

# STA 711: Probability & Measure Theory

Robert L. Wolpert

## 3 Random Variables & Distributions

Let  $\Omega$  be any set,  $\mathcal{F}$  any  $\sigma$ -field on  $\Omega$ , and  $\mathbb{P}$  any probability measure defined for each element of  $\mathcal{F}$ ; such a triple  $(\Omega, \mathcal{F}, \mathbb{P})$  is called a *probability space*. Let  $(E, d)$  be a complete separable metric space with Borel sets  $\mathcal{E}$  (such as the real numbers  $\mathbb{R} = (-\infty, \infty)$  or Euclidean space  $\mathbb{R}^n$ ). If  $E$  isn't specified in the Definition below, then it is taken implicitly to be  $\mathbb{R}$  and  $X$  is said to be a “real-valued Random Variable”.

**Definition 1** *An  $E$ -valued Random Variable is a function  $X : \Omega \rightarrow E$  that is “ $\mathcal{F} \setminus \mathcal{E}$ -measurable”, i.e., that satisfies  $X^{-1}(B) := \{\omega : X(\omega) \in B\} \in \mathcal{F}$  for each Borel set  $B \in \mathcal{E}$ .*

This is sometimes denoted simply “ $X^{-1}(\mathcal{E}) \subset \mathcal{F}$ .” Since the probability measure  $\mathbb{P}$  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  $\mathbb{P}[X \in B]$  for each Borel set  $B$  or, for real-valued RVs, even if we want to have a well-defined distribution function (DF)  $F_X(x) := \mathbb{P}[X \leq x]$  for each  $x \in \mathbb{R}$  since the  $\pi$ -system of sets  $B$  of the form  $(-\infty, b]$  for  $b \in \mathbb{R}$  (or even just  $b \in \mathbb{Q}$ ) generates the Borel sets of  $\mathbb{R}$ .

Set-inverses are rather well-behaved functions from one class of sets to another: for any collection  $\{A_\alpha\} \subset \mathcal{E}$ , countable or not,

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

from which it follows that  $\bigcap_{\alpha} X^{-1}(A_\alpha) = X^{-1}\left(\bigcap_{\alpha} A_\alpha\right)$ . Thus, whether  $X$  is measurable or not,  $X^{-1}(\mathcal{E})$  is a  $\sigma$ -field if  $\mathcal{E}$  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  $(\mathcal{G} \setminus \mathcal{E})$ -measurable. In particular,  $X$  is  $(\mathcal{F} \setminus \mathcal{E})$ -measurable if and only if  $\sigma(X) \subset \mathcal{F}$ .

Warning: The backslash character “ $\setminus$ ” 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  $\varphi$ . 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

An  $E$ -valued random variable  $X$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  induces a measure  $\mu_X$  on  $(E, \mathcal{E})$ , called the *distribution measure* (or simply the *distribution*), via the relation

$$\mu_X(B) := \mathbb{P}[X \in B],$$

sometimes written more succinctly as  $\mu_X = \mathbb{P} \circ X^{-1}$  or even  $\mathbb{P}X^{-1}$ . Distributions of real-valued random variables are Borel measures  $\mu_X$  on the real line  $\mathbb{R}$ .

#### 3.1.1 Functions of Random Variables

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space,  $X$  an  $E$ -valued random variable, and  $g : E \rightarrow \mathbb{R}$  a (real-valued  $\mathcal{E} \setminus \mathcal{B}$ ) 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  $B \in \mathcal{B}$ . How are  $\sigma(X)$  and  $\sigma(Y)$  related?

Pretty much every function  $g : E \rightarrow \mathbb{R}$  you'll ever encounter is Borel measurable. In particular, a real-valued function  $g(x)$  is Borel measurable if it is continuous, or right-continuous, or piecewise continuous, or monotonic, or the countable limits, suprema, *etc.* of such functions.

### 3.2 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  $\{(x, y) : a < x \leq b, c < y \leq d\} = (a, b] \times (c, d]$ . 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  $\lambda$  system containing  $\mathcal{F}_0^2$  coincide, and the assignment  $\lambda_0^2((a, b] \times (c, d]) = (b - a) \times (d - c)$  of area to rectangles has a unique extension to a measure on all of  $\mathcal{B}^2$ , called two-dimensional Lebesgue measure (and denoted  $\lambda^2$ ). Of course, it's just the area of sets in the plane.

An  $\mathcal{F} \setminus \mathcal{B}^2$ -measurable mapping  $X : \Omega \rightarrow \mathbb{R}^2$  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 an  $\mathbb{R}^2$ -RV  $X$  are each RVs, and conversely that for any two random variables  $X_1$  and  $X_2$  the two-dimensional RV  $(X_1, X_2) : \Omega \rightarrow \mathbb{R}^2$  is  $\mathcal{F} \setminus \mathcal{B}^2$ -measurable, *i.e.*, is a  $\mathbb{R}^2$ -RV (how would you *prove* that?).

Also, any Borel measurable (and in particular, any piecewise-continuous) real function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  induces a random variable  $Z := f(X, Y)$ . 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  $(X_1, \dots, X_n)$  as  $\mathbb{R}^n$ -valued RVs, or  $\mathcal{F} \setminus \mathcal{B}^n$ -measurable functions, and will

use Lebesgue  $n$ -dimensional measure  $\lambda^n$  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.

Even if we have countably *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,

$$\begin{aligned} [\omega : \sup_{i \in \mathbb{N}} X_i(\omega) \leq r] &= \bigcap_{i=1}^{\infty} [\omega : X_i(\omega) \leq r] \\ [\omega : \limsup_{i \rightarrow \infty} X_i(\omega) \leq r] &= \bigcup_{j=1}^{\infty} \bigcap_{i=j}^{\infty} [\omega : X_i(\omega) \leq r] = \liminf_{i \rightarrow \infty} [\omega : X_i(\omega) \leq r], \end{aligned}$$

so  $\sup X_i$  and  $\limsup X_i$  are random variables if  $\{X_i\}$  are. 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  $\epsilon_k \rightarrow 0$ , and so is  $\mathcal{F}$ -measurable and has a well defined probability  $\mathbb{P}[\limsup_i X_i = \liminf_i X_i]$ . This is one point where countable additivity (and not just finite additivity) of  $\mathbb{P}$  is crucial, and where  $\mathcal{F}$  must be a sigma field (and not just a field).

### 3.3 Example: Discrete RVs

If an 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:

$$\begin{aligned} X(\omega) &= \sum_i b_i \mathbf{1}_{\Lambda_i}(\omega), \quad \text{where} \quad (*) \\ \mathbf{1}_{\Lambda}(\omega) &:= \begin{cases} 1 & \text{if } \omega \in \Lambda \\ 0 & \text{if } \omega \notin \Lambda \end{cases} \quad (1) \end{aligned}$$

is the so-called *indicator function* of  $\Lambda \in \mathcal{F}$ . Since  $\Omega = \cup \Lambda_i$  and  $\Lambda_i \cap \Lambda_j = \emptyset$  for  $i \neq j$ , the  $\{\Lambda_i\}$  form a “countable partition” of  $\Omega$ . Any RV can be approximated uniformly as well as we like by an 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 (can you show that?). The *distribution* of a discrete RV  $X$  is given for Borel sets  $B \subset \mathbb{R}$  by

$$\mu_X(B) = \sum \{\mathbb{P}(\Lambda_j) : b_j \in B\},$$

the probability  $\mathbb{P}[X \in B] = \mathbb{P}[\cup \{\Lambda_j : b_j \in B\}]$  that  $X$  takes a value in  $B$ .

### Arbitrary Functions of Discrete RVs

If  $Y = \phi(X)$  for *any* function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ , then  $Y$  is a random variable with the discrete distribution:

$$\mu_Y(B) = \sum \{P(\Lambda_j) : \phi(b_j) \in B\}$$

for all Borel sets  $B \in \mathcal{B}$ , the probability  $P[Y \in B] = P[\cup \{\Lambda_j : \phi(b_j) \in B\}] = \mu_X(\phi^{-1}(B))$  that  $Y$  takes a value in  $B$ .

### 3.4 Example: Absolutely Continuous RVs

If there is a nonnegative function  $f(x)$  on  $\mathbb{R}$  with unit integral  $1 = \int_{\mathbb{R}} f(x) dx$  whose definite integral gives the CDF

$$F(x) := P[X \leq x] = \int_{-\infty}^x f(t) dt$$

for  $X$ , then the distribution for  $X$  can be given on Borel sets  $B \subset \mathbb{R}$  by the integral

$$\mu_X(B) := P[X \in B] = \int_B f(x) dx = \int_{\mathbb{R}} f(x) \mathbf{1}_B(x) dx \quad (2)$$

of the pdf  $f(x)$  over the set  $B$ . This is immediate for sets of the form  $B = (-\infty, x]$ , but these form a  $\pi$ -system and so by Dynkin's extension theorem it holds for all sets  $B$  in the  $\sigma$ -field they generate, the Borel sets  $\mathcal{B}(\mathbb{R})$ . Such a distribution is called *absolutely continuous*.

#### Smooth Functions of Absolutely Continuous RVs

If  $Y = \phi(X)$  for a strictly non-decreasing differentiable function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ , and if  $X$  has pdf  $f(x)$ , then  $Y$  has a pdf  $g(y)$  too, for then with  $y = \phi(x)$  we have

$$\begin{aligned} G(y) &:= P[Y \leq y] \\ &= P[\phi(X) \leq \phi(x)] \\ &= P[X \leq x] \\ &= F(x) \\ &= \int_{-\infty}^x f(t) dt \end{aligned}$$

Upon differentiating both sides wrt  $x$ , using the chain rule for  $y = \phi(x)$ ,

$$G'(y)\phi'(x) = f(x),$$

so  $G(y)$  has a pdf  $g(y) = G'(y)$  given by

$$g(y) = f(x)/\phi'(x), \quad x = \phi^{-1}(y).$$

In this context the derivative  $\phi'(x)$  is called the *Jacobian* of the transformation  $X \rightsquigarrow Y := \phi(X)$ . Note this didn't come up in change-of-variables for discrete RVs above.

More generally, if  $\phi$  is everywhere differentiable but not necessarily monotone, with a derivative  $\phi'(x)$  that vanishes on at most countably many points, there can be at most countably many solutions  $x$  to  $\phi(x) = y$  for each  $y \in \mathbb{R}$  and  $Y = \phi(X)$  will have pdf

$$g(y) = \sum_{x \in \phi^{-1}(y)} \frac{f(x)}{|\phi'(x)|} \quad (3)$$

and the distribution of  $Y = \phi(X)$  will be given by

$$\mu_Y(B) = \int_B g(y) dy.$$

### 3.4.1 Specific Absolutely Continuous Examples

- The standard  $\text{No}(0, 1)$  Normal or Gaussian distribution is given by

$$\mu_Z(A) := \int_A f(z | 0, 1) dz, \quad f(z | 0, 1) := (2\pi)^{-1/2} e^{-z^2/2}$$

for all Borel  $A \in \mathcal{B}$ , with pdf  $f(z | 0, 1)$ .

- The Normal  $\text{No}(\mu, \sigma^2)$  distribution is that of  $Y = \phi(Z)$  for  $\phi(z) := \mu + \sigma z$  and  $Z \sim \text{No}(0, 1)$ . By (3) its pdf is

$$\begin{aligned} f(y | \mu, \sigma^2) &= \sum_{z: \phi(z)=y} \frac{f(z | 0, 1)}{|\phi'(z)|} \\ &= \sum_{z: \mu + \sigma z = y} \frac{(2\pi)^{-1/2} e^{-z^2/2}}{\sigma} \\ &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

and the  $\text{No}(\mu, \sigma^2)$  distribution is

$$\mu_Y(A) := \int_A f(y | \mu, \sigma^2) dy.$$

- The chi-squared  $\chi_1^2$  distribution with one degree of freedom is that of  $X = \phi(Z)$  for  $\phi(z) := z^2$  and  $Z \sim \text{No}(0, 1)$ . By (3) its pdf is

$$\begin{aligned} g(x) &= \sum_{z: \phi(z)=x} \frac{f(z | 0, 1)}{|\phi'(z)|} \\ &= \sum_{z: z^2=x} \frac{(2\pi)^{-1/2} e^{-z^2/2}}{|2z|} \\ &= \frac{(2\pi)^{-1/2} e^{-(+\sqrt{x})^2/2}}{|+2\sqrt{x}|} + \frac{(2\pi)^{-1/2} e^{-(-\sqrt{x})^2/2}}{|-2\sqrt{x}|} \text{ if } x > 0 \\ &= (2\pi x)^{-1/2} e^{-x/2} \mathbf{1}_{\{x>0\}}, \end{aligned}$$

the same as the  $\text{Ga}(1/2, 1/2)$ , and the  $\chi_1^2$  distribution is

$$\mu_X(A) := \int_A g(x) dx.$$

### 3.5 Example: Infinite Coin Toss

For each  $\omega \in \Omega = (0, 1]$  and integer  $n \in \mathbb{N}$  let  $\delta_n(\omega)$  be the  $n^{\text{th}}$  bit in the nonterminating binary expansion of  $\omega$ , 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  $0.01111111\dots b$ . 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

$$\delta_n(\omega) = (\lfloor 2^n \omega \rfloor + 1) \pmod{2}. \quad (4)$$

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  $\delta_1$  equal to zero on  $(0, \frac{1}{2}]$  and one on  $(\frac{1}{2}, 1]$ ;  $\delta_2$  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,  $\delta_n$  equal to one on a union of  $2^{n-1}$  left-open 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  $\mathbb{P}$  on  $\Omega$  that just assigns to each event  $E \in \mathcal{B}^1$  its length  $\mathbb{P}(E)$ , we have  $\mathbb{P}[\delta_n = 0] = \mathbb{P}[\delta_n = 1] = \frac{1}{2}$  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  $\{X_\alpha\}$  (finite, countable, or uncountable) is defined to be the smallest sigma field for which each  $X_\alpha$  is measurable, *i.e.*, the smallest sigma field  $\sigma(\mathcal{A})$  containing every set in the collection

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

For each  $n \in \mathbb{N}$  the  $\sigma$ -algebra  $\mathcal{F}_n$  on  $\Omega = (0, 1]$  generated by  $\{\delta_1, \dots, \delta_n\}$  is the field

$$\mathcal{F}_n = \{\cup_i (a_i/2^n, b_i/2^n] : 0 \leq a_i < b_i \leq 2^n; a_i, b_i \in \mathbb{N}_0\} \quad (5)$$

consisting of disjoint unions of left-open intervals in  $\Omega$  whose endpoints are integral multiples of  $2^{-n}$ . Each set in  $\mathcal{F}_n$  can be specified by listing which of the  $2^n$  intervals  $(\frac{i}{2^n}, \frac{i+1}{2^n}]$  ( $0 \leq i < 2^n$ ) 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  $\Omega$  with dyadic rational endpoints. It is a field closed under taking complements and finite unions, 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 their finite intersections like  $E_1 \cap \dots \cap E_n = (1 - 2^{-n}, 1]$ , 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 on  $(0, 1]$ .

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

$$\mathbf{P}[\delta_i = d_i, 1 \leq i \leq n] = p^{\sum d_i} (1-p)^{n-\sum d_i}.$$

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  $\mathbf{P}_p$  on  $\mathcal{F} = \bigvee_n \mathcal{F}_n$ . For  $p = 1/2$  this is Lebesgue Measure, characterized by the property that  $\mathbf{P}\{(a, b]\} = b - a$  for each  $0 \leq a \leq b \leq 1$ , but the other  $\mathbf{P}_p$ s are new. This example (the family  $\delta_n$  of random variables on the spaces  $(\Omega, \mathcal{F}, \mathbf{P}_p)$ ) is an important one, and lets us build other important examples.

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

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

Any probability assignment to intervals  $(a, b] \subset \Omega$  determines *some* joint probability distribution for all the  $\{\delta_n\}$ , but typically the  $\delta_n$  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  $\mathbf{P}_F\{(a, b]\} := F(b) - F(a)$  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  $\delta_1$  and  $\delta_2$  identically distributed? Independent? Find the marginal probability distribution for each  $\delta_n$  under  $\mathbf{P}_F$ .

**Q 3:** For  $F(x) = \mathbf{1}_{\{x \geq 1/3\}}$ , find the distribution of each  $\delta_n$  under  $\mathbf{P}_F$ .

## 3.6 Measurability and Observability

We will often consider a number of different  $\sigma$ -algebras  $\mathcal{F}_n$  on the same set  $\Omega$ — for example, those generated by families of events or random variables. In this section we'll illustrate how  $\sigma$ -fields represent *information*, a theme that will continue into our later study of conditioning.

### 3.6.1 An example: Random Walks and Bernoulli Sequences

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

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

the sum of  $n$  independent terms, each  $\pm 1$  with probability  $1/2$  each. This is the “symmetric random walk” (it would be asymmetric with  $P_p$  for  $p \neq 0.5$ ), 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  $\text{Bi}(n, 0.5)$  random variable  $S_n := \sum_{i=1}^n \delta_i$ , the partial sums of the  $\delta_n$ s.

For each fixed  $n \in \mathbb{N}$  the three sigma fields

$$\mathcal{F}_n := \sigma \{ \delta_i : 1 \leq i \leq n \} = \sigma \{ Y_i : 1 \leq i \leq n \} = \sigma \{ S_i : 1 \leq i \leq n \}$$

are all identical, and in fact coincide with the  $\sigma$ -algebra constructed in Eqn (5): all disjoint unions of half-open intervals with endpoints of the form  $j2^{-n}$ . A random variable  $Z$  on  $(\Omega, \mathcal{F}, \mathbb{P})$  is  $\mathcal{F}_n$ -measurable if *and only if*  $Z$  can be written as a function  $Z = \varphi_n(\delta_1, \dots, \delta_n)$  of the first  $n$   $\delta$ s (see subsection 3.6.2 below). 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  $\delta_i$  (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  $\Omega$  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  $n$   $\delta_i$ s, as  $n \rightarrow \infty$ .

For example, the RVs  $Y_n$  and  $S_n$  are in  $\mathcal{F}_m$  for  $m \geq n$ , but not for  $m < n$ . The RV  $Z := \min\{n : Y_n \geq 1\}$  is in  $\mathcal{F} = \sigma\{ \cup_{n \in \mathbb{N}} \mathcal{F}_n \}$ , but not in  $\mathcal{F}_n$  for any  $n \in \mathbb{N}$ .

### 3.6.2 Sub- $\sigma$ -fields

**Proposition 1** *Let  $X$  and  $Y$  be real-valued random variables on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Then  $\sigma(Y) \subset \sigma(X)$  if and only if there exists a Borel function  $g : \mathbb{R} \rightarrow \mathbb{R}$  for which  $Y = g(X)$ .*

**Proof.** First, suppose  $Y = g(X)$  for a Borel-measurable  $g : \mathbb{R} \rightarrow \mathbb{R}$ . Then for any Borel  $B \in \mathcal{B} = \mathcal{B}(\mathbb{R})$ ,  $g^{-1}(B) \in \mathcal{B}$  so

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

Thus  $\sigma(Y) \subset \sigma(X)$ .

Now suppose  $\sigma(Y) \subset \sigma(X)$ . For each  $j \in \mathbb{Z}$  and  $n \in \mathbb{N}$ , the event

$$A_j^n := \{ \omega : j2^{-n} \leq Y(\omega) < (j+1)2^{-n} \}$$

is in  $\sigma(Y) \subset \sigma(X)$ , so there is a Borel set  $B_j^n \in \mathcal{B}$  for which  $A_j^n = X^{-1}(B_j^n)$ . Since the  $\{A_j^n : j \in \mathbb{Z}\}$  are disjoint for fixed  $n \in \mathbb{N}$ , we may take the  $\{B_j^n : j \in \mathbb{Z}\}$  to be disjoint as well. Set:

$$g^n(x) := \sum_{j \in \mathbb{Z}} j2^{-n} \mathbf{1}_{\{B_j^n\}}(x)$$

and verify that

$$g^n(X) \leq Y < g^n(X) + 2^{-n}.$$

Now set  $g(x) := \limsup_{n \rightarrow \infty} g^n(x)$  and verify that  $Y = g(X)$ . □



### 3.7 Selecting a Probability Space $\Omega$

Let  $\mu$  be a specified probability distribution on some metric space  $E$ , *i.e.*, a probability measure on the Borel sets  $\mathcal{E}$  of  $E$  (for example,  $E$  might be  $\mathbb{R}$  or  $\mathbb{R}^N$ ). How can we construct a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and random element  $X : \Omega \rightarrow E$  with distribution  $\mu$ ?

If  $\mu$  is a discrete measure with finite support, *i.e.*, if  $\mu(S) = 1$  for some finite set  $S = \{x_1, \dots, x_n\} \subset E$ , then one possibility is to let  $\Omega = \{\omega_1, \dots, \omega_n\}$  be *any* finite set with  $n$  elements and set  $\mathcal{F} := 2^\Omega$ ,  $p_i := \mu(\{x_i\})$ ,  $X(\omega_i) := x_i$ , and set

$$\mathbb{P}[A] := \sum_{\omega_i \in A} p_i = \sum_i p_i \mathbf{1}_A(\omega_i).$$

For example, to model the outcome of two distinguishable dice (not necessarily fair ones) we could use any set  $\Omega$  with (at least) 36 distinct elements (for indistinguishable dice we would need only 21 distinct elements; if only the sum is of interest then 11 elements would do). Similarly, if  $\mu$  is any discrete measure then we could construct a suitable model with  $\Omega = \mathbb{N}$  and  $\Omega = 2^\Omega$  by enumerating the support points  $x_n$  of  $\mu$  and setting  $X(n) := x_n$ ,  $\mathbb{P}[A] := \sum \{\mu(\{x_n\}) : n \in A\}$ .

But these aren't the only choices. If  $\mu$  is discrete with a finite number  $n$  of support points, then *any* set  $\Omega$  with  $n$  or more points can serve. Or, we could construct a random variable  $X$  with any distribution at all, on the unit interval  $\Omega = (0, 1]$  with the Borel sets  $\mathcal{F} = \mathcal{B}$  and Lebesgue measure  $\mathbb{P}$  (we do this in Section (3.7.2) below). In any particular problem we are free to choose a space  $(\Omega, \mathcal{F}, \mathbb{P})$  that makes our calculations as clear and simple as possible.

#### 3.7.1 The Canonical Space

One space that will *always* work is to select  $\Omega = E$  itself, with its Borel sets  $\mathcal{F} = \mathcal{E}$ , with  $\mathbb{P} = \mu$  and  $X(\omega) = \omega$ . This is called the “canonical space”. For example, a (real-valued) Random Variable  $X$  can be constructed with any distribution  $\mu$  on  $(\mathbb{R}, \mathcal{B})$  by setting

$$\Omega = \mathbb{R} \quad \mathcal{F} = \mathcal{B} \quad \mathbb{P} = \mu \quad X(\omega) = \omega.$$

#### 3.7.2 The Inverse CDF Method

We can build real-valued random variables with any specified distribution on the unit interval with Lebesgue measure, as follows. Let  $(\Omega, \mathcal{F}, \mathbb{P}) = ((0, 1], \mathcal{B}, \mathbb{P})$  be the unit interval with the Borel sets and Lebesgue measure, and let  $F(x)$  be any DF— non-decreasing, right-continuous function on  $\mathbb{R}$  with limits  $F(-\infty) = 0$  and  $F(\infty) = 1$ . Define a real-valued<sup>1</sup> random variable  $X$  on  $(\Omega, \mathcal{F}, \mathbb{P})$  by

$$X(\omega) = F^{\leftarrow}(\omega) := \inf\{x \in \mathbb{R} : F(x) \geq \omega\}$$

<sup>1</sup>If the support of  $\mu$  is unbounded, *i.e.*, if  $F(x) < 1$  for all  $x \in \mathbb{R}$ , this could be *extended* real-valued since  $X(1)$  would be infinite. Simply set  $X(1) = 0$  (say) and use the given expression for  $\omega \in (0, 1)$  to construct a (finite) real-valued random variable with the same distribution.

Then  $X$  is a random variable on  $(\Omega, \mathcal{F}, \mathbf{P})$  with DF  $F$ , because for any  $x \in \mathbb{R}$

$$\{\omega : X(\omega) \leq x\} = (0, F(x)]$$

whose Lebesgue measure is  $F(x)$ . For continuous and strictly monotone DFs,  $F^{\leftarrow}(\omega)$  coincides with the inverse  $F^{-1}(\omega)$ , so this is called the *inverse CDF method* of generating random variables with specified distributions— but the method still works even if  $F$  isn't continuous or strictly monotone. For some examples, we could take  $X = \Phi^{-1}(\omega)$  to get a  $\text{No}(0, 1)$  RV or  $X = -\log(1 - \omega)$  for one with the unit exponential distribution or  $X = \mathbf{1}_{\{\omega > 1-p\}}$  for the Bernoulli  $\text{Bi}(1, p)$  distribution.

### 3.7.3 Uniforms, Normals, And More

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

$$\begin{aligned} U_1(\omega) &:= \sum_{i=1}^{\infty} 2^{-i} \delta_{2^i}(\omega) & U_3(\omega) &:= \sum_{i=1}^{\infty} 2^{-i} \delta_{5^i}(\omega) \\ U_2(\omega) &:= \sum_{i=1}^{\infty} 2^{-i} \delta_{3^i}(\omega) & U_4(\omega) &:= \sum_{i=1}^{\infty} 2^{-i} \delta_{7^i}(\omega) \end{aligned}$$

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  $\{2, 3, 5, 7, 11, 13, \dots\}$ .

**Q 4:** Why did I use primes in  $\delta_{2^i}$ ,  $\delta_{3^i}$ ,  $\delta_{5^i}$ ,  $\delta_{7^i}$ ? Give another choice that would work.

Using the Inverse CDF method, for any DF  $F(x)$  we can construct independent random variables  $X_n(\omega) = F^{\leftarrow}(U_n) := \inf\{x \in \mathbb{R} : F(x) \geq U_n(\omega)\}$ , each with DF  $F(x) = \mathbf{P}[X_n \leq x]$ ; or, if we have any sequence  $\{F_n\}$  of DFs, we could construct independent random variables  $X_n(\omega) = F_n^{\leftarrow}(U_n)$  with arbitrary specified distributions, all on the same probability space  $(\Omega, \mathcal{F}, \mathbf{P}) = ((0, 1], \mathcal{B}, \mathbf{P})$ . 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 unit exponentially-distributed random variables.

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

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

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

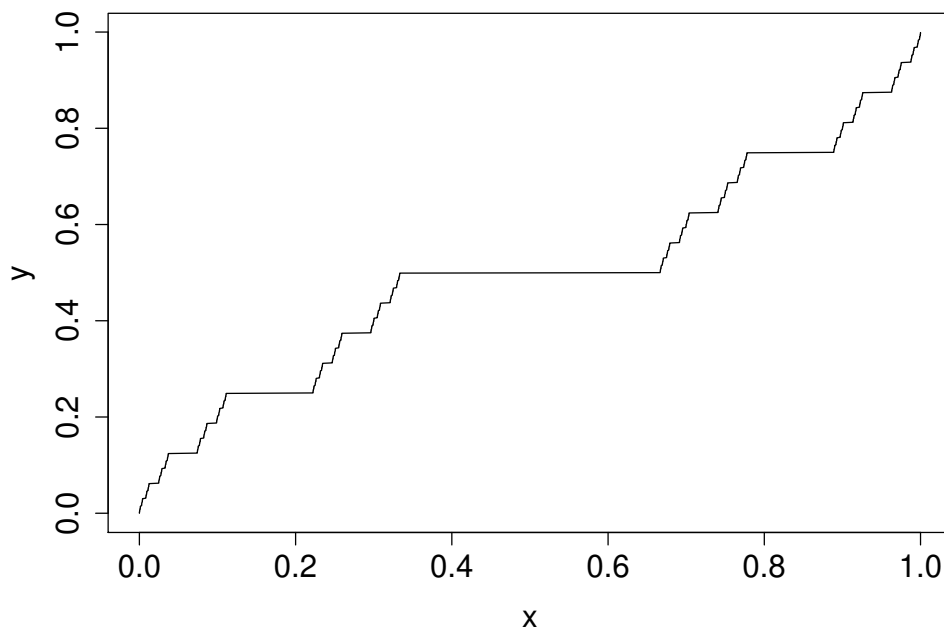
### 3.7.4 The Cantor Distribution

Set  $Y := \sum_{n=1}^{\infty} 2\delta_n 3^{-n}$  for the random variables  $\delta_n(\omega)$  of Eqn (4). 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  $t_n := [3^n y] \pmod{3}$  of  $y = \sum_{n=1}^{\infty} t_n 3^{-n}$  or graphically as

#### Cantor function



Evidently  $F(x)$  is continuous, and has derivative  $F' = 0$  wherever it is differentiable, *i.e.*, outside the Cantor set. This distribution is an example of a *singular continuous* distribution, one that has no absolutely continuous or discrete part. 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*

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

Proof. Easy—pick off the jumps of  $F(x)$  first (at most countably many, by a HW problem), to build  $F_d$  and find  $p_d$ ; then pick off the pdf proportional to  $F'$ , where that exists, to find  $F_{ac}$  and  $p_{ac}$ ; and build  $F_{sc}$  and find  $p_{sc}$  from whatever's left.  $\square$

### 3.8 Expectation and Integral Inequalities

This section is just a peek ahead at material presented in more detail in the lecture notes for Week 4.

#### 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 (3.3) on  $p.3$ )  $Y$  can be represented in the form

$$Y(\omega) = \sum_i b_i \mathbf{1}_{\Lambda_i}(\omega) \quad (6)$$

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 (6) by choosing  $\epsilon > 0$ , setting  $b_i := i\epsilon$  for  $i \in \mathbb{Z}$ , and

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

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

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

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

$$\sum_i i\epsilon \mathbf{P}[i\epsilon \leq X < (i+1)\epsilon] \leq \mathbf{E}X < \sum_i i\epsilon \mathbf{P}[i\epsilon \leq X < (i+1)\epsilon] + \epsilon. \quad (7)$$

This determines the value of  $\mathbf{E}X = \int_{\Omega} X d\mathbf{P}$  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  $\mathbf{E}X := \lim_n \mathbf{E}X_n$ .

Note that even for  $\Omega = (0, 1]$ ,  $\mathbf{P} = \lambda(dx)$  (Lebesgue measure), and  $X$  continuous, the value of the integral may be the same but the *passage to the limit* suggested in (7) 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  $\epsilon$ , 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  $\mathbf{E}X^+ < \infty$  or  $\mathbf{E}X^- < \infty$  we can unambiguously define  $\mathbf{E}X := \mathbf{E}X^+ - \mathbf{E}X^-$ , while if  $\mathbf{E}X^+ = \mathbf{E}X^- = \infty$  we regard  $\mathbf{E}X$  as undefined. For example, if  $U \sim \text{Un}(0, 1)$  then  $\mathbf{E}[1/\sqrt{U(1-U)}]$  and  $\mathbf{E}[1/(U(1-U))]$  are well-defined (can you evaluate them?), but  $\mathbf{E}[1/(1-2U)]$  is not.

For any event  $\Lambda \in \mathcal{F}$  we may write  $\int_{\Lambda} X d\mathbf{P}$  for  $\mathbf{E}X\mathbf{1}_{\Lambda}$ . For  $\Omega \subset \mathbb{R}$ , if  $\mathbf{P}$  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  $\int_a^b X d\mathbf{P}$  isn't expressive enough to distinguish them. Instead we write  $\int_{(a,b)} X d\mathbf{P}$ ,  $\int_{(a,b]} X d\mathbf{P}$ , *etc.* or, equivalently,  $\int \mathbf{1}_{(a,b)} X d\mathbf{P}$ ,  $\int \mathbf{1}_{(a,b]} X d\mathbf{P}$ , *etc.*

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.  $\int_{\Lambda} X d\mathbf{P}$  is well-defined and finite if and only if  $\int_{\Lambda} |X| d\mathbf{P} < \infty$ , and  $\left| \int_{\Lambda} X d\mathbf{P} \right| \leq \int_{\Lambda} |X| d\mathbf{P}$ . We can also define  $\int_{\Lambda} X d\mathbf{P} \leq \infty$  for any  $X$  bounded below by some  $b > -\infty$ .
2. **Lebesgue's Monotone Convergence Thm:** If  $0 \leq X_n \nearrow X$ , then  $\int_{\Lambda} X_n d\mathbf{P} \nearrow \int_{\Lambda} X d\mathbf{P} \leq \infty$ . In particular, the sequence of integrals converges (possibly to  $+\infty$ ).
3. **Lebesgue's Dominated Convergence Thm:** If  $X_n \rightarrow X$ , and if  $|X_n| \leq Y$  for some RV  $Y \geq 0$  with  $\mathbf{E}Y < \infty$  then  $\int_{\Lambda} |X_n - X| d\mathbf{P} \rightarrow 0$ ,  $\int_{\Lambda} X_n d\mathbf{P} \rightarrow \int_{\Lambda} X d\mathbf{P}$ , and  $\int_{\Lambda} |X| d\mathbf{P} \leq \int_{\Lambda} Y d\mathbf{P} < \infty$ . In particular, the sequence of integrals converges to a finite limit,  $\mathbf{E}X_n \rightarrow \mathbf{E}X$  with  $|\mathbf{E}X| \leq \mathbf{E}Y$ .
4. **Fatou's Lemma:** If  $X_n \geq 0$  on  $\Lambda$ , then

$$\int_{\Lambda} (\liminf X_n) d\mathbf{P} \leq \liminf \left( \int_{\Lambda} X_n d\mathbf{P} \right).$$

The two sides may be unequal (example?), and the inequality fails for  $\limsup$ . Is " $X_n \geq 0$ " necessary? Can it be weakened?

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

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

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

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  $E XY \leq E |XY| \leq [E|X|^p]^{\frac{1}{p}} [E|Y|^q]^{\frac{1}{q}}$ . More generally, if  $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$  for  $p, q, r \in [1, \infty]$ , then  $\|XY\|_r \leq \|X\|_p \|Y\|_q$ . In particular, for  $p = q = 2$  and  $r = 1$ ,

**Cauchy-Schwartz Inequality:**  $E XY \leq E |XY| \leq \sqrt{EX^2 EY^2}$ .

9. **Minkowski's Inequality:** Let  $1 \leq p \leq \infty$  and let  $X, Y \in L_p(\Omega, \mathcal{F}, P) := \{Z : E|Z|^p < \infty\}$ . Then

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

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

What if  $0 < p < 1$ ?

10. **Jensen's Inequality:** Let  $\varphi(x)$  be a convex function on  $\mathbb{R}$ ,  $X$  an integrable RV. Then  $\varphi(E[X]) \leq E[\varphi(X)]$ . Examples:  $\varphi(x) = |x|^p$ ,  $p \geq 1$ ;  $\varphi(x) = e^x$ ;  $\varphi(x) = [0 \vee x]$ . 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  $\varphi$  is positive and increasing, then  $P[|X| \geq u] \leq E[\varphi(|X|)]/\varphi(u)$ . In particular  $P[|X - \mu| > u] \leq \frac{\sigma^2}{u^2}$  and  $P[|X| > u] \leq \frac{\sigma^2 + \mu^2}{u^2}$ .

12. **One-Sided Version:**  $P[X > u] \leq \frac{\sigma^2}{\sigma^2 + (u - \mu)^2}$   
(pf:  $P[(X - \mu + t) > (u - \mu + t)] \leq ?$  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.  $P[a_j \leq X_j \leq b_j] = 1$ , then  $(\forall c > 0)$ ,  $S_n := \sum_{j=1}^n X_j$  satisfies the bound  $P[S_n - ES_n \geq c] \leq \exp(-2c^2 / \sum_1^n |b_j - a_j|^2)$ . 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.