}(A_\alpha)]^c = X^{-1}\big((A_\alpha)^c\big) \quad \text{and} \quad \bigcup_\alpha X^{-1}(A_\alpha) = X^{-1}\big(\bigcup_\alpha A_\alpha\big)$$

from which it follows that $\cap_\alpha X^{-1}(A_\alpha) = X^{-1}\big(\cap_\alpha A_\alpha\big)$. Thus, whether X is measurable or not, $X^{-1}(\mathcal{B})$ *is* a σ-field if $\mathcal{B}$ is. It is denoted $\mathcal{F}_X$ (or $\sigma(X)$), called the "sigma field generated by X," and is the smallest sigma field $\mathcal{G}$ such that X is $(\mathcal{G}\backslash\mathcal{B})$- measurable. In particular, X is $(\mathcal{F}\backslash\mathcal{B})$- measurable if and only if $\sigma(X) \subset \mathcal{F}$.
Warning: The backslash character "\" in this notation is entirely unrelated to the backslash character that appears in the common notation for set exclusion, $A \setminus B := A \cap B^c$.

In probability and statistics, sigma fields represent *information*: a random variable Y is measurable over $\mathcal{F}_X$ if and only if the value of Y can be found from that of X, *i.e.*, if $Y = \varphi(X)$ for some function φ. Note the difference in perspective between real analysis, on the one hand, and probability & statistics, on the other; in analysis it is only *Lebesgue* measurability that mathematicians worry about, and only to avoid paradoxes and pathologies. In probability and statistics we study measurability for a variety of sigma fields, and the (technical) concept of measurability corresponds to the (empirical) notion of *observability.*

3.1 Distributions

A random variable X on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ *induces* a measure μ_X on $(\mathbb{R}, \mathcal{B})$, called the *distribution measure* (or simply the *distribution*), via the relation

$$\mu(B) := \mathsf{P}[X \in B],$$

sometimes written more succinctly as $\mu_X = \mathsf{P} \circ X^{-1}$ or even $\mathsf{P}X^{-1}$.

3.1.1 Functions of Random Variables

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, X a (real-valued) random variable, and $g : \mathbb{R} \to \mathbb{R}$ a (real-valued $\mathcal{B}\backslash\mathcal{B}$) measurable function. Then $Y = g(X)$ is a random variable, *i.e.*,

$$Y^{-1}(B) = X^{-1}(g^{-1}(B)) \in \mathcal{F}$$

for any $B \in \mathcal{B}$. Also every continuous or piecewise-continuous real-valued function on $\mathbb{R}$ is $\mathcal{B}\backslash\mathcal{B}$-measurable. How are $\sigma(X)$ and $\sigma(Y)$ related?

3.1.2 Random Vectors

Denote by $\mathbb{R}^2$ the set of points (x, y) in the plane, and by $\mathcal{B}^2$ the sigma field generated by rectangles of the form $\{(x, y) : a < x \le b, c < y \le d\} = (a, b] \times (c, d]$. Note that finite unions of those rectangles (with a, b, c, d in the *extended* reals $[-\infty, \infty]$) form a field $\mathcal{F}_0^2$, so the minimal sigma field and minimal λ system containing $\mathcal{F}_0^2$ coincide, and the assignment $\lambda_0^2\big((a, b] \times (c, d]\big) = (b - a) \times (d - c)$ has a unique extension to a measure on all of $\mathcal{B}^2$, called two-dimensional Lebesgue measure (and denoted λ^2). Of course, it's just the area of sets in the plane.

An $\mathcal{F}\backslash\mathcal{B}^2$-measurable mapping $X : \Omega \to \mathbb{R}^2$ is called a (two-dimensional) *random vector*, or simply an $\mathbb{R}^2$-valued random variable, or (a bit ambiguously) an $\mathbb{R}^2$-RV. It's easy to show that the components X_1, X_2 of a $\mathbb{R}^2$-RV X are each RVs, and conversely that for any two random variables X_1 and X_2 the two-dimensional RV $(X_1, X_2) : \Omega \to \mathbb{R}^2$ is $\mathcal{F}\backslash\mathcal{B}^2$-measurable, *i.e.*, is a $\mathbb{R}^2$-RV.

Also, any Borel measurable (and in particular, any piecewise-continuous) function $f : \mathbb{R}^2 \to \mathbb{R}$ induces a random variable $f(X, Y)$: this shows that such combinations as $X + Y$, X/Y, $X \wedge Y$, $X \vee Y$, *etc.* are all random variables if X and Y are.

The same ideas work in any finite number of dimensions, so without any special notice we will regard n-tuples $(X_1, ..., X_n)$ as $\mathbb{R}^n$-valued RVs, or $\mathcal{F}\backslash\mathcal{B}^n$-measurable functions, and will use Lebesgue n-dimensional measure λ^n on $\mathcal{B}^n$. Again $\sum_i X_i$, $\prod_i X_i$, $\min_i X_i$, and $\max_i X_i$ are all random variables. For any metric space (E, d) with Borel sets $\mathcal{E}$, a $\mathcal{F}\backslash\mathcal{E}$-measurable function $X : \Omega \to E$ will be called an "E-valued random variable" (although some authors prefer the term "random element of E" unless E is $\mathbb{R}$ or perhaps $\mathbb{R}^n$).

Even if we have *infinitely many* random variables we can verify the measurability of $\sum_i X_i$, $\inf_i X_i$, and $\sup_i X_i$, and of $\liminf_i X_i$, and $\limsup_i X_i$ as well: for example,

$$[\omega : \sup_i X_i(\omega) \le r] = \bigcap_{i=1}^{\infty} [\omega : X_i(\omega) \le r]$$

$$[\omega : \limsup_i X_i(\omega) \le r] = \bigcup_{j=1}^{\infty} \bigcap_{i=j}^{\infty} [\omega : X_i(\omega) \le r].$$

The event "X_i converges" is the same as

$$\left[\omega : \limsup_i X_i(\omega) - \liminf_i X_i(\omega) = 0\right] = \bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{i,j=n}^{\infty} [\omega : \ |X_i(\omega) - X_j(\omega)| < \epsilon_k]$$

for any positive sequence $\epsilon_k \to 0$, and so is $\mathcal{F}$-measurable and has a well defined probability $\mathsf{P}[\limsup_i X_i = \liminf_i X_i]$. This is one point where countable additivity (and not just finite additivity) of P is crucial, and where $\mathcal{F}$ must be a sigma field (and not just a field).

3.1.3 Example: Discrete RVs

If a RV X can take on only a finite or countable set of distinct values, say $\{b_i\}$, then each set $\Lambda_i = \{\omega : X(\omega) = b_i\}$ must be in $\mathcal{F}$. The random variable X can be written:

$$X(\omega) = \sum_i b_i \mathbf{1}_{\Lambda_i}(\omega), \qquad \text{where} \tag{*}$$

$$\mathbf{1}_\Lambda(\omega) = \begin{cases} 1 & \text{if } \omega \in \Lambda \\ 0 & \text{if } \omega \notin \Lambda \end{cases} \tag{1}$$

is the so-called *indicator function* of Λ. Since $\Omega = \cup \Lambda_i$, the $\{\Lambda_i\}$ form a "countable partition" of Ω. Any RV can be approximated uniformly as well as we like by a RV of the form $(*)$ (how?). Note that the indicator function $\mathbf{1}_A$ of the limit supremum $A := \limsup_i A_i$ of a sequence of events is equal pointwise to the indicator $\mathbf{1}_A(\omega) = \limsup_i \mathbf{1}_{A_i}(\omega)$ of their limit supremum.

3.2 Explicit Construction of Sigma Fields [optional]

Ordinals and Transfinite Induction

Every finite set S (say, with $n < \infty$ elements) can be *totally ordered* $a_1 \prec a_2 \prec a_3 \prec \cdots \prec a_n$ in $n!$ ways, but in some sense every one of these is the same— if $\prec_1$ and $\prec_2$ are two orderings, there exists a 1–1 order-preserving isomorphism $\varphi : (S, \prec_1) \longleftrightarrow (S, \prec_2)$. Thus *up to isomorphism* there is only one ordering for any finite set.

For countably infinite sets there are many different orderings. The obvious one is $a_1 \prec a_2 \prec a_3 \prec \cdots$, ordered just like the positive integers $\mathbb{N}$; this ordering is called ω, the first *limit ordinal.* But we could pick any element (say, $b_1 \in S$) and order the remainder of S in the usual way, but declare $a_n \prec b_1$ for every $n \in \mathbb{N}$; one element is "bigger" (in the ordering) than all the others. This is *not* isomorphic to ω, and it is called $\omega + 1$, the *successor* to ω. If we set aside two elements (say, $b_1 \prec b_2$) to follow all the others we have $\omega + 2$, and similarly we have $\omega + n$ for each $n \in \mathbb{N}$. The limit of all these is $\omega + \omega$, or 2ω... it is the ordering we would get if we lexicographically ordered the set $\{(i, j) : i = 1, 2 \ j \in \mathbb{N}\}$ of the first two rows of integers in the first quadrant, declaring $(1,j) \prec (2,k)$ for every j,k and otherwise $(i,j) \prec (i,k)$ if $j < k$.

We would get the successor to this, $2\omega+1$, by extending the lexicographical ordering as we add $(3,1)$ to S; in an obvious way we get $2\omega+n$ for every $n \in \mathbb{N}$ and eventually the limit ordinals 3ω, 4ω, *etc.*, and the successor ordinals $m\omega+n$. The limit of all these is $\omega\omega$ or ω^2, the lexicographical ordering of the entire first quadrant of integers (i,j). It too has successors ω^2+n (graphically you can think about integer triplets (i,j,k)), and limits like $\omega^2+\omega$ and ω^3 and ω^ω (which turns out to be the same as 2^ω).

In general an ordinal is a *successor* ordinal if it has a maximal element, and otherwise is a *limit* ordinal. Every ordinal α has a successor $\alpha+1$, and every set of ordinals $\{\alpha_n\}$ has a limit (least upper bound) λ. Let Ω be the first *un*countable ordinal.

Proofs and constructions by *transfinite induction* typically have one step at each successor ordinal, and another at each limit ordinal. The *Borel sets* can be defined by transfinite construction as follows.

Let $\mathcal{F}_0$ be the class of open subsets of some topological space $\mathcal{X}$ (perhaps the real numbers $\mathcal{X} = \mathbb{R}$, for example).

Succ: For any ordinal α, let $\mathcal{F}_{\alpha+1}$ be the class of countable unions of sets $E_n \in \mathcal{F}_\alpha$ and their complements $E_m : E_m^c \in \mathcal{F}_\alpha$.

Lim: For any limit ordinal λ, let $\mathcal{F}_\lambda = \cup_{\alpha \prec \lambda} \mathcal{F}_\alpha$.

Together these define $\mathcal{F}_\alpha$ for all ordinals, limit and successor; the sigma field *generated by* $\mathcal{F}_0$ is just $\mathcal{F}_\Omega$, where Ω is the first uncountable ordinal. It remains to prove that:

1. $\mathcal{F}_0 \subset \mathcal{F}_\Omega$, *i.e.*, $\mathcal{F}_\Omega$ contains the open sets (including $\mathcal{X}$ itself);
2. $E \in \mathcal{F}_\Omega \implies E^c \in \mathcal{F}_\Omega$, *i.e.*, $\mathcal{F}_\Omega$ is closed under complements;
3. $E_n \in \mathcal{F}_\Omega \implies \cup_{n=1}^\infty E_n \in \mathcal{F}_\Omega$, *i.e.*, $\mathcal{F}_\Omega$ is closed under countable unions;
4. $\mathcal{F}_\Omega \subset \mathcal{G}$ for any sigma field $\mathcal{G}$ containing $\mathcal{F}_0$.

Item 1. is trivial since $\mathcal{F}_\Omega = \cup_{\alpha \prec \Omega} \mathcal{F}_\alpha$, and in particular contains $\mathcal{F}_0$. Item 2. follows by transfinite induction upon noting that $E \in \mathcal{F}_\alpha \implies E^c \in \mathcal{F}_{\alpha+1}$. Item 3 follows by noting that $E_n \in \mathcal{F}_\Omega \implies E_n \in \mathcal{F}_{\alpha_n}$ for some $\alpha_n \prec \Omega$, and $\beta = \sup_{n<\infty} \alpha_n$ is an ordinal satisfying $\alpha_n \preceq \beta \prec \Omega$ and hence $E_n \in \mathcal{F}_\beta$ for all n and $\cup_{n=1}^\infty E_n \in \mathcal{F}_{\beta+1}$. Verifying the minimality condition Item 4 is left as an exercise.

It isn't immediately obvious from the construction that we couldn't have stopped earlier—for example, that $\mathcal{F}_2$ or $\mathcal{F}_\omega$ isn't already the Borel sets, unchanging as we allow successively more intersections and unions. In fact that happens if the original space $\mathcal{X}$ is countable or finite; in the case of $\mathbb{R}$, however, one can show that $\mathcal{F}_\alpha \neq \mathcal{F}_{\alpha+1}$ for every $\alpha \prec \Omega$.

Do you think this explicit construction is clearer or more complicated than the completion argument used in the text?

3.3 Infinite Coin Toss

For each $\omega \in \Omega = (0, 1]$ and integer $n \in \mathbb{N}$ let $\delta_n(\omega)$ be the n^{th} *bit* in the nonterminating binary expansion of ω, so $\omega = \sum_n \delta_n(\omega) 2^{-n}$. There's some ambiguity in the expansion of dyadic rationals... for example, one-half can be written either as $0.10b$ or as the infinitely repeating $0.01111111...b$. If we had used the convention that the dyadic rationals have only finitely many 1s in their expansion (so $1/2 = 0.10b$) then $\delta_n(\omega) = \lfloor 2^n \omega \rfloor \pmod 2$; with our convention ("nonterminating") that all expansions must have infinitely many ones, we have

$$\delta_n(\omega) = (\lceil 2^n \omega \rceil + 1) \pmod 2. \tag{2}$$

We can think of $\{\delta_n\}$ as an infinite sequence of *random variables*, all defined on the same measurable space $(\Omega, \mathcal{B}^1)$, with the random variable δ_1 equal to zero on $(0, \frac{1}{2}]$ and one on $(\frac{1}{2}, 1]$; δ_2 equal to zero on $(0, \frac{1}{4}] \cup (\frac{1}{2}, \frac{3}{4}]$ and one on $(\frac{1}{4}, \frac{1}{2}] \cup (\frac{3}{4}, 1]$; and, in general, δ_n equal to one on a union of 2^{n-1} intervals, each of length 2^{-n} (for a total length of $\frac{1}{2}$), and equal to zero on the complementary set, also of length $\frac{1}{2}$. For the Lebesgue probability measure P on Ω that just assigns to each event $E \in \mathcal{B}^1$ its length $\mathsf{P}(E)$, we have $\mathsf{P}[\delta_n = 0] = \mathsf{P}[\delta_n = 1] = \frac{1}{2}$ for each n, independently.

Q 1: If we had used the other convention that every binary expansion must have infinitely many zeroes (instead of ones), so e.g. $1/2 = 0.10b$, *then* what would the event $E_1 := \{\omega : \delta_1(\omega) = 1\}$ have been? How about $E_2 := \{\omega : \delta_2(\omega) = 1\}$?

The sigma field "generated by" any family of random variables $\{X_\alpha\}$ (whether countable or not) is defined to be the smallest sigma field for which each X_α is measurable, *i.e.*, the smallest one $\sigma(\mathcal{A})$ containing every

$$\mathcal{A}_\alpha = \left\{ {X_\alpha}^{-1}(B) : \quad B \in \mathcal{B}(\mathbb{R}) \right\}.$$

For each $n \in \mathbb{N}$ the $\sigma-$algebra $\mathcal{F}_n$ generated by $\delta_1, \cdots, \delta_n$ is the field $\mathcal{F}_n = \{\cup_i (a_i/2^n, b_i/2^n]\}$ consisting of all finite disjoint unions of left-open intervals with both endpoints an integer between zero and 2^n, over 2^n. Each set in $\mathcal{F}_n$ can be specified by listing which of the 2^n intervals $(\frac{i}{2^n}, \frac{i+1}{2^n}]$ $(0 \le i < 2^n)$ it contains, so there are 2^{2^n} sets in $\mathcal{F}_n$ altogether. The union $\cup \mathcal{F}_n$ consists of all finite disjoint unions of left-open intervals in Ω with dyadic rational endpoints. It is closed under taking complements but it still isn't a sigma field, since it isn't closed under *countable* unions and intersections; for example, it contains the set $E_n = \{\omega : \delta_n{=}1\}$ for each $n \in \mathbb{N}$ and finite intersections like $E_1 \cap ... \cap E_n = (1 - 2^{-n}, 1]$, but not their countable intersection $\cap_{n=1}^\infty E_n = \{1\}$. By definition the "join" $\mathcal{F} = \bigvee_n \mathcal{F}_n := \sigma(\cup_n \mathcal{F}_n)$ is the smallest sigma field that contains each $\mathcal{F}_n$ (and so contains their union); this is just the familiar Borel sets in $(0, 1]$.

Lebesgue measure P, which assigns to any interval $(a, b]$ its length, is determined on each $\mathcal{F}_n$ by the rule $\mathsf{P}\big\{ \cup_i (a_i/2^n, b_i/2^n] \big\} = \sum (b_i - a_i) 2^{-n}$ or, equivalently, by the joint distribution of the random variables $\delta_1, ..., \delta_n$: independent Bernoulli RVs, each with $\mathsf{P}[\delta_i = 1] = \frac{1}{2}$. For any number $0 < p < 1$ we can make a similar measure P_p on $(\Omega, \mathcal{F}_n)$ by requiring

$\mathsf{P}_p[\delta_n = 1] = p$ and, more generally,

$$\mathsf{P}[\delta_i = d_i,\ 1 \le i \le n] = p^{\Sigma d_i}(1{-}p)^{n-\Sigma d_i}.$$

The four intervals in $\mathcal{F}_2$ would have probabilities $[(1{-}p)^2,\ p(1{-}p),\ p(1{-}p),\ p^2]$, for example, instead of $[\frac{1}{4},\ \frac{1}{4},\ \frac{1}{4},\ \frac{1}{4}]$. This determines a measure on each $\mathcal{F}_n$, which extends uniquely to a measure P_p on $\mathcal{F} = \bigvee_n \mathcal{F}_n$. For $p = 1/2$ this is Lebesgue Measure, characterized by the property that $\mathsf{P}\{(a,b]\} = b - a$ for each $0 \le a \le b \le 1$, but the other P_ps are new. This example (the family δ_n of random variables on the spaces $(\Omega, \mathcal{F}, \mathsf{P}_p)$) is an important one, and lets us build other important examples.

Under each of these probability distributions all the δ_n are both identically distributed and independent, *i.e.*,

$$\mathsf{P}[\delta_1 \in A_1, \ldots, \delta_n \in A_n] = \prod_{i=1}^{n} \mathsf{P}[\delta_1 \in A_i].$$

Any probability assignment to intervals $(a,b] \subset \Omega$ determines *some* joint probability distribution for all the $\{\delta_n\}$, but typically the δ_n will be neither independent nor identically distributed. For any DF (*i.e.*, non-decreasing right-continuous function $F(x)$ satisfying $F(0) = 0$ and $F(1) = 1$), the prescription $\mathsf{P}_F\{(a,b]\} := F(b) - F(a)$ determines a probability distribution on every $\mathcal{F}_n$ that extends uniquely to $\mathcal{F}$, determining the joint distribution of all the $\{\delta_n\}$.

Q 2: For $F(x) = x^2$, are δ_1 and δ_2 identically distributed? Independent?
Find the marginal probability distribution for each δ_n under P_F.

Q 3: For $F(x) = \mathbf{1}_{\{x \ge 1/3\}}$, find the distribution of each δ_n under P_F.

3.4 Measurability and Observability

Fix any measure P_p on $(\Omega, \mathcal{F})$ (say, Lebesgue measure $\mathsf{P} = \mathsf{P}_{0.5}$), and define a new sequence of random variables Y_n on $(\Omega, \mathcal{F}, \mathsf{P})$ by

$$Y_n(\omega) = \sum_{i=1}^{n} (-1)^{1+\delta_n(\omega)} = \sum_{i=1}^{n} \big(2\delta_n(\omega) - 1\big),$$

the sum of n independent terms, each ± 1 with probability $1/2$ each. This is the "symmetric random walk" (it would be asymmetric with P_p for $p \ne 0.5$), starting at the origin and moving left or right with equal probability at each step. Each Y_n is $2S_n - n$ for the binomial $\mathsf{Bi}(n, 0.5)$ random variable $S_n = \sum_{i=1}^{n} \delta_i$, the partial sums of the δ_ns.

The sigma field $\mathcal{F}_n$ generated by the first n Y_is is the same as that generated by the first n S_is or that generated by the first n δ_is, the finite field $\mathcal{F}_n$ of all disjoint unions of half-open intervals with endpoints of the form $j2^{-n}$. A random variable Z on $(\Omega, \mathcal{F}, \mathsf{P})$ is $\mathcal{F}_n$-measurable if *and only if* Z can be written as a function $Z = \varphi_n(\delta_1, \ldots, \delta_n)$ of the first n δs. Thus "measurability" *means something* for us— Z is **measurable** over $\mathcal{F}_n$ if and only if you can tell its value by **observing** the first n values of δ_i (or, equivalently, of Y_i or S_i— each of these gives the same *information* $\mathcal{F}_n$). We'll see that a function Z on Ω is $\mathcal{F}$-measurable (*i.e.*, is a random variable) if and only if you can approximate it arbitrarily well by a function of the first n δ_is, as $n \to \infty$.

3.5 Uniforms, Normals, And More

From the infinite sequence of independent random bits $\{\delta_n\}$ we can construct as many random variables as we like of *any* distribution, all on the same space $(\Omega, \mathcal{F}, \mathsf{P})$, the unit interval with Lebesgue measure (length). For example, set:

$$U_1(\omega) = \sum_{i=1}^{\infty} 2^{-i}\delta_{2^i}(\omega) \qquad U_3(\omega) = \sum_{i=1}^{\infty} 2^{-i}\delta_{5^i}(\omega)$$
$$U_2(\omega) = \sum_{i=1}^{\infty} 2^{-i}\delta_{3^i}(\omega) \qquad U_4(\omega) = \sum_{i=1}^{\infty} 2^{-i}\delta_{7^i}(\omega)$$

each the sum of *different* (and therefore independent) random bits. It is easy to see that $\{U_n\}$ will be independent, uniformly distributed random variables for $n = 1, 2, 3, 4$, and that we could construct as many of them as we like using successive primes $\{2, 3, 5, 7, 11, 13, ...\}$.

Q 3: Why did I use primes in δ_{2^i}, δ_{3^i}, δ_{5^i}, δ_{7^i}? Give another choice that would have worked.

Let $F(x)$ be any DF (right-continuous, non-decreasing function on $\mathbb{R}$ with limits 0 as $x \to -\infty$ and 1 as $x \to +\infty$) and define:

$$X_1(\omega) := \inf[x \in \mathbb{R} : F(x) \geq U_1(\omega)] \qquad X_3(\omega) := \inf[x \in \mathbb{R} : F(x) \geq U_3(\omega)]$$
$$X_2(\omega) := \inf[x \in \mathbb{R} : F(x) \geq U_2(\omega)] \qquad X_4(\omega) := \inf[x \in \mathbb{R} : F(x) \geq U_4(\omega)].$$

For strictly monotone continuous $F(x)$ this is the same as $X_n := F^{-1}(U_n)$, so this is called the *inverse CDF method* of generating random varaiables with specified distributions. It's not hard to see or show (we'll do it in a week or so) that the $\{X_n\}$ are independent, each with DF $F(x) = \mathsf{P}[X_n \leq x]$. For example, we could take $X_n = \Phi^{-1}(U_n)$ to get independent random variables with the standard normal distribution or $X_n = -\log(1 - U_n)$ for the exponential distribution.

Independent normal random variables can be constructed even more efficiently via:

$$Z_1(\omega) = \cos(2\pi U_1)\sqrt{-2\log U_2} \qquad Z_3(\omega) = \cos(2\pi U_3)\sqrt{-2\log U_4}$$
$$Z_2(\omega) = \sin(2\pi U_1)\sqrt{-2\log U_2} \qquad Z_4(\omega) = \sin(2\pi U_3)\sqrt{-2\log U_4}.$$

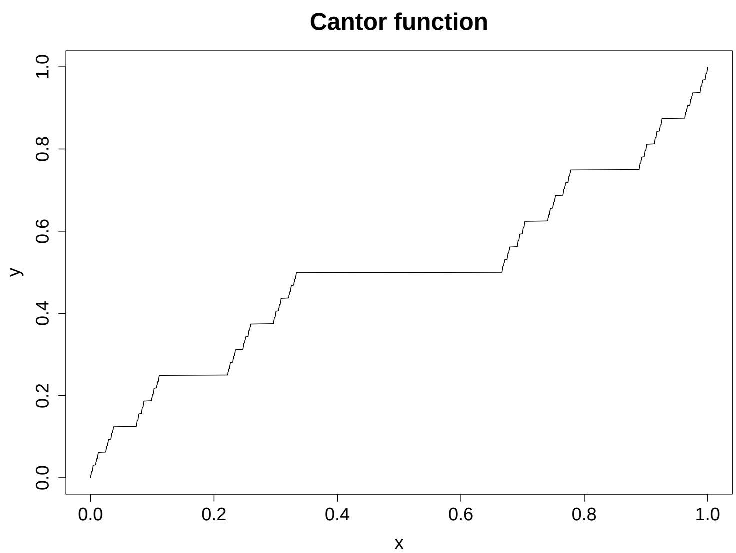
We've seen that from ordinary length (Lebesgue) measure on the unit interval (or, equivalently, from a single uniformly-distributed random variable ω) we can construct first an infinite sequence of independent 0/1 bits δ_n; then an infinite sequence of independent uniform random variables U_n; then an infinite sequence of independent normal random variables Z_n or, more generally, random variables X_n with any distribution(s) we choose.

3.5.1 The Cantor Distribution

Set $Y := \sum_{n=1}^{\infty} 2\delta_n 3^{-n}$ for the random variables $\delta_n(\omega)$ of Eqn (2). Then the ternary expansion of $y = Y(\omega)$ includes only zeroes (where $\delta_n = 0$) and twos (where $\delta_n = 1$), never ones, and so y lies in the Cantor set $C = Y(\Omega)$. Since Y takes on uncountably many different values, it cannot have a discrete distribution. Its CDF can be given analytically by the expression

$$F(y) = \sum_{n=1}^{\infty} \{2^{-n} : t_n > 0,\ t_m \neq 1,\ 1 \leq m < n\},$$

in terms of the ternary expansion $t_n := \lfloor 3^n y \rfloor \pmod 3$ of $y = \sum_{n=1}^{\infty} t_n 3^{-n}$ or graphically as



Evidently $F(x)$ is continuous, and has has derivative $F' = 0$ wherever it is differentiable, *i.e.*, outside the Cantor set. This distribution is an example of a *singular* distribution, one that is neither absolutely continuous nor discrete. We won't see many more of them.

Theorem 1 *Let $F(x)$ be any distribution function. Then there exist unique numbers $p_d \geq 0$, $p_{ac} \geq 0$, $p_{sc} \geq 0$ with $p_d + p_{ac} + p_{sc} = 1$ and distribution functions $F_d(x)$, $F_{ac}(x)$, $F_{sc}(x)$ with the properties that F_d is discrete with some probability mass function $f_d(x)$, F_{ac} is absolutely continuous with some probability density function $f_{ac}(x)$, and F_{sc} is singular continuous, satisfying $F(x) = p_d F_d(x) + p_{ac} F_{ac}(x) + p_{sc} F_{sc}(x)$ and*

$$F_d(x) = \sum_{t \leq x} f_d(t), \qquad F_{ac}(x) = \int_{t \leq x} f_{ac}(t)\, dt, \qquad F'_{sc}(x) = 0 \quad \text{where it exists.}$$

3.6 Expectation and Integral Inequalities

Discrete RVs

A random variable Y is *discrete* if it can take on only a finite or countably infinite set of distinct values $\{b_i\}$. Then (recall Section (3.1.3) on *p.* 3) Y can be represented in the form

$$Y(\omega) = \sum_i b_i \mathbf{1}_{\Lambda_i}(\omega) \tag{3}$$

as a linear combination of indicator functions of the disjoint measurable sets $\Lambda_i := X^{-1}(b_i)$. Any RV X can be approximated as well as we like by a simple RV of the form $(\star)$ by choosing $\epsilon > 0$, setting $b_i := i\epsilon$ for $i \in \mathbb{Z}$, and

$$\Lambda_i := \{\omega : b_i \le X(\omega) < b_i + \epsilon\} \qquad X_\epsilon(\omega) := \sum_{-\infty}^{\infty} b_i \mathbf{1}_{\Lambda_i}(\omega) = \epsilon \lfloor X(\omega)/\epsilon \rfloor$$

so $X - \epsilon < X_\epsilon \le X$. It is easy to define the *expectation* of such a discrete RV, or (equivalently) the *integral* of X_ϵ over $(\Omega, \mathcal{F}, \mathsf{P})$, if X is bounded below or above (to avoid indeterminate sums):

$$\mathsf{E}X_\epsilon := \int_\Omega X_\epsilon(\omega)\, \mathsf{P}(d\omega) := \int_\Omega X_\epsilon(\omega)\, d\mathsf{P}(\omega) := \int_\Omega X_\epsilon\, d\mathsf{P} := \sum_i b_i \mathsf{P}(\Lambda_i),$$

Since $X_\epsilon(\omega) \le X(\omega) < X_\epsilon(\omega) + \epsilon$, we have $\mathsf{E}X_\epsilon \le \mathsf{E}X < \mathsf{E}X_\epsilon + \epsilon$, *i.e.*,

$$\sum_i i\epsilon\, \mathsf{P}[i\epsilon \le X < (i{+}1)\epsilon] \le \mathsf{E}X < \sum_i i\epsilon\, \mathsf{P}[i\epsilon \le X < (i{+}1)\epsilon] + \epsilon. \tag{$\star\star$}$$

This determines the value of $\mathsf{E}\,X = \int_\Omega X\, d\mathsf{P}$ for each random variable X bounded above or below. If we take $\epsilon = 2^{-n}$ above, and simplify the notation by writing X_n for $X_{2^{-n}} = 2^{-n}\lfloor 2^n X \rfloor$, the sequence X_n increases monotonically to X and we can define $\mathsf{E}X := \lim_n \mathsf{E}X_n$.

Note that even for $\Omega = (0, 1]$, $\mathsf{P} = \lambda(dx)$ (Lebesgue measure), and X continuous, the value of the integral may be the same but the *passage to the limit* suggested in $(\star\star)$ is *not* the same as the limit of Riemann sums that is used to introduce integration in undergraduate calculus courses. For the Riemann sum it is the x-axis that is broken up into integral multiples of some ϵ, determining the integral of *continuous* functions, while here it is the y axis that is broken up, determining the integral of all *measurable* functions. The two definitions of integral agree for continuous functions where they are both defined, of course, but the Lebesgue integral is much more general.

If X is *not* bounded below or above, we can set $X^+ := 0 \vee X$ and $X^- := 0 \vee -X$, so that $X = X^+ - X^-$ with both X^+ and X^- bounded below (by zero), so their expectations are well-defined. If either $\mathsf{E}X^+ < \infty$ or $\mathsf{E}X^- < \infty$ we can unambiguously define $\mathsf{E}X := \mathsf{E}X^+ - \mathsf{E}X^-$, while if $\mathsf{E}X^+ = \mathsf{E}X^- = \infty$ we regard $\mathsf{E}X$ as undefined. For example, if $U \sim \mathsf{Un}(0, 1)$ then $\mathsf{E}\big[1/\sqrt{U(1-U)}\big]$ is well-defined (can you evaluate it?).

For any measurable set $\Lambda \in \mathcal{F}$ we write $\int_\Lambda X\,d\mathsf{P}$ for $\mathsf{E}X\mathbf{1}_\Lambda$. For $\Omega \subset \mathbb{R}$, if P gives positive probability to either $\{a\}$ or $\{b\}$ then the integrals over the sets (a,b), $(a,b]$, $[a,b)$, and $[a,b]$ may all be different, so the notation $\int_a^b X\,d\mathsf{P}$ isn't expressive enough to distinguish them.

Frequently in Probability and Statistics we need to calculate or estimate or find bounds for integrals and expectations. Usually this is done through limiting arguments in which a sequence of integrals is shown to converge to the one whose value we need. Here are some important properties of integrals for any measurable set $\Lambda \in \mathcal{F}$ and random variables $\{X_n\}$, X, Y, useful for bounding or estimating the integral of a random variable X. We'll prove each of these in class.

1. $\int_\Lambda X\,d\mathsf{P}$ is well-defined and finite if and only if $\int_\Lambda |X|\,d\mathsf{P} < \infty$, and $\left|\int_\Lambda X\,d\mathsf{P}\right| \le \int_\Lambda |X|\,d\mathsf{P}$. We can also define $\int_\Lambda X\,d\mathsf{P} \le \infty$ for any X bounded below by some $b > -\infty$.

2. **Lebesgue's Monotone Convergence Thm:** If $0 \le X_n \nearrow X$, then $\int_\Lambda X_n\,d\mathsf{P} \nearrow \int_\Lambda X\,d\mathsf{P} \le \infty$. In particular, the sequence of integrals converges (possibly to $+\infty$).

3. **Lebesgue's Dominated Convergence Thm:** If $X_n \to X$, and if $|X_n| \le Y$ for some RV $Y \ge 0$ with $\mathsf{E}Y < \infty$ then $\int_\Lambda |X_n - X|\,d\mathsf{P} \to 0$, $\int_\Lambda X_n\,d\mathsf{P} \to \int_\Lambda X\,d\mathsf{P}$, and $\int_\Lambda |X|\,d\mathsf{P} \le \int_\Lambda Y\,d\mathsf{P} < \infty$. In particular, the sequence of integrals converges to a finite limit, $\mathsf{E}X_n \to \mathsf{E}X$ with $|\mathsf{E}X| \le \mathsf{E}Y$.

4. **Fatou's Lemma:** If $X_n \ge 0$ on Λ, then
$$\int_\Lambda (\liminf X_n)\,d\mathsf{P} \le \liminf \Big(\int_\Lambda X_n\,d\mathsf{P}\Big).$$
The two sides may be unequal (example?), and the result is false for $\limsup$. Is "$X_n \ge 0$" necessary? Can it be weakened?

5. **Fubini's Thm:** If *either* each $X_n \ge 0$, *or* $\sum_n \int_\Lambda |X_n|\,d\mathsf{P} < \infty$, then the order of integration and summation can be exchanged: $\sum_n \int_\Lambda X_n\,d\mathsf{P} = \int_\Lambda \sum_n X_n\,d\mathsf{P}$. If both these conditions fail, the orders may not be exchangeable (example?)

6. For any $p > 0$, $\mathsf{E}|X|^p = \int_0^\infty p\,x^{p-1}\mathsf{P}[|X| > x]\,dx$ and $\mathsf{E}|X|^p < \infty \Leftrightarrow \sum_{n=1}^\infty n^{p-1}\mathsf{P}[|X| \ge n] < \infty$. The case $p = 1$ is easiest and most important: if $S := \sum_{n=1}^\infty \mathsf{P}[|X| \ge n] < \infty$, then $S \le \mathsf{E}|X| < S+1$. If X takes on only integer values, $\mathsf{E}|X| = S$.

7. If μ_X is the distribution of X, and if f is a measurable real-valued function on $\mathbb{R}$, then $\mathsf{E}f(X) := \int_\Omega f(X(\omega))\,d\mathsf{P} = \int_\mathbb{R} f(x)\,\mu_X(dx)$ if either side exists. In particular, $\mu := \mathsf{E}X = \int x\,\mu_X(dx)$ and $\sigma^2 := \mathsf{E}(X-\mu)^2 = \int (x-\mu)^2\,\mu_X(dx) = \int x^2\,\mu_X(dx) - \mu^2$.

8. **Hölder's Inequality:** Let $p > 1$ and $q = \frac{p}{p-1}$ (*e.g.*, $p = q = 2$ or $p = 1.01$, $q = 101$). Then $\mathsf{E}\,XY \le \mathsf{E}\,|XY| \le \big[\mathsf{E}|X|^p\big]^{\frac{1}{p}}\big[\mathsf{E}|Y|^q\big]^{\frac{1}{q}}$. In particular, for $p = q = 2$,
Cauchy-Schwartz Inequality: $\mathsf{E}\,XY \le \mathsf{E}\,|XY| \le \sqrt{\mathsf{E}X^2\,\mathsf{E}Y^2}$.

9. **Minkowski's Inequality:** Let $1 \le p \le \infty$ and let $X, Y \in L_p(\Omega, \mathcal{F}, \mathsf{P})$. Then

$$(\mathsf{E}|X+Y|^p)^{\frac{1}{p}} \le (\mathsf{E}|X|^p)^{\frac{1}{p}} + (\mathsf{E}|Y|^p)^{\frac{1}{p}}$$

so the norm $\|X\|_p := (\mathsf{E}|X|^p)^{\frac{1}{p}}$ obeys the triangle inequality on $L_p(\Omega, \mathcal{F}, \mathsf{P})$.
What if $0 < p < 1$?

10. **Jensen's Inequality:** Let $\varphi(x)$ be a convex function on $\mathbb{R}$, X an integrable RV. Then $\varphi\big(\mathsf{E}[X]\big) \le \mathsf{E}[\varphi(X)]$. Examples: $\varphi(x) = |x|^p$, $p \ge 1$; $\varphi(x) = e^x$; $\varphi(x) = [0 \vee x]$. The equality is *strict* if X has a non-degenerate distribution and $\varphi(\cdot)$ is strictly convex on the range of X.

11. **Markov's & Chebychev's Inequalities:** If φ is positive and increasing, then $\mathsf{P}[|X| \ge u] \le \mathsf{E}[\varphi(|X|)]/\varphi(u)$. In particular $\mathsf{P}[|X-\mu| > u] \le \frac{\sigma^2}{u^2}$ and $\mathsf{P}[|X| > u] \le \frac{\sigma^2+\mu^2}{u^2}$.

12. **One-Sided Version:** $\mathsf{P}[X > u] \le \frac{\sigma^2}{\sigma^2+(u-\mu)^2}$
(pf: $\mathsf{P}[(X-\mu+t) > (u-\mu+t)] \le ?$ for $t \in \mathbb{R}$)

13. **Hoeffding's Inequality:** If $\{X_j\}$ are real-valued, independent and essentially bounded, so $(\exists\ \{a_j, b_j\})$ s.t. $\mathsf{P}[a_j \le X_j \le b_j] = 1$, then $(\forall c > 0)$, $S_n := \sum_{j=1}^n X_j$ satisfies the bound $\mathsf{P}[S_n - \mathsf{E}S_n \ge c] \le \exp\big(-2c^2/\sum_1^n |b_j - a_j|^2\big)$. Hoeffding proved this improvement on Chebychev's inequality (at UNC) in 1963. See also related **Azuma**'s inequality (1967), **Bernstein**'s inequality (1937), and **Chernoff** bounds (1952).

    The importance of this result is that it offers an *exponentially small* (in c^2) bound for tail probabilities, while Chebychev offers only an algebraic bound on the order of $1/c^2$. Later we will find needs for the bound to be *summable* in c^2; Hoeffding's satisfies this condition, while Chebychev's does not.

Last edited: September 18, 2013