LECTURE 2

Sums of independent random variables. Suppose $X_1, X_2, \dots$ is a sequence of independent, identically distributed random variables (i.i.d.) , with common mean m. We expect

$$\xi_n = \frac{X_1 + \cdots + X_n}{n} - m \to 0.$$

But what does it mean exactly for $\xi_n \to \xi$ for random variables? Three possibilities (out of many) are

1. $P(\xi_n \to \xi) = 1$ (*almost sure convergence*).
2. $P(|\xi_n - \xi| > \epsilon) \to 0$ as $n \to \infty$ for each fixed $\epsilon > 0$ (*convergence in probability*).
3. $\int f dF_n \to \int f dF$ for every continuous bounded function f where F_n is the distribution function of ξ_n, and F is the distribuition function of ξ. (*convergence in distribution*). One writes $F_n \Rightarrow F$.

Note that it has to be almost surely in the first and not everywhere because changing a random variable on a set of measure zero does not really change anything.

Each type of convergence above implies those below it, but not vice versa.

Lemma. (Tchebyshev's inequality) For any random variable X,

$$P(X \geq a) \leq E[X^2]/a^2.$$

Proof. $P(X \geq a) = \int_{X \geq a} dP \leq \int_{X \geq a} \frac{X^2}{a^2} dP \leq \int \frac{X^2}{a^2} dP$.

Theorem. (Weak law of large numbers) Let $X_1, X_2, \dots$ be i.i.d. with mean m and variance $\sigma^2 < \infty$. Then $\frac{X_1 + \cdots + X_n}{n} \to m$ in probability.

Proof. Let $Y_i = X_i - m$. Then Y_i are i.i.d. with mean 0 and variance σ^2 and $\frac{X_1 + \cdots + X_n}{n} - m = \frac{Y_1 + \cdots + Y_n}{n}$. Now

$$E[(Y_1 + \cdots + Y_n)^2] = \sum_{i,j=1}^{n} E[Y_i Y_j].$$

If $i \neq j$ we have $E[Y_i Y_j] = E[Y_i]E[Y_j] = 0$ since they are independent and mean 0. So only the diagonal terms $E[Y_i^2]$ survive. So

$$E[(\frac{X_1 + \cdots + X_n}{n} - m)^2] = \sigma^2/n.$$

Let $\epsilon > 0$. By Tchebyshev,

$$P\left(|\frac{X_1 + \cdots + X_n}{n} - m| > \epsilon\right) \leq E[(\frac{X_1 + \cdots + X_n}{n} - m)^2]/\epsilon^2 = \sigma^2/n\epsilon^2 \to 0$$

as $n \to \infty$.

Borel-Cantelli Lemma. Let $A_1, A_2, \dots$ be events in $\mathcal{F}$. Suppose that

$$\sum_{i=1}^{\infty} P(A_i) < \infty.$$

Then

$$P(\bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty} B_n) = 0.$$

Proof. $P(\bigcup_{n=k}^{\infty} B_n) \leq \sum_{n=k}^{\infty} P(B_n) \to 0$ as $k \to \infty$.

The set $\bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty} B_n$ is the event that B_n occurs infinitely many times. It is called B_n *infinitely often*, or B_n i.o. for short.

Theorem. (Strong law of large numbers) Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables in L^4. Then $\frac{X_1+\cdots+X_n}{n} \to m$ a.s. (Note: Actually L^1 is enough, but it is hard to prove.)

Proof. The transformation $Y_i = X_i - m$ makes it clear that we can assume without loss of generality that $m = 0$. Now

$$E\left[(X_1+\cdots+X_n)^4\right] = \sum_{i,j,k,\ell} E[X_iX_jX_kX_\ell].$$

If any one of i, j, k, ℓ is different from the others, than that variable pops out of the expectation by independence, and you get 0. Otherwise either you have two pairs and you get a square of second moments, or you have four of a kind and you get a fourth moment. Whatever you get is bounded by some number we can call C. The number of two-pairs is $n(n-1)$ and the number of four-of-a-kinds is n. So $E\left[(X_1+\cdots+X_n)^4\right] \leq Cn^2$ and therefore by Tchebyshev,

$$P\left(\left|\frac{X_1+\cdots+X_n}{n}\right| \geq \epsilon\right) \leq E\left[\left(\frac{X_1+\cdots+X_n}{n}\right)^4\right]/\epsilon^4 \leq C/n^2\epsilon^4.$$

So $P\left(\left|\frac{X_1+\cdots+X_n}{n}\right| \geq \epsilon\right)$ are summable, and therefore by Borel-Cantelli, $P(\left|\frac{X_1+\cdots+X_n}{n}\right| \geq \epsilon\ i.o.) = 0$ which is the same thing as $\frac{X_1+\cdots+X_n}{n} \to m$ a.s.

Lemma. Let F_n be a sequence of distribution functions and $\phi_n(z) = \int e^{izx}dF_n(x)$ the corresponding sequence of characteristic functions. Then $F_n \Rightarrow F$ if and only if $\phi_n(z) \to \phi(z)$ for each $z \in \mathbf{R}$.

Central limit theorem. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean m and variance $\sigma^2 < \infty$. Then $\frac{X_1+\cdots+X_n-nm}{\sigma\sqrt{n}}$ converges in distribution to a standard normal.

Proof. By transforming to $z-$scores $Z_i = \frac{X_i-m}{\sigma}$ we can assume without loss of generality that the X_i have mean 0 and variance 1. Let ϕ be the characteristic function of X_i and ϕ_n be the characteristic function of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$. Then we have

$$\phi_n(z) = [\phi(z/\sqrt{n})]^n.$$

But

$$\phi(z) = E[e^{izX}] = E[1 + izX - \frac{z^2X^2}{2} + o(z^2X^2)] = 1 + \frac{z^2}{2} + o(z^2)$$

where $o(h)$ means something that goes to zero faster than h as h goes to zero (or sometimes as h goes to infinity). So

$$\phi_n(z) = [1 + \frac{z^2}{2n} + o(\frac{1}{n})]^n \to e^{-z^2/2}$$

as $n \to \infty$ for each fixed z.

3. DEPENDENT RANDOM VARIABLES. Let S_n be a simple symmetric random walk on the integers $\mathbf{Z}$. Clearly if S_n is very large then S_{n+1} will be as well. The sequence of differences $X_n = S_n - S_{n-1}$ are independent, but the S_n are not.

Conditional probability. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and A and B are events with $P(A) > 0$. The *conditional probability of B given A* is

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}.$$

Problem. A family has two children. What is the probability that both are girls given that one of them is a girl?

Solution. The sample space is $\Omega = \{bb, bg, gb, gg\}$ where for example bg means the younger child is a boy and the older a girl. So each event has probability $1/4$ (this would not be the case if the children were unordered).

$$P(\{gg\} \mid \{bg, gb, gg\}) = \frac{P(\{gg\})}{P(\{bg, gb, gg\})} = \frac{1/4}{3/4} = 1/3.$$

Conditional probability given a $\sigma-$field. Recall that for any event A, $P(A) = E[\mathbf{1}_A]$ where $\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A; \\ 0 & \text{otherwise.} \end{cases}$ Last time I called it χ_A but I noticed it looks lousy. Sometimes its also called I_A.

We could also make the following equivalent definition of conditional probability. Let $\mathcal{G} = \{\emptyset, A, A^C, \Omega\}$, the smallest $\sigma-$field containing A. We define the *random variable* $P(B \mid \mathcal{G})$ in any of the following equivalent ways:

1. $P(B \mid \mathcal{G})$ is the projection of $\mathbf{1}_B \in L^2(\Omega, \mathcal{F}, P)$ onto the subspace $L^2(\Omega, \mathcal{G}, P)$.
2. For each C in $\mathcal{G}$ define $Q(C) = \int_C \mathbf{1}_B dP = P(C \cap B)$. Then Q is a measure on $(\Omega, \mathcal{G}, P)$ which is absolutely continuous with respect to P, i.e. $P(C) = 0 \Rightarrow Q(C) = 0$. By the Radon-Nikodym theorem there exists a random variable $P(B \mid \mathcal{G})$ on $(\Omega, \mathcal{G}, P)$ such that for any C in $\mathcal{G}$,
$$Q(C) = \int_C P(B \mid \mathcal{G}) dP.$$
3. $P(B \mid \mathcal{G})(\omega) = \begin{cases} \frac{P(B \cap A)}{P(A)} & \text{if } \omega \in A; \\ \frac{P(B \cap A^C)}{P(A^C)} & \text{if } \omega \in A^C. \end{cases}$

Conditional expectation. In the equivalent definitions 1 and 2 above one can replace the random variable $\mathbf{1}_B$ by more or less any random variable (all you need is $X \in L^2$ in 1 and $X \in L^1$ in 2) and the special $\sigma-$field $\mathcal{G}$ by any sub-$\sigma-$field of $\mathcal{F}$. Since 2 is the most general we use it as the definition.

Let $X \in L^1(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subset \mathcal{F}$ a $\sigma-$field. The *conditional expectation of X given $\mathcal{G}$* is the unique random variable $E[X \mid \mathcal{G}] \in L^1(\Omega, \mathcal{G}, P)$ satisfying

$$\int_A E[X \mid \mathcal{G}] dP = \int_A X dP \qquad \text{for all } A \in \mathcal{G}.$$

$E[X \mid \mathcal{G}]$ is the best estimate of X given only the information $\mathcal{G}$.

If X and Y are two random variables then the conditional expectation $E[X \mid Y]$ of X given Y means $E[X \mid \sigma(Y)]$ where $\sigma(Y)$ is the $\sigma-$field generated by Y. Similarly $E[X \mid Y_1, \ldots, Y_n]$ is defined to be the conditional expectation of X given the smallest σ-field with respect to which $Y_1, \ldots, Y_n$ are all measurable.

Basic properties of conditional expectation.

1. If X is $\mathcal{G}$ measurable then $E[X \mid \mathcal{G}] = E[X]$.
2. If $\mathcal{G}$ is the trivial $\sigma-$field $\{\emptyset, \Omega\}$ then $E[X \mid \mathcal{G}] = E[X]$.
3. (linearity) $E[aX + bY \mid \mathcal{G}] = aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]$.
4. If Y is $\mathcal{G}$ - measurable then $E[YX \mid \mathcal{G}] = YE[X \mid \mathcal{G}]$.
5. (repeated conditioning) If $\mathcal{H}$ is a sub-$\sigma-$field of $\mathcal{G}$ then $E[E[X \mid \mathcal{G}] \mid \mathcal{H}] = E[X \mid \mathcal{H}]$.
6. (Jensen's inequality) If ϕ is a convex function then $\phi(E[X \mid \mathcal{G}]) \leq E[\phi(X) \mid \mathcal{G}]$.

Proof. We use the fact that ϕ is convex if and only if $\phi(x) = \sup_{\substack{\ell \text{ linear} \\ \ell(x) \leq \phi(x) \text{ for all } x}} \ell(x)$. We have

$$\ell(E[X \mid \mathcal{G}]) = E[\ell(X) \mid \mathcal{G}] \leq E[\phi(X) \mid \mathcal{G}].$$

Taking the sup on the left hand side gives the result.

Note that all the equality signs above only hold up to sets of measure 0. The conditional expectation is anyway only defined and unique up to sets of measure 0.

Variance of sums of i.i.d. 's. Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with mean 0 and variance 1. Think of them as the amount a game pays back if you bet \$1. So if you keep betting \$1 repeatedly your fortune at time n is $S_n = \sum_{j=1}^n X_j$. But suppose that you bet $\$a_j$ dollars on the jth game. Then your fortune is

$$S_n = \sum_{j=1}^n a_j X_j.$$

a_j could be a function of what has happened up to time $j-1$, i.e. $a_j = a_j(X_1, \ldots, X_{j-1})$. We want to compute the mean and variance of S_n. By linearity

$$E[\sum_{j=1}^n a_j X_j] = \sum_{j=1}^n E[a_j X_j].$$

By repeated conditioning

$$E[a_j X_j] = E[E[a_j X_j \mid X_1, \ldots, X_{j-1}]].$$

Since a_j is measurable with respect to the σ-field generated by $X_1, \ldots, X_{j-1}$,

$$E[a_j X_j \mid X_1, \ldots, X_{j-1}] = a_j E[X_j \mid X_1, \ldots, X_{j-1}].$$

But X_j is independent of $X_1, \ldots, X_n$ so this gives $a_j E[X_j] = 0$. So $E[S_n] = 0$. Now we do the variance the same way.

$$E[(\sum_{j=1}^n a_j X_j)^2] = 2 \sum_{1 \le i < j \le n} E[(a_j X_j)(a_i X_i)] + \sum_{j=1}^n E[(a_j X_j)^2].$$

But if $i < j$,

$$E[(a_j X_j)(a_i X_i)] = E[E[(a_j X_j)(a_i X_i) \mid X_1, \ldots, X_{j-1}]].$$

Everything except X_j pops out of the inner conditional expectation, so inside we are left with $E[X_j \mid X_1, \ldots, X_{j-1}] = 0$. We conclude that all of the off-diagonal terms vanish. So we are left with

$$\sum_{j=1}^n E[(a_j X_j)^2].$$

But

$$E[(a_j X_j)^2] = E[E[(a_j X_j)^2 \mid X_1, \ldots, X_{j-1}]] = E[a_j^2 E[X_j^2 \mid X_1, \ldots, X_{j-1}]] = E[a_j^2].$$

So $E[S_n^2] = \sum_{j=1}^n E[a_j^2]$.

The above calculation is very important and it is worth looking at what we really used. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. We assumed a_j was measurable with respect to $\mathcal{F}_{j-1}$ and all we used for the X_i was that

$$E[X_i \mid \mathcal{F}_{i-1}] = 0.$$

Such a sequence is called a sequence of *martingale differences.*

Martingales. Let $M_1, M_2, \ldots$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{F}_j$, $j = 1, 2, \ldots$ be a filtration. The sequence M_j is called a *martingale with respect to* $\mathcal{F}_j$ if M_j is $\mathcal{F}_j$-measurable and

$$E[M_j \mid \mathcal{F}_{j-1}] = M_{j-1}.$$

If a_j is $\mathcal{F}_{j-1}$-measurable and

$$N_j = \sum_{j=1}^n a_j (M_j - M_{j-1})$$

then the first calculation above shows that N_j is another martingale and the second shows that

$$E[N_j^2] = \sum_{j=1}^n E[a_j^2 \sigma_j^2]$$

where $\sigma_j^2 = E[(M_j - M_{j-1})^2 \mid \mathcal{F}_{j-1}]$. N_j is called the *martingale transform* of M_j. We can write it as

$$\Delta N = a \Delta M.$$

Stochastic differential equations are just continuous time versions of this.

If you are playing a fair game and M_n is your fortune after the nth game then M_n is a martingale.

Submartingales. If we are in the situation above but we only have we have $E[M_j \mid \mathcal{F}_{j-1}] \geq M_{j-1}$ then we say M_j is a submartingale. If the inequality goes the other way then it is called a supermartingale.

Suppose M_j is a martingale and ϕ is a convex function. Then by Jensen's inequality, $X_j = \phi(M_j)$ is a submartingale.

Suppose M_j is a martingale and A_j is $\mathcal{F}_{j-1}$ measurable and increasing in j. Then

$$E[M_j + A_j \mid \mathcal{F}_{j-1}] = M_{j-1} + A_j \geq M_{j-1} + A_{j-1}$$

so $X_j = M_j + A_j$ is a submartingale. Suppose that X_j is a submartingale and let

$$A_j = \sum_{i=1}^{j} E[X_i - X_{i-1} \mid \mathcal{F}_{i-1}].$$

Then A_j are increasing $\mathcal{F}_{j-1}$ measurable, and

$$E[X_j - A_j \mid \mathcal{F}_{j-1}] = X_{j-1} + E[X_j - X_{j-1} \mid \mathcal{F}_{j-1}] - A_{j-1} - E[X_j - X_{j-1} \mid \mathcal{F}_{j-1}] = X_{j-1} - A_{j-1}$$

so $M_j = X_j - A_j$ is a martingale. This shows that X_j is a submartingale if and only if it is the sum of a martingale and an increasing predictable process.

PROBLEMS.

1. On *Let's make a deal* there are three doors. Behind two are goats and behind one is a car. Lets say the car is worth \$30,000 and the goats are worth \$200. You pick a door at random. Then Monty Hall opens one of the other doors to show you a goat and asks you if you want to pay \$1000 to change your door to the remaining one. Should you change it?
2. An urn contains two coins. One is fair and the other has probability 1/3 of falling heads. You draw a coin at random and toss it. It falls heads. What is the probability it is the fair coin?
3. Let $X_1, X_2, \ldots$ be independent random variables with $P(X_i = 1) = P(X_i = -1) = 1/2$ (i.e. *Bernoulli trials*). Let $S_N = X_1 + \cdots + X_N$ (i.e. *symmetric simple random walk*). Calculate $E[S_m \mid S_n]$ for all values of n and m.
4. Let $\Omega = [0, 1)$, $\mathcal{B}$ = the Borel sets, and P = Lebesgue measure. For each positive integer n let $\mathcal{F}_n$ be the smallest σ−field containing the dyadic intevals at level n, i.e., $[0, 1/2^n), [1/2^n, 2/2^n), \ldots, [(2^n - 1)/2^n, 1)$. Let $f \in L^1(\Omega, \mathcal{B}, P)$. Calculate $f_n = E[f \mid \mathcal{F}_n]$.
5. Let $X_1, X_2, \ldots$ be random variables in L^2 with common mean m which are *negatively correlated*, i.e. $Cov(X_i, X_j) < 0$. Show that they satisfy the weak law of large numbers.
6. i. Let X and Y be two random variables. We want to estimate X given knowledge of Y. For any Borel measurable function f we call $f(Y)$ an estimator of X in

terms of Y and $E[(X - f(Y))^2]$ the error of the estimator. An estimator $f^*(Y)$ is called an *optimal estimator* if

$$E[(X - f^*(Y))^2] \leq E[(X - f(Y))^2]$$

for any other estimator $f(Y)$. Show that such an optimal estimator always exists and is unique (up to sets of measure zero). (Hint: conditional expectation.)

ii. Show that if (X, Y) is a Gaussian vector then the optimal estimator of X in terms of Y is a linear function of Y.

7. Suppose that X is a random variable which is uniformly distributed between 0 and 1. Find a transformation $Y = g(X)$ such that Y has a standard normal distribution.