CALIFORNIA INSTITUTE OF TECHNOLOGY
Division of the Humanities and Social Sciences

A Few Things You Should Know About Probability*

KCB

Revised July 1998, October 2001

*First typed draft of notes from 1983. This will almost surely be updated.

Contents

Lecture 1

Random Variables and Expectation

1.1 Metaphysical nonsense

The modern way to think about uncertainty is to take the view that there are several conceivable outcomes to any uncertain phenomenon. We name these outcomes "states of the world" or "states of the nature," though some philosophers prefer the term "possible worlds." The canonical way to create a mathematical object to represent the set of states of nature is take an appropriate language, and identify consistent sets of sentences with the state of nature or possible world in which these sentences are true. (We won't do that here, but if you like that sort of thing see, e.g., Chellas [5, Chapter 1].) The important thing about states is that they are assumed to be mutually exclusive and exhaustive of all the possibilities. (This last assumption has to be suspect, since the set of states of nature is a modeling construction.)

We know some things about the state of nature, but perhaps not enough to know the state of nature exactly. The sorts of things we can know about are called "events," which are just sets of states that are consistent with our knowledge. The fact that we describe states of the world as a collection of sentences means that by using the conjunctions "or" and "and" and the modifier "not" means that the set of events ought to possess a certain minimal set theoretic structure known as a *field* or *algebra* of sets. These properties essentially require closure of the collection of events under pairwise unions and intersections and under complementation.

We express our degree of belief about the likelihoods of events in terms of a probability measure, which is a kind of set function defined on the field

of events. It assigns values between 0 and 1, although certain Bayesian statisticians entertain probability measures assuming values between 0 and ∞ inclusive. We shan't do that here. The main property of a probability measure is *additivity*: the probability of disjoint events is the sum of their probabilities.

For technical reasons beyond the scope of these notes, probabilists prefer to work with a little more structure. The collection of events is assumed to be a σ-field (closed under unions and intersections of infinite sequences of events) and the probability measure is assumed to be countably additive (the probability of the union of a sequence of pairwise disjoint events is the infinite series of their probabilities).

Why you may ask, do we take this approach? There are several physical phenomena such as coin tossing, V2 hits on London, diffusion of ink in water, etc., that seem to be well described by this approach. There are actually deep physical and mathematical reasons why something as theoretically deterministic as coin tossing, which is after all governed by Newton's laws, should be best described probabilistically. For an introduction to this, see Lasota and Mackey [12]. There are reasonable, if not compelling, reasons why purely subjective beliefs may well be described by probabilities. The classic works here are of course de Finetti [9] and Savage [18].

1.2 Probability Spaces

A **probability space** is a triple $(\Omega, \mathcal{E}, P)$ where Ω is a nonempty set of "states of nature," $\mathcal{E}$ is a collection of subsets of Ω called "events," and P is a probability measure.

Elaborate on states of the world. The σ-field of events represents what is observable. E.g., two identical dice $\{(1,6),(6,1)\}$ can not be distinguished. Voting returns from secret ballots.

We assume that $\mathcal{E}$ is a $\boldsymbol{\sigma}$**-field** (also known as a $\boldsymbol{\sigma}$**-algebra**). That is, it satisfies:

1. $\varnothing \in \mathcal{E}$, $\Omega \in \mathcal{E}$.
2. If $E_i, i = 1, 2, 3, \ldots$ belong to $\mathcal{E}$, then $\cap_{i=1}^{\infty} E_i$ and $\cup_{i=1}^{\infty} E_i$ belong to $\mathcal{E}$.
3. If $E \in \mathcal{E}$, then $E^c \in \mathcal{E}$.

A set function $P: \mathcal{E} \to [0,1]$ is a **probability measure** if:

1. $P(\varnothing) = 0$; and $P(\Omega) = 1$
2. $P(\cup_{i=1}^{\infty} E_i) = \Sigma_{i=1}^{\infty} P(E_i)$ provided $E_i \cap E_j = \varnothing$ for $i \neq j$.

To cut down on the number of delimiters, when a set is delimited with braces, we may omit the parentheses surrounding it and simply write something like $P\{\omega : f(\omega) = 1\}$ instead of $P(\{\omega : f(\omega) = 1\})$. You will appreciate this later.

Elementary consequences:

1. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$, even if $E \cap F \neq \varnothing$.
2. $E_n \downarrow$ implies $P(\cap_n E_n) = \lim_n P(E_n)$.
3. $E_n \uparrow$ implies $P(\cup_n E_n) = \lim_n P(E_n)$.

1.3 Lebesgue measure

Far and away the most important example of a probability space is the unit interval of the real line under Lebesgue measure, $([0,1], \mathcal{B}, \lambda)$. In fact, probability spaces are called *standard* if they look like a subset of this one.

First we have to describe the events. The smallest σ-field of subsets of $\boldsymbol{R}$ containing all the intervals is called the **Borel σ-field** of $\boldsymbol{R}$ and is denoted $\mathcal{B}$. More generally the smallest σ-field of subsets of $\boldsymbol{R}^n$ containing all the open sets is called the **Borel σ-field** of $\boldsymbol{R}^n$ and is denoted $\mathcal{B}^n$. Sets belonging to $\mathcal{B}$ or $\mathcal{B}^n$ are called Borel sets. The Borel subsets of the unit interval $[0,1]$ comprise a σ-field that we also denote $\mathcal{B}$. These sets are our events.

Our probability is **Lebesgue measure** λ. It has the property that for any interval $[a,b]$ with $a \leqslant b$, its probability is its length, $\lambda([a,b]) = b - a$. This is sometimes called the **uniform** probability distribution. It turns out that there is only measure measure on $\mathcal{B}$ with this property.

The curious among you may ask, "Are all subsets of the interval Borel sets?" The answer is no. If you are a diligent student you will immediately demand, "Name one that isn't." The problem is that until you take a serious course in real analysis, you will never be able to think of a set that is not a Borel set. Really, you have to trust me (although I do have some good notes on the subject). Then you might ask, "Why don't we take the collection of all subsets to be our σ-field of events?" The answer is that there is no countably additive way to extend Lebesgue measure to all subsets of the unit interval.[1] Finally, seeking a way around this madness you might ask, "Why don't we just use additivity instead of countable additivity?" Indeed some brave souls have taken this approach, notably Dubins and Savage [6], but they encounter problems related to conditional expectations and uniqueness of infinite product measures that are not obviously easier to deal with than restricting our attention to Borel sets.

The long and the short of it is that you have to learn to do things this way before you can hope to understand the alternatives, and besides I had to learn it, so you should as well.

1.4 Random Variables and Vectors

A **measurable space** is a probability space without the probability measure. That is, it is a pair $(\Omega, \mathcal{E})$, where Ω is a nonempty set and $\mathcal{E}$ is σ-field of subsets of Ω. Given two measurable spaces $(\Omega, \mathcal{E})$ and $(\Lambda, \mathcal{F})$, a function $\varphi \colon \Omega \to \Lambda$ is **measurable**, more specifically **$(\mathcal{E}, \mathcal{F})$-measurable**, if for every set F belonging to $\mathcal{F}$, its inverse image $\varphi^{-1}[F]$ belongs to $\mathcal{E}$.

Here is a fact: A function φ from the measurable space $(\Omega, \mathcal{E})$ into the measurable space $(\boldsymbol{R}, \mathcal{B})$ is $(\mathcal{E}, \mathcal{B})$-measurable if for every open subset U of $\boldsymbol{R}$, its inverse image $\varphi^{-1}[F]$ belongs to $\mathcal{E}$. That is we don't have to check the inverse image of every Borel set. In fact, it is sufficient to check only the inverse images of intervals of the form $(-\infty, a)$. It is also sufficient to check only the inverse images of intervals of the form $(-\infty, a]$. It is a useful exercise to convince yourself of this. Actually, we always assume that $\boldsymbol{R}$ is a measurable space with its Borel sets, and we often say that φ is just plain $\mathcal{E}$-measurable.

A function $g \colon \boldsymbol{R}^n \to \boldsymbol{R}^m$ is **Borel-measurable** if $g^{-1}[u] \in \mathcal{B}^n$ for any open $U \subset \boldsymbol{R}^m$. It is not too hard to show that every continuous function from $\boldsymbol{R}^n$ into $\boldsymbol{R}^m$ is Borel measurable. (Try to prove it.)

A **real random variable** on $(\Omega, \mathcal{E})$ is a measurable function from the measurable space $(\Omega, \mathcal{E})$ into the measurable space $(\boldsymbol{R}, \mathcal{B})$. For reasons I don't understand, probabilists and statisticians like to use uppercase Latin letters near the end of the alphabet to denote random variables, that is, functions. This has confused generations of students, who have trouble thinking of random variables as functions. For the sake of tradition, and so that you get used to it, we follow suit. Thus a random variable X is a function $X: \Omega \to \boldsymbol{R}$ that is **$\mathcal{E}$-measurable**, that is, $X^{-1}[U] \in \mathcal{E}$ for every open set $U \subset \boldsymbol{R}$. A **real random vector** is a function $X: \Omega \to \boldsymbol{R}^n$ that is $\mathcal{E}$-measurable, that is, $X^{-1}[U] \in \mathcal{E}$ for every open set $U \subset \boldsymbol{R}^n$.

The smallest σ-field $\mathcal{F}$ such that X is $\mathcal{F}$-measurable is called the **σ-field generated by** X, and is denoted $\mathcal{F}(X)$. Clearly $\mathcal{F}(X) \subset \mathcal{E}$ whenever X is a random vector.

1 Definition (Statisticians' Notation) *Statisticians often use the notation*

$$[X \in E] \quad \textit{to mean} \quad \{\omega \in \Omega : X(\omega) \in E\} = X^{-1}[E].$$

Likewise $[X \leqslant t]$ *means* $\{\omega \in \Omega : X(\omega) \leqslant t\}$, *etc. We may also omit parentheses when referring to the probability of such an event, like so:* $P[X \leqslant t]$.

2 Definition *A property* $Q(\omega)$ *parameterized by states of the world, is said to hold* ***almost surely****, abbreviated* Q *a.s. if the set of states of the world for which it fails to hold is a subset of an event of probability zero. That is,*

$$Q \text{ a.s.} \iff \exists E \in \mathcal{E}\ P(E) = 0 \;\&\; \{\omega \in \Omega : \neg Q(\omega)\} \subset E.$$

Facts About Random Vectors

Here are some basic facts about random vectors and random variables. Remember the hard part is showing measurability. Let $X, Y : (\Omega, \mathcal{E}) \to \boldsymbol{R}^n$ be random vectors. Then:

1. $\alpha X + \beta Y$ is a random vector, where $(\alpha X + \beta Y)(\omega) = \alpha X(\omega) + \beta Y(\omega)$. That is, the collection of random vectors is a vector space.

2. $X \cdot Y$ is a random variable, where $\cdot$ is the standard dot product of vectors.

3. $X \vee Y$ and $X \wedge Y$ are random vectors.

   If X_n is a sequence of random variables, then

4. $\sup_n X_n$ and $\inf X_n$ are random variables.

5. $\limsup_n X_n$, $\liminf_n X_n$, $\lim X_n$ are random variables.

6. $X^{-1}[I] \in \mathcal{E}$ for any interval I (open, closed, half-open, degenerate).

7. If E belongs to $\mathcal{E}$, then its **indicator function** $\mathbf{1}_E$ is a random variable, where

$$\mathbf{1}_E(\omega) = \begin{cases} 0 & \omega \notin E \\ 1 & \omega \in E. \end{cases}$$

8. If $X_1, \cdots, X_n$ are random variables, then

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

is a random vector, and conversely.

1.5 Integration of Random Variables

A **simple random variable** is one that takes on only finitely many values. That is, it is of the form $\sum_{i=1}^{n} \alpha_i \mathbf{1}_{E_i}$, where each E_i belongs to $\mathcal{E}$.

3 Lemma *If X is a nonnegative random variable, then*

$$X(\omega) = \sup\{Y(\omega) : Y \text{ is a simple random variable and } Y \leqslant X\}.$$

Proof: Set $A_{n,k} = \{\omega : \frac{k}{n} \leqslant X(\omega) < \frac{k+1}{n}\}$. Observe that X is the supremum of the sequence of simple functions $X_n = \sum_{k=0}^{n^2} \frac{k}{n} \mathbf{1}_{A_{n,k}}$. ∎

4 Definition *Let $\sum_{i=1}^{n} \alpha_i \mathbf{1}_{E_i}$ be a simple random variable and define its* ***integral*** *or* ***expectation*** *by*

$$\int_\Omega \sum_{i=1}^{n} \alpha_i \mathbf{1}_{E_i}\, dP = \sum_{i=1}^{n} \alpha_i P(E_i)$$

It is a fact that $\int_\Omega X\, dP$ is well defined. That is, even though we can write a simple random variable in several ways (e.g., $\mathbf{1}_\Omega = \mathbf{1}_A + \mathbf{1}_{A^c}$) the integral always turns out the same.

5 Definition *If X is a nonnegative random variable, define its integral by*

$$\int_\Omega X\, dP = \sup \left\{ \int_\Omega Y\, dP : Y \leqslant X \text{ and } Y \text{ is a simple random variable} \right\}$$

We say that a nonnegative random variable X is **integrable** if $\int_\Omega |X|\, dP < \infty$.

For more general random variables X, set $X^+ = X \vee 0$ and $X^- = (-X \vee 0)$. Note that X^+, X^- are nonnegative random variables by a previous fact. Unless neither $\int_\Omega X^+\, dP$ nor $\int_\Omega X^-\, dP$ is integrable define

$$\int_\Omega X\, dP = \int_\Omega X^+\, dP - \int_\Omega X^-\, dP.$$

We say that X itself is **integrable** if both X^+, X^- are integrable. We can define integration over measurable subsets of Ω as follows.

$$\int_E X\, dP = \int_E X \cdot \mathbf{1}_E\, dP$$

Notations for an integral

The following notations all mean the same thing:

$$\int_\Omega X\, dP, \quad \int_\Omega X(\omega)\, dP(\omega), \quad \int_\Omega X(\omega)\, P(d\omega), \quad \int X\, dP.$$

1.6 Facts about Integrals

1. $\int_E \alpha X + \beta Y\, dP = \alpha \int_E X\, dP + \beta \int_E Y\, dP$.
2. $X \geqslant Y$ implies $\int_E X\, dP \geqslant \int_E Y\, dP$.
3. $X \geqslant 0$ and $\int_\Omega X\, dP = 0$ imply $X = 0$ a.s.
4. $P(E \cap F) = 0$ implies $\int_{E \cup F} X\, dP = \int_E X\, dP + \int_F X\, dP$.

Define Riemann integral?

1.7 Interchanging Integration and Limits

If X_n is a sequence of integrable random variables and X_n converges pointwise to an integrable random variable X, does it follow that $\int X_n\, dP$ converges to $\int X\, dP$? The answer is unfortunately *not necessarily*.

6 Example (You can't always interchange limits and integration) In this example $\Omega = [0, 1]$, $\mathcal{E} = \mathcal{B}$, and P is Lebesgue measure. Define the sequence X_n by

$$X_n(\omega) = \begin{cases} 2n & \frac{1}{2n} \leqslant \omega \leqslant \frac{1}{n} \\ 0 & \text{otherwise.} \end{cases}$$

Then each X_n is an integrable random variable with $\int X_n\, dP = 1$, but X_n converges pointwise to $X = 0$, which has $\int X\, dP = 0$. □

There are however some useful sufficient conditions for being able to interchange limits and integration. They are given in the next two theorems.

7 Monotone Convergence Theorem *If $X_n \uparrow X$, all X_n nonnegative, then* $\lim_n \int_E X_n\,dP = \int_E X\,dP$

8 Lebesgue's Dominated Convergence Theorem *If Y is integrable and dominates each X_n, that is, $|X_n| \leqslant Y$, and if $X_n \to X$ pointwise, then* $\lim_n \int_E X_n\,dP = \int_E X\,dP$.

The next result is instrumental in proving the above and occasionally is useful in its own right.

9 Fatou's Lemma $X_n \geqslant 0, X_n \to X$ *implies* $\int_E X\,dP \leqslant \liminf \int_E X_n\,dP$.

Add a section on differentiating under an integral.

Lecture 2

Distributions and densities

2.1 The Distribution of a Random Variable

Let φ be a measurable function from the probability space $(\Omega, \mathcal{E}, P)$ into the measurable space $(\Lambda, \mathcal{F})$. Then φ induces a probability measure Q on $(\Lambda, \mathcal{F})$ by the formula

$$Q(F) = P(\varphi^{-1}[F])$$

for every F in $\mathcal{F}$. This measure Q is often called $P\varphi^{-1}$. For random vectors we have a special terminology. The **distribution** of the random vector $X \colon \Omega \to \boldsymbol{R}^n$ is the probability measure P_X defined on $(\boldsymbol{R}^n, \mathcal{B}^n)$ by

$$P_X(E) = P\{\omega \in \Omega : X(\omega) \in E\} = P[X \in E] = P(X^{-1}[E]).$$

10 Definition *For a random variable $X \colon \Omega \to \boldsymbol{R}^1$, the* ***cumulative distribution function*** *F_X of X is a function $F_X \colon \boldsymbol{R} \to [0,1]$ defined by*

$$F_X(t) = P_X((-\infty, t]) = P\{\omega \in \Omega : X(\omega) \leqslant t\}.$$

N.B. *Rao [16] and Breiman [4] and the French define the cumulative distribution function using the strict inequality $<$ rather than $\leqslant$.*

11 Fact *The cumulative distribution function F_X is nondecreasing, right continuous function, and satisfies $\lim_{t\to-\infty} F_X(t) = 0$ and $\lim_{t\to\infty} F_X(t) = 1$.*

Constructing a random variable with a given distribution

The converse of Fact 11 is true.

12 Proposition *Given any function $F \colon \boldsymbol{R} \to [0,1]$ that is nondecreasing, right continuous, and satisfies $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$, there is a random variable X on the standard probability space $([0,1], \mathcal{B}, \lambda)$ with $F = F_X$.*

Proof: Given such an F, define $X\colon [0,1] \to \boldsymbol{R}$ by

$$X(t) = \inf\{x \in \boldsymbol{R} : F(x) \geqslant t\}.$$

(This makes $X(0) = -\infty$, but that's okay since $\lambda\{0\} = 0$.) When F is strictly increasing and maps onto $[0,1]$, then X is just F^{-1}. More generally, flat spots in F correspond to jumps in X and vice-versa. See Figure 2.1.

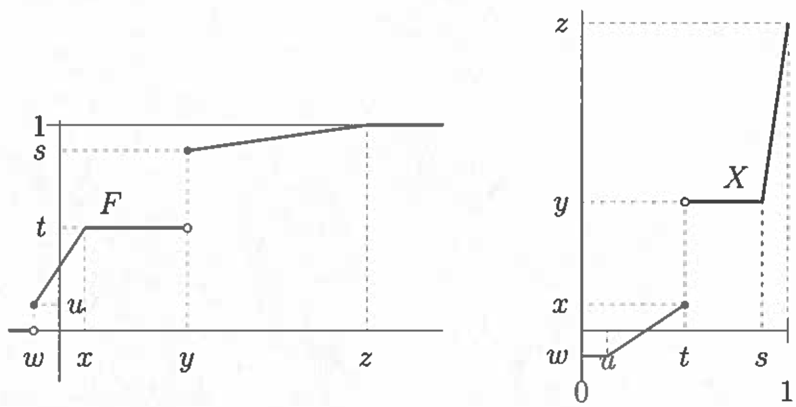

Figure 2.1: Construction of X from F.

First note that X is nondecreasing, and therefore Borel measurable (inverse images of intervals are intervals). To see this, let $t < s$. Then

$$\{z \in \boldsymbol{R} : F(z) \geqslant s\} \subset \{z \in \boldsymbol{R} : F(z) \geqslant t\},$$

so

$$X(t) = \inf\{z \in \boldsymbol{R} : F(z) \geqslant t\} \leqslant \inf\{z \in \boldsymbol{R} : F(z) \geqslant s\} = X(s).$$

Thus X is a random variable on the probability space$[0,1]$. Since F is right continuous, another key property is that

$$X(t) \leqslant y \iff t \leqslant F(y),$$

which implies that

$$\{t \in [0,1] : X(t) \leqslant y\} = [0, F(y)],$$

so

$$\lambda\{t \in [0,1] : X(t) \leqslant y\} = F(y).$$

In other words, F is the cdf of X. ∎

Fractiles are uniformly distributed

Another interesting property of the cdf of X is that if the cdf F_X is continuous, so that its range has no gaps (that is, includes $(0,1)$), then $F_X \circ X$ is a random variable that is uniformly distributed on $[0,1]$. In other words,

$$P\{\omega \in \Omega : F_X(X(\omega)) \leqslant t\} = t, \qquad 0 < t < 1.$$

To see this, fix $t \in [0,1]$, and let $x = \sup\{z \in \boldsymbol{R} : F_X(z) \leqslant t\}$. By continuity, $F_X(x) = t$. Then $F(z) \leqslant t \iff z \leqslant x$, so $F_X(X(\omega)) \leqslant t \iff X(\omega) \leqslant x$. Thus $P\{\omega \in \Omega : F_X(X(\omega) \leqslant t\}) = P\{\omega \in \Omega : X(\omega) \leqslant x\} = F_X(x) = t$.

The Change of Variables Formula #1

The following formula actually allows one to convert integrals on abstract probability spaces to integrals on the real lines, and so allow the use of calculus.

13 Proposition

$$\int_\Omega X(\omega)\, dP(\omega) = \int_{\boldsymbol{R}} t\, dP_X(t)$$

and for any Borel-measurable $g \colon \boldsymbol{R} \to \boldsymbol{R}$,

$$\int_\Omega g \circ X\, dP = \int_{\boldsymbol{R}} g(t)\, dP_X.$$

In other words, *Only distributions matter!* Some people write $\int g\, dF$ for $\int g\, dP$.

2.2 Densities

A measure P on $(\boldsymbol{R}, \mathcal{B})$, has a **density** f (with respect to λ) provided

$$P(E) = \int_{\boldsymbol{R}} \mathbf{1}_E(t) f(t) d\lambda(t)$$

for every set E in $\mathcal{B}$. Note that densities are not unique, they may differ on sets of measure zero.

14 Fact *Let F be the cumulative distribution function of the measure P, that is, $F(t) = P((-\infty, t])$. If P has a density f, then $F'(t)$ exists a.s. and $f = F'(t)$ a.s.*

15 Fact *If X is a random variable whose distribution P_X has a density f_X, and $g \colon \boldsymbol{R} \to \boldsymbol{R}$ is Borel measurable, then*

$$\int_\Omega g(X(\omega))\, dP(\omega) = \int_{\boldsymbol{R}} g(t) f_X(t)\, d\lambda(t).$$

2.3 Marginal Distributions

Move to a section on product measures.

Let

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

be a random vector with distribution ν on $\boldsymbol{R}^n$ (with density $g(x_1, \dots x_n)$).

The **marginal distribution of X_i** is given by

$$\begin{aligned} P[X_i \in B] &= \int_{\boldsymbol{R}^n} 1\boldsymbol{R} \times \cdots \times \underbrace{\boldsymbol{R}\times B\times \boldsymbol{R}}_{i\text{th place}} \times \cdots \times \boldsymbol{R}(x_1, \dots, x_n)\, d\nu(x_1, \dots, x_n) \\ &= \int_{\boldsymbol{R}} \dots \int_B \dots \int_{\boldsymbol{R}} g(x_1, \dots, x_n)\, dx_1 \dots dx_n \end{aligned}$$

Then the density $g_i(x_i)$ of X_i is given by

$$g_i(x_i) = \underbrace{\int_{\boldsymbol{R}} \dots \int_{\boldsymbol{R}}}_{n-1 \text{ times}} g(x_1, \dots, x_n)\, \underbrace{dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n}_{x_i \text{ is not included}}.$$

2.4 Moments of a Random Variable or Distribution

Let P be a probability measure on $(\boldsymbol{R}, \mathcal{B})$. The **first moment** of P is

$$\int_{\boldsymbol{R}} t\, dP(t).$$

If P has a density f_P, this becomes

$$\int_{\boldsymbol{R}} t f_P(t)\, d\lambda(t).$$

The first moment of a random variable X is just the first moment of its distribution. The first moment is also called the **mean** or **expectation**.

The **second moment** of P is $\int_\Omega t^2\, dP(t)$. The **second central moment** of P is

$$\int_{\boldsymbol{R}} (t-m)^2\, dP(t),$$

where m is the first moment of P. This is also called the **variance** of X.

We will now adopt the following notation for random variables.

$$Eg(X) = \int_\Omega g(X(\omega))\, dP(\omega).$$

Thus the variance can be written

$$E((X - EX)^2).$$

The **n^{th} moment** of X is EX^n. The **n^{th} central moment** is

$$E[(X - EX)^n]$$

Do the moments determine the distribution? In general, no—see [4, 8.12]. But if m_k denotes the k^{th} moment and $\limsup_k |m_k|^{\frac{1}{k}} < \infty$ then there is at most one distribution with all these moments.

We can rewrite the variance:

$$\begin{aligned} \operatorname{var} X = E[(X - EX)^2] &= E[X^2 - 2X(EX) + (EX)^2] \\ &= E(X^2) - 2(EX)EX + E[(EX)^2] \\ &= E(X^2) - (EX)^2 \end{aligned}$$

2.5 Some Important Examples

Bernoulli random variables

Let $(\Omega, \mathcal{E}, P)$ be a probability space. Let $E \in \mathcal{E}$ have $P(E) = p$, and let $X = \mathbf{1}_E$. Then

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p. \end{cases}$$

The cumulative distribution function of X is given in Figure 2.2 Such a random variable

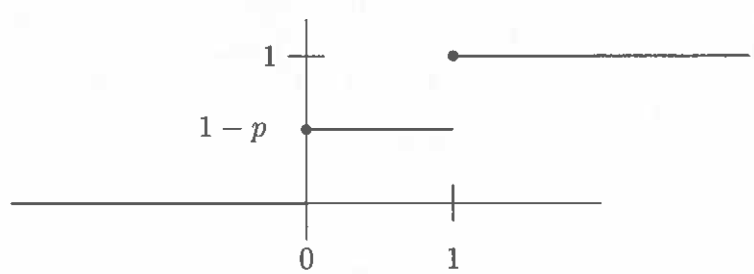

Figure 2.2: The cumulative distribution of a Bernoulli random variable.

is called a **Bernoulli random variable with parameter p**, or a **Bernoulli trial**. The value 1 is often referred to as a **success**.

A Bernoullie random variable X does not have a density. But $\int_{\boldsymbol{R}} g(t)\,dP_X = (1-p)g(0) + pg(1)$. Clearly,

$$EX = \int_\Omega X\,dP = \int_{\boldsymbol{R}} t\,dP_X(t) = (1-p)0 + p1 = p.$$

$$EX^2 = p1^2 = p,$$

and

$$\operatorname{var} X = E(X^2) - (EX)^2 = p - p^2.$$

Normal random variables

A random variable X is said to have a **normal** (or **Gaussian**) **distribution $N(\mu, \sigma^2)$** if P_X has a density f_X given by

$$f_X(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$$

The **standard normal** case is $\mu = 0$ and $\sigma = 1$. A graph of the standard normal density is given in Figure 2.3. The value of the density at $t = 0$ is about 0.39894 and at $t = \pm 3$ is about 0.00443. We show in Proposition 18 below that

$$\frac{1}{\sqrt{2\pi}} \int_{\boldsymbol{R}} e^{-\frac{t^2}{2}}\,dt = 1.$$

The cumulative distribution function for the standard normal is graphed in Figure 2.4. Its value at -3 is about 0.00135 and at 3 is 0.99865. (Note that the scale of the vertical axis exaggerates the curvature in both graphs.) For a more precise estimate of the behavior of the tails, Pollard [15, p. 191] proves that

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{\exp(-\frac{1}{2}t^2)}{\sqrt{2\pi}} \leqslant \text{Prob}\,[N(0,1) \geqslant t] \leqslant \frac{1}{t}\frac{\exp(-\frac{1}{2}t^2)}{\sqrt{2\pi}}$$

for $t > 0$.

The relationship between the standard normal distribution and the others is quite simple.

16 Proposition *Let Z be a standard normal. Then $\sigma Z + \mu$ has the $N(\mu, \sigma^2)$ normal distribution.*

Conversely, if X has the $N(\mu, \sigma^2)$ normal distribution, then $\dfrac{X-\mu}{\sigma}$ is a standard normal.

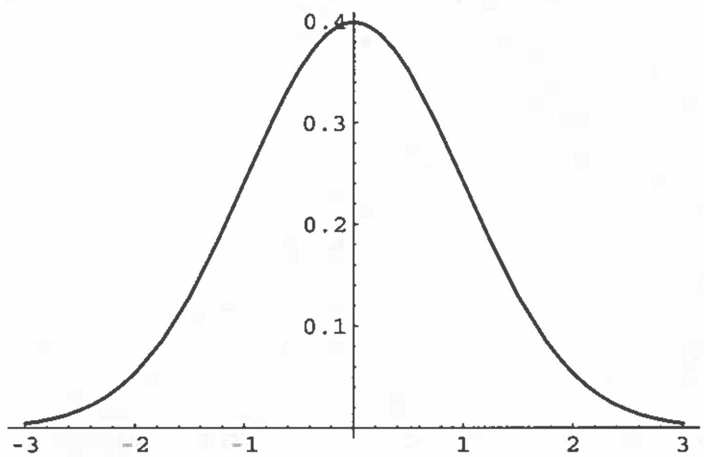

Figure 2.3: The Gaussian density function

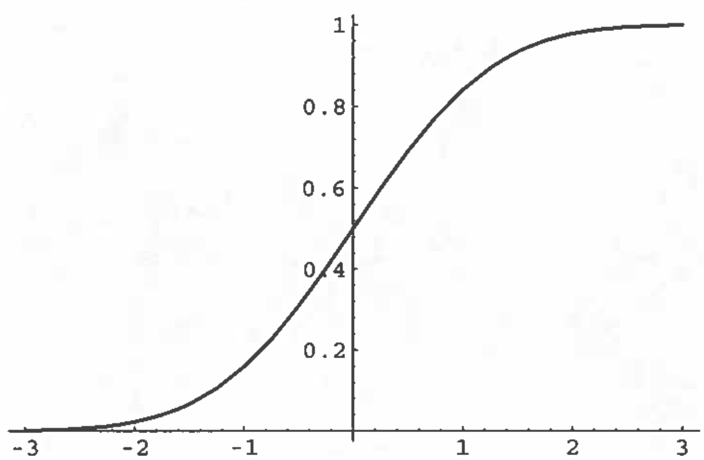

Figure 2.4: The Gaussian cumulative distribution function.

Proof: We prove only the first half. Let $Y = \sigma X + \mu$. We wish to show that $f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

We know $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

$$\begin{aligned} F_Y(t) &= P[Y \leqslant t] = P[\sigma X + \mu \leqslant t] \\ &= P\left[X \leqslant \frac{t-\mu}{\sigma}\right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{t-\mu}{\sigma}} e^{-\frac{1}{2}x^2}\, dx \end{aligned}$$

This is differentiable so the density of Y is $F_Y'(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \frac{1}{\sigma}$. ■

What is the mean of a normal random variable X?

$$\begin{aligned} EX = \int_\Omega X\, dP &= \int_{\boldsymbol{R}} t f(t)\, d\lambda(t) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{\boldsymbol{R}} t e^{-\frac{1}{2\sigma^2}(t-\mu)^2}\, dt, \end{aligned}$$

setting $z = \dfrac{t-\mu}{\sigma}$, so that $dz = \dfrac{1}{\sigma}\, dt$,

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi}} \int_{\boldsymbol{R}} (\mu + \sigma z) e^{-\frac{1}{2}z^2}\, dz \\ &= \frac{\mu}{\sqrt{2\pi}} \int_{\boldsymbol{R}} e^{-\frac{1}{2}z^2}\, dz + \underbrace{\frac{\sigma}{\sqrt{2\pi}} \int_{\boldsymbol{R}} z e^{-\frac{1}{2}z^2}\, dz}_{\frac{\sigma}{\sqrt{2\pi}}\left(\int_0^\infty z e^{-\frac{z^2}{2}}\, dz - \int_{-\infty}^0 z e^{z^2/2}\, dz\right) = \frac{\sigma}{\sqrt{2\pi}}(1-1) \text{ as } \int z e^{-z^2/2} = -e^{-\frac{z^2}{2}}} \\ &= \frac{\mu}{\sqrt{2\pi}} \int_{\boldsymbol{R}} e^{-\frac{1}{2}z^2}\, dz \\ &= \mu, \end{aligned}$$

by Proposition 18 below.

The variance is given by $E(X - EX)^2 = E(X-\mu)^2$

$$= \int_{\boldsymbol{R}} (t-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)}\, dt$$

Set $z = \frac{t-\mu}{\sigma}$ and use integration by parts:

$$\operatorname{var} X = \frac{1}{\sqrt{2\pi}} \int_{\boldsymbol{R}} \sigma^2 z^2 e^{-\frac{z^2}{2}}\, dz$$

Integrate by parts, $\int v\,du = uv - \int u\,dv$ with $u = e^{-\frac{z^2}{2}}$, $\frac{du}{dz} = -ze^{-\frac{z^2}{2}}$, $du = -ze^{-\frac{z^2}{2}}dz$, $v = -z$, $\frac{dv}{dz} = -1$, $dv = -dz$, to get

$$\begin{aligned}
\operatorname{var} X &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{\boldsymbol{R}} z^2 e^{-\frac{z^2}{2}}\,dz \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{\boldsymbol{R}} (-z)\left(-ze^{-\frac{z^2}{2}}\right) dz \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \qquad \underbrace{-ze^{-\frac{z^2}{2}}\Big|_{-\infty}^{+\infty}}_{\substack{=0-0 \\ \text{as exponential dominates} \\ \text{(use l'Hôpital's rule)}}} \qquad -\frac{\sigma^2}{\sqrt{2\pi}} \int_{\boldsymbol{R}} -e^{-\frac{z^2}{2}}\,dz \\
&= \frac{\sigma^2}{\sqrt{2\pi}}(0 + \int_{\boldsymbol{R}} e^{-\frac{z^2}{2}}\,dz) \\
&= \sigma^2.
\end{aligned}$$

Poisson random variables

A random variable X has a **Poisson distribution** if

$$P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}.$$

2.6 Change of Variables in a Multiple Integral

Whoa! What is a multiple integral?

$\boldsymbol{R}^n$ has its own collection of Borel sets $\mathcal{B}^n$ generated by rectangles. n-dimensional Lebesgue measure λ^n, is defined for rectangles to be their volume. It has a unique extension to all of $\mathcal{B}^n$.

For $f\colon \boldsymbol{R}^n \to \boldsymbol{R}$ that is Borel measurable, that is, $f^{-1}(B) \in \mathcal{B}^n$ for every $B \in \mathcal{B}^1$, we can define $\int_{\boldsymbol{R}^n} f(x)\,d\lambda^n(x)$ in the usual way.

There are however some computational tricks. Let's let $n = 2$. Consider

$$\int_{A\times B\subset \boldsymbol{R}^2} f(x,y)d\lambda^2(x,y)$$

Put

$$g(x) = \int_B f(x,y)\,d\lambda^1(y)$$

Is

$$\int_{A\times B} f(x,y)\,d\lambda^2(x,y) \stackrel{?}{=} \int_A g(x)\,d\lambda^1(x)$$

YES, provided f is integrable with respect to λ^2. This result is known as **Fubini's theorem**. Also, if $f(x,y) = g(x)h(y)$, then

$$\int_{A\times B} f(x,y)\,d\lambda^2(x,y) = \int_A g(x)d\lambda^1(x) \times \int_B h(y)\,d\lambda^1(y).$$

Thus we can reduce an integral over $\boldsymbol{R}^n$ to an iterated sequence of 1-dimensional integrals.

2.7 Change of Variables

17 Theorem (Change of Variables # 2) *cf. Hogg and Craig [11, G4.3 and 4.5].*

Let $E \subset \boldsymbol{R}^n$, $H: E \to F \subset \boldsymbol{R}^n$, and let x denote a typical element of F and u a typical element of E. Let $f: F \to \boldsymbol{R}$ be integrable. Then

$$\int_F f(x)\,d\lambda^n(x) = \int_E f(H(u))|J_H(u)|\,d\lambda^n(u)$$

where $J_H(u) = \det\left[\frac{\partial H^j}{\partial u_i}(u)\right]$ provided H is one-to-one and $J_H(u) \neq 0$ except for a set of n-dimensional Lebesgue measure zero.

18 Proposition (The Gaussian Density) *The standard normal density*

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{z^2}{2}\right)}$$

is truly a density. That is, $\int_{\boldsymbol{R}} f(z)\,dz = 1$.

Proof: Set

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left(\frac{z^2}{2}\right)}\,dz$$

We will show $I^2 = 1$ hence $I = 1$. Now

$$\begin{aligned} I^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx \\ &= \frac{1}{2\pi} \int_{\boldsymbol{R}^2} e^{-\frac{1}{2}(x^2+z^2)} d\lambda^2(x,z). \end{aligned}$$

Let

$$H: (r,\theta) \mapsto (r\cos\theta, r\sin\theta)$$

so $H: (0,\infty) \times [0, 2\pi) \to \boldsymbol{R}^2 \setminus \{0\}$ is one-to-one, and

$$J_H(x,\theta) = \det\begin{bmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{bmatrix} = r\cos^2 + r\sin^2\theta = r.$$

By the change of variables formula

$$\frac{1}{2\pi}\int_{R^2} e^{-\frac{1}{2}(x^2+z^2)} d\lambda^2(x,z)$$

$$= \frac{1}{2\pi}\int_{(0,\infty)\times[0,2\pi)} e^{-\frac{1}{2}((r\cos\theta)^2+(r\sin\theta)^2)}|r|\,d\lambda^2(r,\theta)$$

$$= \frac{1}{2\pi}\int_{(0,\infty)\times[0,2\pi)} e^{-\frac{1}{2}r^2} r\,d\lambda^2(r,\theta)$$

by Fubini's Theorem this is equal to

$$= \frac{1}{2\pi}\int_0^{2\pi}\left(\int_0^{\infty} e^{-\frac{1}{2}r^2} r\,d\lambda^1(r)\right) d\lambda'(\theta)$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \underbrace{\left[-e^{\frac{-r^2}{2}}\Big|_0^{\infty}\right]}_{0-(-1)} d\lambda^1(\theta)$$

$$= \frac{1}{2\pi}\int_0^{2\pi} 1\,d\theta = 1$$

Thus $I^2 = 1$, and since clearly $I \geqslant 0$, $I = 1$. ∎

Lecture 3

More on Integration

3.1 Integration by Parts

Let $f, g\colon \boldsymbol{R} \to \boldsymbol{R}$ continuously differentiable. Then

$$(f(t)g(t))' = f'(t)g(t) + g'(t)f(t)$$

so

$$fg = \int f'g\,dt + \int g'f\,dt$$

so

$$\int_a^b f(t)g'(t)\,dt = f(b)g(b) - f(a)g(a) - \int_a^b g(t)f'(t)\,dt$$

19 Fact *Let F, G be cumulative distribution functions. By $\int dF$ we mean $\int dP$ where F is the cumulative distribution function of P. Then*

$$\int_{(a,b]} F(t)\,dG(t) = F(b)G(b) - F(a)G(a) - \int_{(a,b]} G(y^-)\,dF(y),$$

where $G(y^-) = \lim_{\delta \downarrow 0} G(y - \delta)$. The limit y^- can be replaced by y if either F or G is continuous.

20 Corollary *Let F be a cumulative distribution function with $F(0) = 0$, that is, the cumulative distribution function of a nonnegative random variable Then for any $p > 0$,*

$$\int_{[0,\infty)} x^p\,dF(x) = p\int_0^\infty (1 - F(x))x^{p-1}\,dx$$

Proof: Set $G_b(x) = x^p \wedge b^p$. By Integration by Parts (2)

$$\begin{aligned}
\int_0^b x^p\,dF &= F(b)b^p - \int_0^b F(x)\,dG_b(x). \\
&= F(b)b^p - p\int_0^b F(x)x^{p-1}\,dx = p\int_0^b (F(b) - F(x))x^{p-1}\,dx.
\end{aligned}$$

since G_b has derivative px^{p-1}. Now let $b \to \infty$. ∎

3.2 Chebychev's Inequality

Let X be a random variable on $(\Omega, \mathcal{E}, P)$, let φ be a nonnegative and nondecreasing function, and let $\alpha \in \boldsymbol{R}$ satisfy $\varphi(\alpha) > 0$. Then

$$P[|X| > \alpha] \leqslant \frac{E\varphi(|x|)}{\varphi(\alpha)}$$

Proof: Now

$$\begin{aligned} E\varphi(|X|) &= \int_\Omega \varphi(|X|)\, dP \\ &= \int_{[|X|>\alpha]} \varphi(|X|) + \underbrace{\int_{[|X|\leqslant\alpha]} \varphi(|X|) dp}_{\geqslant 0}. \end{aligned}$$

Therefore

$$E(\varphi(|X|)) \geqslant \int_{[|X|>\alpha]} \varphi(|X|)\, dP \geqslant \int_{[|X|>\alpha]} \varphi(\alpha)\, dP = \varphi(\alpha)P[|X| > \alpha],$$

where the second inequality follows from the monotonicity of φ. Now divide by $\varphi(\alpha)$. ∎

21 Corollary *Let X be a random variable with mean μ and variance $\sigma^2 > 0$. Let $k > 0$. Then*

$$P[|X - \mu| \geqslant k\sigma] \leqslant \frac{1}{k^2}.$$

Proof: Apply Chebychev's Inequality to $\varphi(x) = x^2$ and $Y = |X - \mu|$. ∎

3.3 Jensen's Inequality

Let X be a random vector on $(\Omega, \mathcal{E}, P)$ and let the range of X lie in some convex subset C of $\boldsymbol{R}^n$. Let $\varphi: C \to \boldsymbol{R}$ be convex. That is, $\varphi(tx + (1-t)y) \leqslant t\varphi(x) + (1-t)\varphi(y)$ for $t \in [0, 1]$. If X is integrable and φ is subdifferentiable at EX (e.g., if EX is interior to C), then $E(\varphi \circ X)$ is defined (but possibly $+\infty$), and

$$\varphi(EX) \leqslant E\varphi(X).$$

Proof: Let $p \in \boldsymbol{R}^n$ be a subdifferential for φ at EX. That is,

$$\varphi(x) \geqslant p \cdot (x - EX) + \varphi(EX)$$

for all x in C. Then

$$E\varphi(X) \geqslant p \cdot (\underbrace{EX - EX}_{=0}) + E\varphi(EX) = \varphi(EX).$$

∎

3.4 An Inner Product for Random Variables

Let $L_2(\Omega, \mathcal{E}, P)$ denote the set of random variables X on $(\Omega, \mathcal{E}, P)$ such that $E(X^2) < \infty$ (modulo null sets). Define

$$(X, Y) = \int_\Omega X(\omega)Y(\omega)\, dP(\omega) = E(XY).$$

I claim (X, Y) is an inner product. That is,

1. $(X, X) = EX^2 \geqslant 0$ with $= 0$ only if $X = 0$ a.s.
2. $(X, Y) = (Y, X)$.
3. $(\alpha X, Y) = \alpha(X, Y)$.

It follows from the Cauchy–Schwartz inequality that:

$$|(X, Y)|^2 \leqslant (X, X)(Y, Y)$$

with equality only if X and Y are linearly dependent, that is, there exist a, b not both zero $aX + bY = 0$ a.s. We also have a notion of **orthogonality**, etc.

Covariance

Let X, Y be random variables on $(\Omega, \mathcal{E}, P)$. Then

$$E((X - EX)(Y - EY))$$

is called the **covariance** of X and Y (provided the integral exists). We can rewrite this as

$$\begin{aligned}
\operatorname{cov}(X, Y) &= E((X - EX)(Y - EY)) \\
&= E(XY - Y(EX) - X(EY) + (EX)(EY)) \\
&= E(XY) - (EX)(EY)
\end{aligned}$$

The **correlation coefficient** between X and Y is defined to be

$$\frac{\operatorname{cov}(X,Y)}{(\operatorname{var} X)^{\frac{1}{2}}(\operatorname{var} Y)^{\frac{1}{2}}}$$

provided $\operatorname{var} X > 0$ and $\operatorname{var} Y > 0$. By the Cauchy–Schwartz Inequality

$$-1 \leqslant \frac{\operatorname{cov}(X,Y)}{(\operatorname{var} X)^{\frac{1}{2}}(\operatorname{var} Y)^{\frac{1}{2}}} \leqslant 1$$

When is the correlation coefficient equal to ± 1? By Cauchy–Schwartz this occurs only if $(X - EX)$ and $(Y - EY)$ are linearly dependent, that is, there are a, b not both zero such that $a(X - EX) + b(Y - EY) = 0$. Suppose $a \neq 0$. Then

$$X = -\frac{b}{a}Y + (EX - \frac{b}{a}EY)$$

or in other words

$$X = cY + d.$$

When $EX = EY = 0$, then $d = 0$ and $\operatorname{cov}(X,Y) = EXY = cEY^2$ so $c > 0$ if the correlation $= +1$ and $c < 0$ if correlation $= -1$. If the correlation between X and Y is zero, then $X - EX$, $Y - EY$ are orthogonal in terms of our inner product. Expand on this point?

Let $Y, X_1, \ldots, X_n \in L_2$ have mean 0. Let $M = \operatorname{span}(X_1, \ldots, X_n)$. We can orthogonally project Y onto M to get $Y = \sum_{i=1}^n a_i X_i + U$ where $U \perp M$, that is, U is uncorrelated with each X_i.

Variance-Covariance Matrices

Let $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ be a random vector on $(\Omega, \mathcal{E}, P)$. Define $EX = \begin{bmatrix} EX_1 \\ \vdots \\ EX_n \end{bmatrix}$.

22 Exercise *Let $Y = AX$. Then $EY = A(EX)$*

23 Definition *For a random vector X, define $\operatorname{var}(X)$ to be the matrix defined by*

$$\begin{aligned} \operatorname{var}(X) &= E((X - EX)(X - EX)') \\ &= E[(X_i - EX_i)(X_j - EX_j)] \\ &= [\operatorname{cov}(X_i, X_j)] \\ &= E(XX') - (EX)(EX)' \end{aligned}$$

3.5 Stochastic Independence

A family $\mathcal{X}$ of random variables on $(\Omega, \mathcal{E}, P)$ is **(stochastically) independent** if for all $X_1, \ldots, X_n \in \mathcal{X}$, and all $E_1, \ldots, E_n \in \mathcal{E}$,

$$P[X_1 \in E_1, X_2 \in E_2, \ldots, X_n \in E_n]$$

$$= \prod_{i=1}^{n} P[X_i \in E_i].$$

Let $X_1, \ldots, X_n$ be independent and let ν be the distribution of

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

Then

$$\nu(A_1 \times A_2 \times \cdots \times A_n) = \nu_1(A_1)\nu_2(A_2) \cdots \nu_n(A_n)$$

where ν_i is the distribution of X_i on $\boldsymbol{R}$.

Under independence, if ν has a density $f \colon \boldsymbol{R}^n \to \boldsymbol{R}$, then the density $f(t_1, \ldots, t_n)$ factors, i.e.,

$$f(t_1, \ldots, t_n) = f_1(t_1) f_2(t_2) \cdots f_n(t_n)$$

where f_i is a density for ν_i. The converse, that is, factorization of densities or distributions implies independence, is true (provided all the random variables are defined on the same probability space).

Let X, Y be independent random variables on $(\Omega, \mathcal{E}, P)$ with joint distribution ν on $\boldsymbol{R}^2$. Then

$$\begin{aligned} E(XY) &= \int_{\boldsymbol{R}^2} xy \, d\nu(x,y) \\ &= \int_{\boldsymbol{R}} \int_{\boldsymbol{R}} xy \, d\nu_X(x) \, d\nu_Y(y) \quad \text{by Fubini and independence} \\ &= EX \int_{\boldsymbol{R}} y \, d\nu_Y(y) \\ &= EX \, EY. \end{aligned}$$

Therefore

$$\operatorname{cov}(X,Y) = E(XY) - E_X E_Y = 0.$$

The converse is not true in general.

24 Corollary *If X, Y random variables on $(\Omega, \mathcal{E}, P)$ are independent and $E|f(x)| < \infty$ and $E|g(Y)| < \infty$, then $E(f(X) \cdot g(Y)) = (Ef(x)) \cdot (Eg(Y))$.*

Also $f(X)$ and $g(Y)$ are independent for f, g Borel measurable.

3.6 Moment Generating Functions

25 Proposition *Let F, G be cumulative distribution functions with $F(0) = G(0) = 0$. If*

$$\int_{\mathbf{R}} e^{-tx}\, dF(x) = \int_{\mathbf{R}} e^{-tx}\, dG(x) \text{ for all } t \geqslant 0,$$

then $F = G$.

Proof: Breiman [4, Ch. 8.13, Prop 8.51]. ∎

26 Definition *The **moment generating function** M_X of a random variable X is defined by*

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \int_{\mathbf{R}} e^{tx}\, dF_X(x) \end{aligned}$$

provided the integral exists.

27 Fact *If there is some $\delta > 0$ such that $M_F(t) = M_G(t)$ for all $t \in [0, \delta)$, then $F = G$.*

See, e.g., Billingsley [3, p. 305, prob. 26.7].

28 Proposition *Suppose $M_X(t)$ is defined and n-times differentiable at $t = 0$. Then*

$$\begin{aligned} M_X'(t) &= \int_{\mathbf{R}} (e^{tx})'\, dF(x) \\ &= \int_{\mathbf{R}} x e^{tx}\, dF(x). \end{aligned}$$

Thus $M_X'(0) = \int_{\mathbf{R}} x\, dF(x) = EX$. Similarly,

$$M_X^{(n)}(t) = \int x^n e^{tx}\, dF(x)$$

so $M_X^{(n)}(0) = EX^n$.

3.7 The Characteristic Function of X

The **characteristic function** $\varphi_X \colon \mathbf{R} \to \mathbb{C}$ of the random variable X is a complex-valued function defined on the real line by

$$\varphi_X(u) = Ee^{iuX},$$

where of course $i = \sqrt{-1}$. This expectation always exists as a complex number (integrate real and imaginary parts separately) as $|e^{iux}| = 1$ for all u, x.

29 Proposition (Properties of φ_X) *The characteristic function φ satisfies*

1. $\varphi_X(0) = 1$.
2. $|\varphi_X(u)| \leqslant 1$ *for all* u.
3. $\varphi_X(-u) = \overline{\varphi_X(u)}$, *where the bar denotes complex conjugation.*
4. φ *is uniformly continuous.*

Also if X and Y have the same characteristic function, they have the **same distribution**. Thus we can talk about the characteristic functions of distributions. For a random vector

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

define $\varphi_X : \boldsymbol{R}^n \to \mathbb{C}$ by

$$\varphi_X(u_1 \dots, u_n) = E[e^{i(\sum_{i=1}^n u_i X_i)}] = Ee^{iu \cdot X}.$$

Observe that in the one-dimensional case

$$\frac{d}{du}\varphi_X(u) = EiXe^{iuX} = \int ixe^{iux}\, dF_X(x)$$

so

$$\frac{d}{du}\varphi_X(0) = i(EX)$$

and more generally

$$\frac{d^n}{du^n}\varphi_X(0) = i^n(EX^n).$$

30 Fact *If F has a density f and characteristic function φ, then*

$$\varphi(u) = \int_{-\infty}^{\infty} e^{iux} f(x)\, dx$$

and we have the ***inversion formula***

$$f(x) = \int_{-\infty}^{\infty} e^{-iux}\varphi(u)\, du.$$

Characteristic Function of $aX + b$

$$\varphi_{aX+b}(u) = e^{iub}\varphi_X(au).$$

Setting $a = 0$ we get the characteristic function of a constant random variable as

$$\varphi_b(u) = e^{iub}.$$

Independence and Characteristic Functions

Let $X_1, \ldots, X_n$ be random variables on $(\Omega, \mathcal{E}, P)$.

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

31 Fact *Then $X_1, \ldots, X_n$ are stochastically independent if and only if*

$$\varphi_X(u_1, \ldots, u_n) = \prod_{i=1}^{n} \varphi_{X_i}(u_i)$$

Partial proof: ($\Longrightarrow$) Assume a density f:

$$\begin{aligned}
\varphi_X(u_1, \ldots, u_n) &= \int_{R^n} e^{i \sum_{j=1}^{n} u_j x_j} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\
&= \int \prod_{j=1}^{n} e^{iu_j x_j} f_j(x_j)\, dx_1, \ldots, dx_n \quad \text{by independence} \\
&= \prod_{j=1}^{n} \int e^{iu_j x_j} f_j(x_i)\, dx_j \quad \text{by Fubini's Theorem} \\
&= \prod_{j=1}^{n} \varphi_{X_j}(u_j).
\end{aligned}$$

∎

Characteristic Function of an Independent Sum

32 Lemma *Let X, Y be independent random variables on $(\Omega, \mathcal{E}, P)$. Then*

$$\begin{aligned}
\varphi_{X+Y}(u) &= E(e^{iu(X+Y)}) \\
&= E(e^{iuX} e^{iuY}) \\
&= E^{iuX} E e^{iuY} \quad \textit{by independence} \\
&= \varphi_X(u) \varphi_Y(u).
\end{aligned}$$

Characteristic Function of a Bernoulli Random Variable

Let X be a Bernoulli random variable, that is,

$$X = \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p. \end{cases}$$

Then

$$\begin{aligned}\varphi_X(u) &= Ee^{iuX} \\ &= pe^{iu1} + (1-p)e^{iu0} \\ &= pe^{iu} + (1-p)\end{aligned}$$

Characteristic Function of a Normal Random Variable

Let X be a standard normal random variable. Then

$$\varphi_X(u) = Ee^{iuX} = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{iux} e^{\frac{1}{2}x^2}\, dx.$$

Now use the fact for any complex number z,

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

(see, e.g., Rudin [17, p. 1]) to get

$$\begin{aligned}\varphi(u) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} \sum_{n=0}^{\infty} \frac{i^n u^n x^n}{n!} e^{-\frac{1}{2}x^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} \lim_{k\to\infty} \sum_{n=0}^{k} \left(\frac{i^n u^n x^n}{n!} e^{-\frac{1}{2}x^2} \right) dx\end{aligned}$$

Now

$$\left| \sum_{n=0}^{k} \frac{i^n u^n x^n}{n!} e^{-\frac{1}{2}x^2} \right| \leqslant e^{|ux| - \frac{1}{2}x^2},$$

and $e^{|ux|-\frac{1}{2}x^2}$ is integrable:

$$\begin{aligned}\int_{\mathbf{R}} e^{|ux|-\frac{1}{2}x^2}\, dx &= 2 \int_0^{\infty} e^{ux-\frac{1}{2}x^2}\, dx \\ &= \int_0^{2(u+1)} e^{ux-\frac{1}{2}x^2}\, dx + \int_{2(1+u)}^{\infty} e^{ux-\frac{1}{2}x^2}\, dx \\ &\leqslant \int_{2(1+u)}^{\infty} e^{-x}\, dx \\ &= e^{-2(1+u)}\end{aligned}$$

Therefore by the Dominated Convergence Theorem

$$\varphi(u) \xleftarrow[k\to\infty]{} \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{k} \frac{(iu)^n}{n!} \int_{\mathbf{R}} x^n e^{-\frac{1}{2}x^2}\, dx.$$

Let us break this into its pieces. Set $I_n = \int_{\boldsymbol{R}} x^n e^{-\frac{1}{2}x^2}\,dx$. Using integration by parts $\int u\,dv = uv - \int v\,du$, with $u = -x^{n-1}$, $du = -(n-1)x^{n-2}\,dx$, $v = e^{-\frac{x^2}{2}}$, $dv = -xe^{-\frac{x^2}{2}}\,dx$, we get

$$\begin{aligned} I_n &= \int_{\boldsymbol{R}^n} x^n e^{-\frac{1}{2}x^2}\,dx \\ &= \underbrace{-x^{n-1}e^{-\frac{x^2}{2}}\Big|_{-\infty}^{+\infty}}_{=0} - \int_{\boldsymbol{R}} -(n-1)x^{n-2}e^{-\frac{x^2}{2}}\,dx \\ &= (n-1)I_{n-2} \end{aligned}$$

But $I_0 = \int e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi}$ and $I_1 = \int xe^{-\frac{x^2}{2}}\,dx = 0$. (We've done these before.) Therefore:

For n odd, $I_n = 0$.

For n even, $n = 2m$,

$$\begin{aligned} I_{2m} &= \left\lceil (2m-1)\cdots 5\cdot 3\cdot 1\cdot \sqrt{2\pi} \right\rceil \\ &= \frac{(2m)!}{(2m)(2m-2)\ldots 2}\sqrt{2\pi} \\ &= \frac{(2m)!}{2^m(m!)}\sqrt{2\pi}. \end{aligned}$$

So

$$\begin{aligned} \varphi(u) &= \lim_{m\to\infty} \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{m} \frac{(iu)^{2n}}{(2n)!} I_{2n} \\ &= \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{(iu)^{2n}}{(2n)!}\,\frac{(2n)!}{2^n n!}\sqrt{2\pi} \\ &= \sum_{n=0}^{\infty} \left\lceil \frac{(iu)^2}{2} \right\rceil \frac{n}{n!} \\ &= \sum_{n=0}^{\infty} \left(\frac{-u^2}{2} \right) \frac{n}{n!} \\ &= e^{-\frac{u^2}{2}}. \end{aligned}$$

Now for X normal $N(0,1)$, $\sigma X + \mu$ is normal (μ, σ^2), but $\varphi_{\sigma X+\mu} = e^{iu\mu}\varphi_X(\sigma u)$, so

$$\varphi_{N(\mu,\sigma^2)} = e^{-\frac{1}{2}\sigma^2 u^2 + iu\mu}.$$

Characteristice Function of a Poisson Random Variable

The characteristic function is given by

$$
\begin{aligned}
\varphi(u) = Ee^{iuX} &= \sum_{k=0}^{\infty} e^{iuk} e^{-\lambda} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{iu})^k}{k!} = e^{-\lambda}(e^{\lambda e^{iu}}) \\
&= e^{\{\lambda(e^{iu}-1)\}}
\end{aligned}
$$

Lecture 4

The Law of Averages

What is the probability that in $2n$ independent Bernoulli random variables with $p = \frac{1}{2}$ that there are n 1s? (coin-tossing)

$$\binom{2n}{n}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^n = \frac{(2n)!}{n!\,n!}\frac{1}{2^{2n}}$$

Use Stirling's approximation $n! = e^{-n}n^n\sqrt{2\pi n}(1+\varepsilon_n)$ where $\varepsilon_n \to 0$ as $n \to \infty$ (Feller [8, p. 52]) so

$$\frac{(2n)!}{n!\,n!} = \frac{e^{-2n}(2n)^{2n}\sqrt{4\pi n}}{e^{-n}e^{-n}n^n n^n\sqrt{2\pi n}\sqrt{2\pi n}}(1+\delta_n),$$

where $\delta_n \to 0$ as $n \to \infty$. So the probability of n 1's in $2n$ attempts is

$$\begin{aligned} & \frac{2^{2n}}{\sqrt{\pi n}}2^{-2n}(1+\delta_n) \\ = & \frac{1}{\sqrt{\pi n}}(1+\delta_n) \\ \longrightarrow & \ 0 \end{aligned}$$

as $n \to \infty$.

What about the probability of between $n-k$ and $n+k$ heads?

$$\begin{aligned} \leqslant & \frac{2k}{\sqrt{\pi n}}(1+\delta_n) \\ \longrightarrow & \ 0 \end{aligned}$$

as $n \to \infty$.

To get a law of averages, we need to let k grow with n. Hmmm.

Let $\varepsilon > 0$ be given. What is

$$\text{Prob}\left[\#\text{ of ones}/n \in \left(\tfrac{1}{2}-\varepsilon, \tfrac{1}{2}+\varepsilon\right)\right]?$$

This turns out to go to one as $n \to \infty$ for any $\varepsilon > 0$.

This sort of phenomenon is called a **weak law of large numbers**. It says if you look at all n-length sequences of independent Bernoulli random variables the proportion of ones is close to $\frac{1}{2}$ (within ε) with probability $1-\varepsilon$ if n is big enough. A **strong law** would say that in almost every infinite sequence the fraction of ones in the first n values $\to \frac{1}{2}$ as $n \to \infty$.

4.1 A Weak Law

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ, variance $\sigma^2 < \infty$, and let $S_n = \sum_{i=1}^n X_i$. Then for any $\varepsilon > 0$

$$P\left[\left|\frac{s^n}{n} - \mu\right| \leqslant \varepsilon\right] \to 1 \quad \text{as } n \to \infty.$$

Proof: By Chebychev's Inequality ($P[|X| > \alpha] \leqslant \frac{E\varphi(|X|)}{\varphi(\alpha)}$),

$$P\left[\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right] \leqslant \frac{E(\frac{S_n}{n} - \mu)^2}{\varepsilon^2}$$

What is $E\left(\frac{S_n}{n} - \mu\right)^2$?

$$\begin{aligned}
E\left(\frac{S_n}{n} - \mu\right)^2 &= E\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)^2 \\
&= E\frac{1}{n^2}\left(\sum_i (X_i - \mu)\left(\sum_j X_j - \mu\right)\right) \\
&= \frac{1}{n^2}\left(\sum_i \sum_j \underbrace{E(X_i - \mu)(X_j - \mu)}_{\substack{=0 \text{ unless } i=j \\ \text{by independence}}}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \operatorname{var} X_i \\
&= \frac{1}{n}\sigma^2
\end{aligned}$$

Thus

$$P\left[\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right] \leqslant \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty.$$

∎

How about a strong law?

Time to talk about different notions of convergence.

Convergence in Probability

We say that Y_n **converges in probability** to Y, written $Y_n \xrightarrow{P} Y$ or $Y = \operatorname{plim} Y_n$ if

$$\forall \varepsilon > 0 \; P\{\omega : |Y_n(\omega) - Y(\omega)| \leqslant \varepsilon\} \to 1 \quad \text{as } n \to \infty.$$

The above weak law says $\operatorname{plim} \frac{S_n}{n} = \mu$.

Almost Sure Convergence

We say that Y_n **converges almost surely** to Y, written $Y_n \xrightarrow{\text{a.s.}} Y$ or $Y = \lim Y_n$ a.s. if

$$P[Y_n \to Y] = P\{\omega : Y_n(\omega) \to Y(\omega)\} = 1.$$

Comparison

$$Y_n \xrightarrow{P} Y$$
means
$$\forall \delta > 0 \; \forall \varepsilon > 0 \; \exists N \; \forall n \geqslant N \; P\{\omega : |Y_n(\omega) - Y(\omega)| \leqslant \varepsilon\} \geqslant (1 - \delta)$$

$$Y_n \xrightarrow{\text{a.s.}} Y$$
means
$$P\{\omega : \forall \varepsilon > 0 \; \exists N \; \forall n \geqslant N |Y_n(\omega) - Y(\omega)| \leqslant \varepsilon\} = 1$$

33 Proposition

$$Y_n \xrightarrow{\text{a.s.}} Y \Longrightarrow Y_n \xrightarrow{P} Y$$

Proof: Put $A_N = \{\omega : \forall n \geqslant N \; |Y_n(\omega) - Y(\omega)| \leqslant \varepsilon\}$. We want that $P\left(\bigcup_{N=1}^{\infty} A_N\right) = 1$. This is clear from $Y_n \xrightarrow{\text{a.s.}} Y$. ∎

Converse Not True

Consider the sequence defined by

Draw a picture.

$$\begin{aligned}
&Y_1 = \mathbf{1}_{[0,\frac{1}{2})},\; Y_2 = \mathbf{1}_{[\frac{1}{2},1)} \\
&Y_3 = \mathbf{1}_{[0,\frac{1}{4})},\; Y_4 = \mathbf{1}_{[\frac{1}{4},\frac{1}{2})},\; Y_5 = \mathbf{1}_{[\frac{1}{2},\frac{3}{4})},\; Y_6 = \mathbf{1}_{[\frac{3}{4},1)} \\
&Y_7 = \mathbf{1}_{[0,\frac{1}{8})},\; \text{etc.}
\end{aligned}$$

Then $Y_n \xrightarrow{P} 0$, but $Y_n \mathrel{\not\xrightarrow{\text{a.s.}}} 0$.

However we do have the following result.

34 Proposition *If $Y_n \xrightarrow{P} Y$, then for some subsequence $Y_{n_k} \xrightarrow{\text{a.s.}} Y$.*

Proof: Later. It uses the Borel–Cantelli Lemma, [4, p. 44]. ∎

4.2 Kolmogorov's Strong Law of Large Numbers

35 Theorem (Strong Law) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables and set $S_n = \sum_{i=1}^n X_i$. Then*

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu < \infty \iff E|X_i| < \infty \text{ and } EX_i = \mu.$$

Proof: Hard. ∎

4.3 L_p Convergence

We say that Y_n converges in the q^{th} mean, written

$$Y_n \xrightarrow{q} Y$$

if

$$\int_\Omega |Y_n(\omega) - Y(\omega)|^q \, dP(\omega) \xrightarrow[n\to\infty]{} 0.$$

36 Proposition

$$Y_n \xrightarrow{q} Y \quad \Longrightarrow \quad Y_n \xrightarrow{P} Y$$

Proof: By Chebychev's Inequality,

$$P[|Y_n - Y| > \varepsilon] \leqslant \frac{E|Y_n - Y|^q}{\varepsilon^q} \to 0.$$

∎

37 Proposition

$$Y_n \xrightarrow{\text{a.s.}} Y, \; |Y_n|^q \leqslant Z, \; EZ < \infty \quad \Longrightarrow \quad Y_n \xrightarrow{q} Y$$

Proof: Use Lebesgue Dominated Convergence Theorem. ∎

4.4 Convergence in Distribution

Let F_n, F be cumulative distribution functions. We say that F_n **converges in distribution** to F, written

$$F_n \xrightarrow{\mathcal{D}} F$$

if $F_n(t) \to F(t)$ for all t at which F is continuous.

Let X_n, X be random variables we say

$$X_n \xrightarrow{\mathcal{D}} X$$

if

$$F_{X_n} \xrightarrow{\mathcal{D}} F_X.$$

For example, if $X_n(\omega) = \frac{1}{n}$ for all ω, and $X(\omega) = 0$ for all ω, then $X_n \xrightarrow{\mathcal{D}} X$ since $F_{X_n}(t) \to F_X(t)$ for all t except 0, but F isn't continuous at 0.

38 Proposition *If $F_n \xrightarrow{\mathcal{D}} F$, then $\int g\,dF_n \to \int g\,dF$ for any bounded continuous g.*

39 Corollary *If $F_n \xrightarrow{\mathcal{D}} F$, and if the k^{th} moments of F_n and F exist, then the moments converge.*

Proof: Use Proposition 38 applied to truncated t^k, and use monotone convergence. ∎

40 Proposition *Let $X_n \xrightarrow{\mathcal{D}} X$, and let $g\colon \boldsymbol{R} \to \boldsymbol{R}$ be continuous a.e.-F. Then*

$$g(X_n) \xrightarrow{\mathcal{D}} g(X).$$

Proof: Breiman [4, 8.19]. ∎

Note: $X_n \xrightarrow{\mathcal{D}} X$ does not imply that $X_n - X \xrightarrow{\mathcal{D}} 0$. For example, let $X, X_n \sim N(0,1)$. Then $X_n - X \sim N(0,2)$.

4.5 Convergence in Distribution and Characteristic Functions

Let F_n, F be cumulative distribution functions and φ_n, φ be the corresponding characteristic function

$$\varphi(u) = \int_{\boldsymbol{R}} e^{iut}\,dF(t).$$

41 Theorem

$$F_n \xrightarrow{\mathcal{D}} F \iff \varphi_n(u) \to \varphi(u) \quad \textit{for all } u.$$

Proof: Breiman [4, 8.30 and 8.31]. ∎

42 Theorem *Let F_n, F be cumulative distribution functions and suppose that there is some interval $(-\delta, \delta)$ about 0 on which all their moment generating functions*

$$M(t) = \int_{\boldsymbol{R}} e^{tx} dF(x)$$

are defined. Then

$$F_n \xrightarrow{\mathcal{D}} F \iff M_n(t) \to M(t) \textit{ for all } t \in (-\delta, \delta).$$

Proof: Billingsley [3, p. ?]. ∎

Method of Moments

Let $\mu_k = \int_{\boldsymbol{R}} t^k \, dF(t)$. Suppose $\limsup_{k\to\infty} \frac{|\mu_k|^{\frac{1}{k}}}{k} < \infty$. If $\int_{\boldsymbol{R}} t^k \, dF_n(t) \xrightarrow[n\to\infty]{} \mu_k$ for every integer $k \geqslant 0$, then $F_n \xrightarrow{\mathcal{D}} F$.

Proof: Breiman [4, 8.48 and 8.49]. ∎

Lecture 5

Central Limit Theorem(s)

43 Theorem (Version 1) *(Loève [13, 21.1]) Let $X_1, X_2 \ldots$ be i.i.d. with $EX_i = 0$ $EX_i^2 = \sigma^2 < \infty$. Then*

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}\sigma} \xrightarrow{\mathcal{D}} N(0,1).$$

Compare this to the Strong Law, which states $\frac{X_1+\cdots+X_n}{n} \xrightarrow{\text{a.s.}} 0$.

Sketch of proof: Compute the characteristic function φ_n of $\dfrac{X_1 + \cdots + X_n}{\sqrt{n}\sigma}$.

$$\begin{aligned}
\varphi_n(u) &= E\, e^{iu\left\{\frac{X_1+\cdots+X_n}{\sqrt{n}\sigma}\right\}} \\
&= E\left(\Pi_{k=1}^n e^{iu\left\{\frac{X_k}{\sqrt{n}\sigma}\right\}}\right) \\
&= \Pi_{k=1}^n E e^{iu\left\{\frac{X_k}{\sqrt{n}\sigma}\right\}} \quad \text{(independence)} \\
&= \left[E e^{iu\left\{\frac{X_1}{\sqrt{n}\sigma}\right\}}\right]^n \quad \text{(identical distribution)} \\
&= \left[E\left(1 + iu\frac{X_1}{\sqrt{n}\sigma} + \frac{1}{2}\left(iu\frac{X_1}{\sqrt{n}\sigma}\right)^2 + o\left[\left(\frac{iuX_1}{\sqrt{n}\sigma}\right)^2\right]\right)\right]^n \\
&\quad \text{(Taylor Series as } EX_1 = 0\ EX_1^2 = \sigma^2\text{)} \\
&= \left[1 + 0 - \frac{u^2}{2n}\frac{\sigma^2}{\sigma^2} + o\left(\frac{-u^2}{2n}\right)\right]^n
\end{aligned}$$

Now use the following well known fact: $\lim_{x\to 0}(1 + ax + o(bx))^{\frac{1}{x}} \to e^a$, so as $n \to \infty$, $\varphi_n(u) \to e^{-\frac{u^2}{2}}$, which is the characteristic function of $N(0,1)$. ∎

Proof of the well known fact:

$$\ln(1 + ax + o(bx))^{\frac{1}{x}} = \frac{1}{x}\ln(1 + ax + o(bx))$$

$$\begin{aligned}
&= \tfrac{1}{x}[\ln 1 + ax\ln'(1) + o(x)] \\
&= \tfrac{1}{x}[ax + o(x)] \\
&= a + \tfrac{o(x)}{x} \to a \text{ as } \to 0.
\end{aligned}$$

Therefore $e^{(1+ax+\gamma(bx))\frac{1}{x}} \to e^a$ as $x \to 0$. ∎

5.1 A More General CLT

What if the X_is are not identically distributed? We can deal with this using **triangular arrays**. A triangular array is sequence of **rows**, that is, finite sequences, of random variables, where each row is independent.

$$\begin{array}{lllll}
X_{1,1} & \dots & X_{1,k_1} & & \\
X_{2,1} & \dots & \dots & X_{2,k_2} & \\
\vdots & & & \ddots & \\
X_{n,1} & \dots & & \dots & X_{n,k_n} \\
\vdots & & & & \ddots
\end{array}$$

A sequence of independent random variables generates a triangular array in the following natural fashion.

$$\begin{array}{lll}
X_1 & & \\
X_1 & X_2 & \\
X_1 & X_2 & X_3 \\
\vdots & & \ddots
\end{array}$$

We still need to assume that all random variables have mean zero, and finite second movements. Let $S_n = \sum_{j=1}^{k_n} X_{n,j}$ denote the sum of n^{th} row and $c_n^2 = \operatorname{var} S_n$. Assume $c_n^2 > 0$ from some n on.

44 Theorem (Lindeberg's CLT) *If for every $\varepsilon > 0$,*

$$\frac{1}{c_n^2}\sum_{j=1}^{k_n}\int_{[t:|t|\geqslant\varepsilon c_n]} t^2\, dF_{n,j}(t) \to 0 \text{ as } n \to \infty$$

(that is, if the fraction of the variance in the tails goes to zero) then

$$\frac{S_n}{c_n} \to N(0,1).$$

Proof: Loève [13, 21.B, p. 292]; Hall and Heyde [10, 3.2]. ∎

Hall and Heyde argue that this implies $\frac{\max_i |X_{n,i}|}{c_n} \xrightarrow{P} 0$. (?)
This condition is not necessary. For instance, let

$$X_{n,j} = \begin{cases} 0 & j < k_n \\ N(0,1) & j = k_n \end{cases}$$

Then S_n is $N(0,1)$ but $\int_{[t:|t|\geqslant\varepsilon]} t^2\, dF(t) \not\to 0$ as $n \to \infty$.

Lindeberg's Theorem Implies Version I

Assume the hypothesis of version 1, and form the triangular array

$$\begin{array}{lll} X_1 & & \\ \frac{1}{\sqrt{2}\sigma}X_1 & \frac{1}{\sqrt{2}\sigma}X_2 & \\ \vdots & & \\ \frac{1}{\sqrt{n}\sigma}X_1 & \cdots & \frac{1}{\sqrt{n}\sigma}X_n \end{array}$$

Then $c_n^2 = 1$ for each n. We want to show that $S_n \xrightarrow{\mathcal{D}} N(0,1)$. Now ???

$$\begin{aligned} &\sum_{j=1}^{n} \int_{[t:|t|\geqslant\varepsilon]} t^2\, dF_{nj}(t) \\ &= \sum_{j=1}^{n} \int_{\left[\left|\frac{X_j}{\sqrt{n}\sigma}\right|\geqslant\varepsilon\right]} \left|\frac{X_j}{\sqrt{n}\sigma}\right|^2 dP \\ &= \frac{1}{n} \int_{[|X_1|\geqslant\varepsilon\sqrt{n}\sigma]} \frac{X_1^2}{\sigma^2} dP \to 0 \end{aligned}$$

finite as $EX_i^2 < \infty$

45 Corollary *Let $EX^2 = \sigma^2 < \infty$; $F_X = F_{X'} = F_{X''}$ X', X'' independent. Suppose*

$$X \stackrel{\mathcal{D}}{=} \frac{X' + X''}{\sqrt{2}}$$

Then $\frac{1}{\sigma}X \sim N(0,1)$

Proof:

$$EX = \frac{1}{\sqrt{2}}(EX' + EX'') = \frac{2}{\sqrt{2}}EX \quad \text{as } F_X = F_{X'} = F_{X''}$$

$$\Longrightarrow EX = 0$$

For $n = 2^m$

$$\frac{1}{\sigma}X \stackrel{\mathcal{D}}{=} \frac{1}{\sigma}\frac{X_1 + \cdots + X_n}{\sqrt{n}} \to N(0,1).$$

∎

A Converse

46 Proposition (Jain–Feller) $X_1, X_2, \dots$ *i.i.d. Then*

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1) \iff EX_1 = 0 \text{ and } EX_1^2 = 1$$

We also have the following result.

47 Proposition (Breiman [4, 9.6]) *Let* $S_n = \sum_{i=1}^{k_n} X_{n,i}$, *where* $X_{n,i}$ *are i.i.d.*

$$M_n = \max_{1 \leqslant i \leqslant k_n n} |X_{n,i}|$$

Suppose $S_n \xrightarrow{\mathcal{D}} X$.

Then $M_n \xrightarrow{\mathcal{D}} 0 \iff X$ *is normal.*

Rate of Convergence to Normal

48 Theorem (Berry–Esseen) $X_1, \dots$ *i.i.d.*, $EX_i = 0$, $EX_i^2 = \sigma^2$, $E|X_i|^3 < \infty$. *Then*

$$\sup_t |F_n(t) - G(t)| \leqslant \frac{c}{\sigma^3 \sqrt{n}} E|X_i|^3$$

Cf. Breiman [4, p. 184], Loève [13, p. 300]. Le Cam proved $c \leqslant 4$; Breiman claims unpublished work gives $c \leqslant 2.05$.

5.2 Binomial and Poisson Convergence

Proposition 48 leads to the question, what happens if $|X_{n,i}|$ does not get small?

Let $X_{n,i}$ be a triangular array of Bernoulli trials with prob of success p_n in row n. Suppose $p_n k_n \to \lambda > 0$. Put

$$\begin{aligned} S_n &= \sum_{i=1}^{k_n} x_{n,i} \\ &= (1 - p_n)^{k_n} \\ &\to (1 - \frac{\lambda}{k_n})^{k_n} \\ &\to e^{-\lambda} \neq 0 \end{aligned}$$

(Note: $P[M_n = 0] = P[S_n = 0]$.) Then $S_n \xrightarrow{\mathcal{D}} \mathcal{P}(\lambda)$.

Proof:

$$
\begin{aligned}
\text{Characteristic Function of } S_n &= \prod_{i=1}^{k_n} Ee^{iuX_{n,i}} \quad \text{i.i.d.} \\
&= \prod_{i=1}^{k_n} (p_n e^{iu} + (1 - p_n)) \\
&= \prod_{i=1}^{k_n} [1 + p_n(e^{iu} - 1)] \\
&\to [1 + \frac{\lambda}{k_n}(e^{iu} - 1)]^{k_n} \\
&\to e^{\lambda(e^{iu}-1)}.
\end{aligned}
$$

■

The next result is a converse of sorts.

49 Theorem (Poisson Convergence Theorem) *Triangular array $X_{n,i}$ independent in rows such that*

$$\max_i \sigma^2_{n,i} \xrightarrow[n\to\infty]{} 0, \qquad \sum_{i=1}^{k_n} \sigma^2_{n,i} \xrightarrow[n\to\infty]{} \lambda.$$

Then $S_n \xrightarrow{\mathcal{D}} \mathcal{P}(\lambda)$ if and only if

$$E(S_n) \to \lambda \text{ and } \forall \varepsilon > 0 \ \sum_{k=1}^{n_k} \int_{[|t-1| \geqslant \varepsilon]} t^2 dF(t + \mu_{n,k}) \to 0$$

See Loève [13, p. 308].

Lecture 6

Conditioning: The straight dope

Cf. Hogg and Craig [11, ch 2].

Let $X\colon \Omega \to \boldsymbol{R}$ be a random variable, i.e., $X^{-1}(B) \in \mathcal{E}$ for any $B \in \mathcal{B}$. Recall that the collection $\{X^{-1}[B] : B \in \mathcal{B}\}$ is a σ-field contained in $\mathcal{E}$ and is denoted $\mathcal{F}(X)$, the σ-field generated by X.

E.g, let $A \in \mathcal{F}$ and let $X = \mathbf{1}_A$. Then $\mathcal{F}(X) = \{\Omega, \varnothing, A, A^c\}$.

50 Fact *Let $X\colon \Omega \to \boldsymbol{R}$ be a random variable. Let $Y\colon \Omega \to \boldsymbol{R}$ be a random variable that is $\mathcal{F}(X)$-measurable. Then there is a Borel-measurable function $g\colon \boldsymbol{R} \to \boldsymbol{R}$ such that*

$$Y(\omega) = g(X(\omega)) \quad \textit{for all } \omega \in \Omega.$$

Proof: Breiman [4, Appendix A.21]. ∎

51 Definition *Let $X, Y\colon \Omega \to \boldsymbol{R}$ be random variables such that EY exists. The* ***expectation of Y given X****, $E(Y|X)$, is a random variable(!) satisfying*

1. *$E(Y|X)$ is $\mathcal{F}(X)$-measurable*
2. *$\int_A E(Y|X)(\omega)\,dP(\omega) = \int_A Y(\omega)\,dP(\omega)$ for any $A \in \mathcal{F}(X)$.*

52 Fact *$E(Y|X)$ exists (provided EY exists). Several "versions" do, but any two versions disagree only on a set of probability 0.*

Since $E(Y|X)$ is $\mathcal{F}(X)$-measurable there is some function $g\colon \boldsymbol{R} \to \boldsymbol{R}$ such that

$$E(Y|X)(\omega) = g(X(\omega)) \text{ for all } \omega.$$

Let $B \in \mathcal{B}$. Then $E = X^{-1}[B] \in \mathcal{F}(X)$ so

$$\begin{aligned} \int_E Y(\omega)\,dP(\omega) &= \int_{E=[X\in B]} E(Y|X)(\omega)\,dP(\omega) \\ &= \int_E g(X(\omega))\,dP(\omega) \\ &= \int_B g(x)\,d\nu_X(x), \end{aligned}$$

where ν_X is the distribution of X. It is customary to write $g(x)$ as $E(Y|X=x)$. Thus

$$\int_{[X\in B]} Y(\omega)\,dP(\omega) = \int_B E(Y|X=x)d\nu_X(x) \quad \text{for any } B \in \mathcal{B}.$$

Suppose $\begin{bmatrix} X \\ Y \end{bmatrix}$ has a density $f(x,y)$, and X has density $f(x)$. (Then $f(x) = \int_{\boldsymbol{R}} f(x,y)\,dy$.) Now for any $B \in \mathcal{B}$,

$$\begin{aligned}
\int_B E(Y|X=x)f(x)\,dx &= \int_B E(Y|X=x)d\nu_X(x) \\
&= \int_{[X\in B]} Y(\omega)\,dP(\omega) \\
&= \int_{B\times\boldsymbol{R}} \int_{\boldsymbol{R}} yf(x,y)\,dx\,dy \\
&= \int_B \left(\int_{\boldsymbol{R}} yf(x,y)\,dy \right) dx
\end{aligned}$$

Thus

$$E(Y|X=x)f(x) = \int_{\boldsymbol{R}} yf(x,y)\,dy \quad \text{a.e.-}\lambda$$

or

$$E(Y|X=x) = \int_{\boldsymbol{R}} y\frac{f(x,y)}{f(x)}\,dy \quad \text{provided } f(x) > 0.$$

Define

$$f(y|x) = \frac{f(x,y)}{f(x)}.$$

This is called the **conditional density** of Y given X.

Let $h\colon \boldsymbol{R} \to \boldsymbol{R}$ be Borel measurable, and Y be $\mathcal{F}(X)$-measurable. Then $h(Y)$ is $\mathcal{F}(X)$-measurable as well. So what is $E(h(Y)|X)$? A random variable satisfying

$$\int_{[X\in B]} E(h(Y)|X)(\omega)\,dP(\omega) = \int_{[X\in B]} h(Y)(\omega)\,dP(\omega).$$

Similarly $E(h(Y)|X=x)$ satisfies

$$\int_{[X\in B]} h(Y)(\omega)\,dP(\omega) = \int_B E(h(Y)|X=x)d\nu_x(x).$$

Suppose $\begin{bmatrix} X \\ Y \end{bmatrix}$ has density $f(x,y)$, X having density $f(x)$. Then

Bad Notation City!)

$$\begin{aligned}
\int_{B\times\boldsymbol{R}} h(y)f(x,y)\,dx\,dy &= \int_{[X\in B]} h(Y)(\omega)\,dP(\omega) \\
&= \int_B E(h(Y)|X=x)f(x)\,dx.
\end{aligned}$$

Conclude as before that

$$E(h(Y)|X=x)f(x) = \int_{\boldsymbol{R}} h(y)f(x,y)\,dy$$

or

$$E(h(Y)|X=x) = \int_{\boldsymbol{R}} h(y)\frac{f(x,y)}{f(x)}\,dy = \int_{\boldsymbol{R}} h(y)f(y|x)\,dy$$

Some Facts

1. $E(\alpha Y_1 + \beta Y_2|X) = \alpha E(Y_1|X) + \beta E(Y_2|X)$ a.s.
2. $E(E(Y|X)) = EY$ (Since $\Omega \in \mathcal{F}(X)$.)
3. If X and Y are independent, then $E(Y|X) = EY$ a.s.
4. $E(h(X)Y|X) = h(X)E(Y|X)$ a.s.

Conditional Expectation as Projection

Let

$$L_2(\Omega, \mathcal{F}, P) = \{X | X\colon \Omega \to \boldsymbol{R} \text{ is } \mathcal{F}\text{-measurable and } E|X^2| < \infty\}.$$

Then $L_2(\Omega, \mathcal{F}, P)$ is a vector space, and we can define an inner product by

$$(X, Y) = \int_{\Omega} X(\omega)Y(\omega)\,dP(\omega)$$

Let $\mathcal{F}' \subset \mathcal{F}$. Then $L_2(\Omega, \mathcal{F}', P)$ is a linear subspace of $L_2(\Omega, \mathcal{F}, P)$.

53 Proposition *$E(Y|X)$ is the orthogonal projection of Y on $L_2(\Omega, \mathcal{F}(X), P)$.*

6.1 Back to Basics

Whatever happened to $P(B|X) = \dfrac{P(B \cap A)}{P(A)}$ for $P(A) > 0$?

Look at $E(\mathbf{1}_B|\mathbf{1}_A = 1)$. Let ν be the distribution of $\mathbf{1}_A$. Then $\nu\{0\} = P(A^c)$ and $\nu\{1\} = P(A)$. We know that $\mathcal{F}(\mathbf{1}_A) = \{\varnothing, \Omega, A, A^c\}$ so

$$\begin{aligned}
P(B \cap A) &= \int_A \mathbf{1}_B(\omega)\,dP(\omega) \\
&= \int_{A=[\mathbf{1}_A=1]} E(\mathbf{1}_B|\mathbf{1}_A)(\omega)\,dP(\omega) \\
&= \int_{\{1\}} E(\mathbf{1}_B|\mathbf{1}_A = x)\,d\nu(x) \\
&= P(A) \cdot E(\mathbf{1}_B|\mathbf{1}_A = 1).
\end{aligned}$$

Therefore

$$E(\mathbf{1}_B|\mathbf{1}_A = 1) = \frac{P(B \cap A)}{P(A)} \quad \text{if } P(A) > 0.$$

Similarly

$$E(\mathbf{1}_B|\mathbf{1}_A = 0) = \frac{P(B \cap A^c)}{P(A^c)} \quad \text{if } P(A^c) > 0.$$

Likewise

$$E(Y|A) = \frac{S_A Y(\omega)\, dP(\omega)}{P(A)} = E(Y|\mathbf{1}_A = 1).$$

So all our simple ideas are included.

What is $E(Y|\mathbf{1}_A)$?

$$E(Y|\mathbf{1}_A)(\omega) = \begin{cases} E(Y|A) & \omega \in A \\ E(Y|A^c) & \omega \in A^c. \end{cases}$$

It's a random variable, you know.

Lecture 7

Metric spaces of probability measures

Let (X, d) be a metric space. (See my notes on metric spaces.) Recall that a set U is **open** if for every $x \in U$ there is an $\varepsilon > 0$ such that $B_\varepsilon(x) = \{y \in X : d(x,y) < \varepsilon\}$ is included in U. The ε-neighborhood $N_\varepsilon(A)$ of a set A is the open set defined by

$$N_\varepsilon(A) = \{y \in X : \exists x \in A \; d(x,y) < \varepsilon\} = \bigcup_{x \in A} B_\varepsilon(x).$$

A set is **closed** if its complement (in X) is open. The **closure** of a set A , denoted $\overline{A}$, is the smallest closed set that includes A—it consists of the limit points of sequences in A. A set D is **dense** if its closure is X. The space X is **separable** if it includes a countable dense subset. The Euclidean space $\boldsymbol{R}^n$ is separable.

The **Borel σ-algebra** of X is the smallest σ-algebra that contains all the open sets. Members of this σ-algebra are called **Borel sets**. The collection of bounded continuous real-valued functions on X is denoted $C^*(X)$. A **Borel probability** on X is a probability measure on the Borel σ-algebra of X. Every continuous real-valued function on X is a random variable (that is, measurable) with respect to this σ-algebra. Let $\mathcal{P}(X)$ denote the set of all Borel probability measures on X.

We say that a sequence P_n of Borel probabilities **converges weakly** to a Borel probability P, written $P_n \xrightarrow{w} P$, if

$$\int_X f \, dP_n \to \int_X f \, dP \quad \text{for every } f \in C^*(X).$$

This convergence can be described by a metric, provided X is separable.

The **Prohorov metric** on $\mathcal{P}(X)$ is defined by

$$\rho(P, Q) = \inf\{\varepsilon > 0 : \text{for all Borel sets } B \;\; P(B) \leqslant Q \leqslant (N_\varepsilon(B)) + \varepsilon\}.$$

It is not obvious that this is symmetric, but it is indeed a metric, see Dudley [7, Theorem 11.3.1, p. 309].

54 Theorem *For a separable metric space (X, d),*

$$P_n \xrightarrow{w} P \iff \rho(P_n, P) \to 0.$$

For a proof see Dudley [7, Theorem 11.3.3, p. 310].

55 Theorem *The metric space $(\mathcal{P}(X), \rho)$ is separable if and only if X is separable, and compact if and only X is compact.*

See Aliprantis and Border [1, Theorems 14.11, 14.12, p. 482] or Parthasarathy [14, Theorems 6.2, 6.4, pp. 43–44].

When the space X is the real line there is another metric that is useful. Let P and Q be Borel probabilities on the real line $\boldsymbol{R}$. Define the **Lévy metric** λ by

$$\lambda(P, Q) = \inf\{\varepsilon > 0 : \text{for all } x \in \boldsymbol{R} \ \ F_P(x - \varepsilon) \leqslant F_Q(x) \leqslant F_P(x + \varepsilon) + \varepsilon\}.$$

Then λ is a metric and $P_n \xrightarrow{w} P \iff \lambda(P_n, P) \to 0$. See Dudley [7, Problem 8, p. 313] or Billingsley [2, Problem 4, p. 21].

References

[1] Aliprantis, C. D. and K. C. Border. 1999. *Infinite dimensional analysis: A hitchhiker's guide*, 2d. ed. Berlin: Springer–Verlag.

[2] Billingsley, P. 1968. *Convergence of probability measures.* Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

[3] Billingsley, P. 1979. *Probability and measure.* Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

[4] Breiman, L. 1968. *Probability.* Reading, Massachusetts: Addison Wesley.

[5] Chellas, B. F. 1980. *Modal logic: An introduction.* Cambridge: Cambridge University Press.

[6] Dubins, L. E. and L. J. Savage. 1976. *Inequalities for stochastic processes (How to gamble if you must).* New York: Dover. Reprint of *How to gamble if you must: Inequalities for stochastic processes* published by McGraw–Hill, New York, 1965. With a bibliographic supplement.

[7] Dudley, R. M. 1989. *Real analysis and probability.* Pacific Grove, California: Wadsworth & Brooks/Cole.

[8] Feller, W. 1970. *An introduction to probability theory and its applications*, 3d. ed., volume 1. New York: Wiley.

[9] de Finetti, B. 1970. *Theory of probability*, volume 1. London: Wiley.

[10] Hall, P. and C. C. Heyde. 1980. *Martingale limit theory and its application.* Probability and Mathematical Statistics. New York: Academic Press.

[11] Hogg, R. V. and A. T. Craig. 1978. *Introduction to mathematical statistics*, 4th. ed. New York: Macmillan.

[12] Lasota, A. and M. C. Mackey. 1994. *Chaos, fractals, and noise*, 2d. ed. New York: Springer–Verlag. Second edition of *Probabilistic Properties of Deterministic Systems*, published by Cambridge University Press, 1985.

[13] Loève, M. 1977. *Probability theory*, 4th. ed. Number 1 in Graduate Texts in Mathematics. Berlin: Springer–Verlag.

[14] Parthasarathy, K. R. 1967. *Probability measures on metric spaces.* Probability and Mathematical Statistics. New York: Academic Press.

[15] Pollard, D. 1984. *Convergence of stochastic processes.* Springer Series in Statistics. Berlin: Springer–Verlag.

[16] Rao, C. R. 1973. *Linear statistical inference and its applications*, 2d. ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

[17] Rudin, W. 1966. *Real and complex analysis.* New York: McGraw Hill.

[18] Savage, L. J. 1954. *Foundations of statistics.* New York: John Wiley and Sons. Reprint. New York: Dover, 1972.