Chapter 5

Joint Distributions and Convergence

5.1 Product measures and Independence

In this section we duscuss the problem of constructing measures on a Cartesian product space and the properties that these measures possess. Such a discussion is essential if we wish to determine probabilities that depend on two or more random variables; for example calculating $P[|X-Y|>1]$ for random variables $X,\ Y$. First consider the analogous problem in $\Re^2$. Given Lebesgue measure λ on $\Re$ how would we construct a similar measure, compatible with the notion of area in two-dimensional Euclidean space? Clearly we can begin with the measure of rectangles or indeed any *product* set of the form $A\times B=\{(x,y); x\in A,\ y\in B\}$ for arbitrary Borel sets $A\subset\Re, B\subset\Re$. Clearly the measure of a product set $\mu(A\times B)=\lambda(A)\lambda(B)$. This defines a measure for any product set and by the extension theorem, since finite unions of product sets form a Boolean algebra, we can extend this measure to the sigma algebra generated by the product sets.

More formally, suppose we are given two measure spaces $(M,\mathcal{M},\mu)$ and $(N,\mathcal{N},\nu)$. Define the *product space* to be the space consisting of pairs of objects, one from each of M and N,

$$\Omega=M\times N=\{(x,y);\ x\in M,\ y\in N\}.$$

The Cartesian product of two sets $A\subset M,\quad B\subset N$ is denoted $A\times B=\{(a,b); a\in A, b\in B\}$. This is the analogue of a rectangle, a subset of $M\times N$, and it is easy to define a measure for such sets as follows. Define the *product measure* of product sets of the above form by $\pi(A\times B)=\mu(A)\nu(B)$. The following theorem is a simple consequence of the Caratheodory Extension Theorem.

Theorem 79 *The product measure π defined on the product sets of the form $\{A\times B; A\in\mathcal{N},\ B\in\mathcal{M}\}$ can be extended to a measure on the sigma algebra*

$\sigma\{A \times B; A \in \mathcal{N}, B \in \mathcal{M}\}$ *of subsets of* $M \times N$.

There are two cases of product measure of importance. Consider the sigma algebra on $\Re^2$ generated by the product of the Borel sigma algebras on $\Re$. This is called the Borel sigma algebra in $\Re^2$. We can similarly define the Borel sigma algebra on $\Re^n$.

In an analogous manner, if we are given two probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ we can construct a *product measure* Q on the Cartesian product space $\Omega_1 \times \Omega_2$ such that $Q(A \times B) = P_1(A)P_2(B)$ for all $A \in \mathcal{F}_1,\ B \in \mathcal{F}_2$. This guarantees the existence of a product probability space in which events of the form $A \times \Omega_2$ are independent of events of the form $\Omega_1 \times B$ for $A \in \mathcal{F}_1,\ B \in \mathcal{F}_2$.

Definition 80 *(Independence, identically distributed) A sequence of random variables* $X_1, X_2, \ldots$ *is* independent *if the family of sigma-algebras* $\sigma(X_1), \sigma(X_2), \ldots$ *are independent. This is equivalent to the requirement that for every finite set* $B_n, n = 1, \ldots N$ *of Borel subsets of* $\Re$*, the events* $[X_n \in B_n], n = 1, ..., N$ *form a mutually independent sequence of events. The sequence is said to be* identically distributed *every random variable* X_n *has the same c.d.f.*

Lemma 81 *If* X, Y *are independent integrable random variables on the same probability space, then* XY *is also integrable and*

$$E(XY) = E(X)E(Y).$$

Proof. Suppose first that X and Y are both simple functions, $X = \sum c_i I_{A_i}, Y = \sum d_j I_{B_j}$. Then X and Y are independent if and only if $P(A_i B_j) = P(A_i)P(B_j)$ for all i, j and so

$$\begin{aligned} E(XY) &= E[(\sum c_i I_{A_i})(\sum d_j I_{B_j})] \\ &= \sum\sum c_i d_j E(I_{A_i} I_{B_j}) \\ &= \sum\sum c_i d_j P(A_i)P(B_j) \\ &= E(X)E(Y). \end{aligned}$$

More generally suppose X, Y are non-negative random variables and consider independent simple functions X_n increasing to X and Y_n increasing to Y. Then $X_n Y_n$ is a non-decreasing sequence with limit XY. Therefore, by monotone convergence

$$E(X_n Y_n) \to E(XY).$$

On the other hand,

$$E(X_n Y_n) = E(X_n)E(Y_n) \to E(X)E(Y).$$

Therefore $E(XY) = E(X)E(Y)$. The case of general (positive and negative random variables X, Y we leave as a problem. ■

5.1.1 Joint Distributions of more than 2 random variables.

Suppose $X_1, ..., X_n$ are random variables defined on the same probability space $(\Omega, \mathcal{F}, P)$. The joint distribution can be characterised by the *joint cumulative distribution function*, a function on $\Re^n$ defined by

$$F(x_1, ..., x_n) = P[X_1 \leq x_1, ..., X_n \leq x_n] = P([X_1 \leq x_1] \cap .... \cap [X_n \leq x_n]).$$

Example 82 *Suppose $n = 2$. Express the probability*

$$P[a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2]$$

using the joint cumulative distribution function.

Notice that the joint cumulative distribution function allows us to find $P[a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n]$. Using inclusion-exclusion,

$$\begin{aligned} &P[a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n] \\ &\quad = F(b_1, b_2, \ldots b_n) - \sum_j F(b_1, ..., a_j, b_{j+1}, ...b_n) \\ &\quad + \sum_{i<j} F(b_1, ..., a_i, b_{i+1}...a_j, b_{j+1}, ...b_n) - ... \end{aligned} \tag{5.1}$$

As in the case $n = 1$, we may then build a probability measure on an algebra of subsets of $\Re^n$. This measure is then extended to the Borel sigma-algebra on $\Re^n$.

Theorem 83 *The joint cumulative distribution function has the following properties:*

(a) $F(x_1, ..., x_n)$ is right-continuous and non-decreasing in each argument x_i when the other arguments $x_j,\ j \neq i$ are fixed.

(b) $F(x_1, ..., x_n) \to 1$ as $min(x_1, ..., x_n) \to \infty$ and $F(x_1, ..., x_n) \to 0$ as $\min(x_1, ..., x_n) \to -\infty$.

(c) The expression on the right hand side of (5.1) is greater than or equal to zero for all $a_1 \leq b_1, \ldots a_n \leq b_n$.

The joint probability distribution of the variables $X_1, ..., X_n$ is a measure on $\mathcal{R}^n$. It can be determined from the cumulative distribution function since (5.1) gives the measure of rectangles, these form a π-system in $\mathcal{R}^n$ and this permits extension first to an algebra and then the sigma algebra generated by these intervals. This sigma algebra is the Borel sigma algebra in $\mathcal{R}^n$. Therefore, in order to verify that the random variables are mutually independent, it is sufficient to verify that the joint cumulative distribution function factors;

$$F(x_1, ..., x_n) = F_1(x_1)F_2(x_2)...F_n(x_n) = P[X_1 \leq x_1] \ldots P[X_n \leq x_n]$$

for all $x_1, ..., x_n \in \Re$.

The next theorem is an immediate consequence of Lemma 81 and the fact that $X_1, ..., X_n$ independent implies that $g_1(X_1), g_2(X_2), ..., g_n(X_n)$ are independent for arbitrary measurable functions $g_i, i = 1, ..., n$.

Theorem 84 *If the random variables* $X_1, ..., X_n$ *are mutually independent, then*

$$E[\prod g_j(X_j)] = \prod E[g_j(X_j)]$$

for any Borel measurable functions $g_1, \ldots, g_n$.

We say an infinite sequence of random variables $X_1, X_2, \ldots$ is mutually independent if every finite subset is mutually independent.

5.2 Strong (almost sure) Convergence

Definition 85 *Let* X *and* $X_n, n = 1, 2, \ldots$ *be random variables all defined on the same probability space* $(\Omega, \mathcal{F})$. *We say that the sequence* X_n *converges* almost surely *(or* with probability one*) to* X *(denoted* $X_n \to X$ *a.s.) if the event*

$$\{\omega; X_n(\omega) \to X(\omega)\} = \cap_{m=1}^{\infty}[|X_n - X| \leq \frac{1}{m} \quad a.b.f.o.]$$

has probability one.

In order to show a sequence X_n converges almost surely, we need that X_n are (measurable) random variables for all n, and to show that there is some measurable random variable X for which the set $\{\omega; X_n(\omega) \to X(\omega)\}$ is measurable and hence an event, and that the probability of this event $P[X_n \to X]$ is 1. Alternatively we can show that for each value of $\epsilon > 0$, $P[|X_n - X| > \epsilon$ i.o.$] = 0$. It is sufficient, of course, to consider values of ϵ of the form $\epsilon = 1/m$, m=1,2,... above.

The law of large numbers (sometimes called the law of averages) is the single most important and well-known result in probability. There are many versions of it but the following is sufficient, for example, to show that the average of independent Bernoulli random variables, or Poisson, or negative binomial, or Gamma random variables, to name a few, converge to their expected value **with probability one**.

Theorem 86 *(Strong Law of Large Numbers) If* X_n, $n = 1, 2, \ldots$ *is a sequence of independent identically distributed random variables with* $E|X_n| < \infty$, *(i.e. they are integrable) and* $E(X_n) = \mu$, *then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad a.s. \quad as \quad n \to \infty$$

Proof. We shall prove this result in the special case $E(X^4) < \infty$. The more general proof will be left for later. First note that, by replacing X_i by $X_i - \mu$ we may assume that $\mu = 0$ without any loss of generality. Now note that with $S_n = \sum_{i=1}^n X_i$, and letting $var(X_i) = \sigma^2$ and $E(X_i^4) = K$, we have

$$E(S_n^4) = nK + 3n(n-1)\sigma^4 = d_n, \quad say.$$

Therefore for each $\epsilon > 0$, we have by Chebyschev's inequality

$$P\{|\frac{S_n}{n}| > \epsilon\} \leq \frac{E(S_n^4)}{\epsilon^4 n^4} = \frac{d_n}{\epsilon^4 n^4}$$

Note that since $\sum_n \frac{d_n}{n^4} < \infty$ we have by the first Borel Cantelli Lemma,

$$P\{|\frac{S_n}{n}| > \epsilon \ i.o.\} = 0.$$

Since this holds for all $\epsilon > 0$ it follows that the probability that $\frac{S_n}{n}$ does not converge to 0 is 0 and so the probability that it does converge is 1. ■

5.3 Weak Convergence (Convergence in Distribution)

Consider random variables that are constants; $X_n = 1 + \frac{1}{n}$. By any sensible definition of convergence, X_n converges to $X = 1$ as $n \to \infty$. Does the cumulative distribution function of X_n, F_n, say, converge to the cumulative distribution function of X pointwise? In this case it is true that $F_n(x) \to F(x)$ at all values of x except the value $x = 1$ where the function $F(x)$ has a discontinuity. Convergence in distribution (weak convergence, convergence in Law) is defined as pointwise convergence of the c.d.f. at all values of x except those at which $F(x)$ is discontinuous. Of course if the limiting distribution is absolutely continuous (for example the normal distribution as in the Central Limit Theorem), then $F_n(x) \to F(x)$ does hold for all values of x.

Definition 87 *(Weak Convergence) If $F_n(x)$, $n = 1, \ldots$ is a sequence of cumulative distribution functions and if F is a cumulative distribution function, we say that F_n converges to F* weakly *or* in distribution *if $F_n(x) \to F(x)$ for all x at which $F(x)$ is continuous. We will sometimes denote weak convergence of a sequence of random variables X_n whose c.d.f. converges in the above sense by $X_n \Rightarrow X$.*

Example 88 *(Maximum of independent exponential(α)) Suppose $(X_1, ..., X_n)$ are independent exponentially distributed random variables all with the exponential cumulative distribution function*

$$F(x) = 1 - e^{-\alpha x}.$$

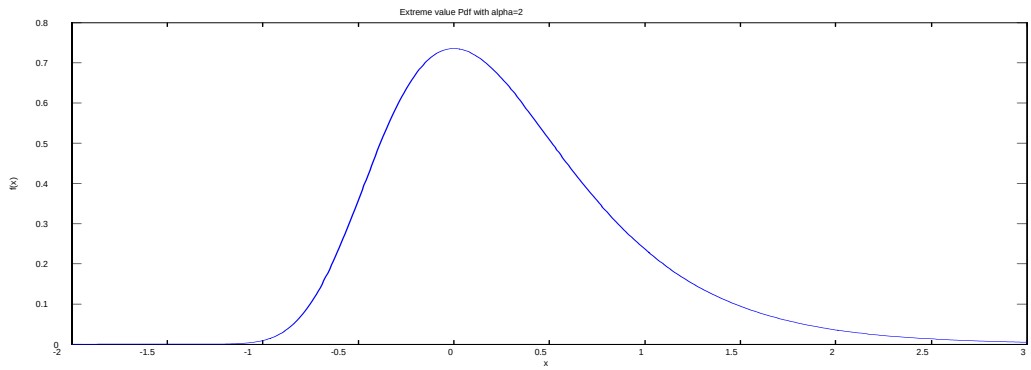

Figure 5.1: The Extreme Value Probability Density Function

Define $M_n = max(X_1, ..., X_n)$. *Then the c.d.f. of* $M_n - (log\ n)/\alpha$ *is*

$$F_{M_n}(x) = (1 - e^{-(\alpha x + log\ n)})^n \rightarrow F(x) = e^{-e^{-\alpha x}}$$

Proof. Note that for arbitrary $x \in \mathcal{R}$

$$\begin{aligned} P[M_n - \frac{\ln n}{\alpha} \leq x] = P[M_n \leq x + \frac{\ln n}{\alpha}] &= [F(x + \frac{\ln n}{\alpha})]^n \\ &= (1 - e^{-\alpha x - \ln n})^n = (1 - \frac{1}{n} e^{-\alpha x})^n \\ &\rightarrow \exp(-e^{-\alpha x}) \text{ as } n \rightarrow \infty. \end{aligned}$$

■

For any independent identically distributed random variables such that the cumulative distribution function satisfies $1 - F(x) \sim e^{-\alpha x}$, the same result holds. The limiting distribution whose cumulative distribution function is of the form $F(x) = \exp(-e^{-\alpha x})$ is called an *extreme value distribution* and is commonly used in environmental, biostatistical and engineering applications of statistics. The corresponding probability density function is

$$\frac{d}{dx} e^{-e^{-\alpha x}} = \alpha \exp\left(-\alpha x - e^{-\alpha x}\right), -\infty < x < \infty$$

and is shaped like a slightly skewed version of the normal density function (see Figure 1 for the case $\alpha = 2$).

This example also shows approximately how large a maximum will be since $M_n - (\ln n)/\alpha$ converges to a proper distribution. Theoretically, if there were no improvement in training techniques over time, for example, we would expect that the world record in the high jump or the shot put at time t (assuming the number of competitors and events occurred at a constant rate) to increase like $ln(t)$. However, records in general have increased at a much higher rate,

indicating higher levels of performance, rather than just the effect of the larger number of events over time. Similarly, record high temperatures since records in North America were begun increase at a higher rate than this, providing evidence of global warming.

Example 89 *Suppose* $1 - F(x) \sim x^{-\alpha}$ *for* $\alpha > 0$. *Then the cumulative distribution function of* $n^{-1/\alpha} M_n$ *converges weakly to* $F(x) = e^{-x^{-\alpha}}$, *for* $x > 0$ *(The distribution with the cumulative distribution function* $F(x) = e^{-x^{-\alpha}}$ *is called the* Weibull *distribution).*

Proof. The proof of the convergence to a Weibull is similar to that for the extreme value distribution above.

$$\begin{aligned} P[n^{-1/\alpha} M_n \leq x] &= [F(n^{1/\alpha} x)]^n \\ &= [1 - (n^{1/\alpha} x)^{-\alpha} + o(n^{-1})]^n \\ &= [1 - \frac{1}{n} x^{-\alpha} + o(n^{-1})]^n \\ &\to \exp(-x^{-\alpha}) \quad \text{as } n \to \infty \end{aligned}$$

■

We have used a slight extension of the well-known result that $(1+c/n)^n \to e^c$ as $n \to \infty$. This result continues to hold even if we include in the bracket and additional term $o(n^{-1})$ which satisfies $no(n^{-1}) \to 0$. The extension that has been used (and is easily proven) is $(1 + c/n + o(n^{-1}))^n \to e^c$ as $n \to \infty$.

Example 90 *Find a sequence of cumulative distribution functions* $F_n(x) \to F(x)$ *for some limiting function* $F(x)$ *where this limit is not a proper c.d.f.*

There are many simple examples of cumulative distribution functions that converge pointwise but not to a genuine c.d.f. All involve some of the mass of the distribution "escaping" to infinity. For example consider F_n the $N(0, n)$ cumulative distribution function. Of more simply, use F_n the cumulative distribution function of a point mass at the point n. However there is an additional condition that is often applied which insures that the limiting distribution is a "proper" probability distribution (i.e. has total measure 1). This condition is called tightness.

Definition 91 *A sequence of probability measures* P_n *defined on a probability space which is also a metric space is tight if for all* $\epsilon > 0$, *there exists a compact set* K *such that* $P_n(K^c) \leq \epsilon$ *for all* n.

A sequence of cumulative distribution functions F_n is tight if it corresponds to a sequence of tight probability measures on $\mathcal{R}$. If these are the cumulative distribution functions of a sequence of random variables X_n, then this is equivalent to the requirement that for every $\epsilon > 0$, there is a value of $M < \infty$ such

that the probabilities outside the interval $[-M, M]$ are less than ϵ. In other words

$$P[X_n \notin [-M, M]] \leq \epsilon \text{ for all } n = 1, 2, ...\text{or}$$
$$F_n(-M-) + (1 - F_n(M)) \leq \epsilon \quad \text{for all } n = 1, 2, ...$$

If a sequence F_n converges to some limiting right-continuous function F at continuity points of F and if the sequence is tight, then F is a c.d.f. of a probability distribution and the convergence is in distribution or weak (see Problem 6).

Lemma 92 *If $X_n \Rightarrow X$, then there is a sequence of random variables Y, Y_n on some other probability space (for example the unit interval) such that Y_n has the same distribution as X_n and Y has the same distribution as X but $Y_n \to Y$ almost surely.*

Proof. Suppose we take a single uniform[0,1] random variable U. Recall the definition of pseudo inverse used in Theorem 58, $F^{-1}(y) = \sup\{z; F(z) < y\}$. Define $Y_n = F_n^{-1}(U)$ and $Y = F^{-1}(U)$ where F_n and F are the cumulative distribution functions of X_n and X respectively. We need to show that if $F_n(x) \to F(x)$ at all x which are continuity points of the function F, then $F_n^{-1}(U) \to F^{-1}(U)$ almost surely. First note that the set of $y \in [0, 1]$ such that $\{x; F(x) = y\}$ contains more than one point (i.e. is an interval) has Lebesgue measure 0. For example the function $F^{-1}(y)$ is a monotone function and therefore has at most a countable number of discontinuities and the discontinuities of $F^{-1}(y)$ are the "flat portions " of F.. Suppose we choose y such that $\{x; F(x) = y\}$ has at most one point in it. Then for any $\varepsilon > 0$, putting $x = F^{-1}(y)$, it is easy to see that $F(x + \varepsilon) > y$ and $F(x - \varepsilon) < y$. If we now choose N sufficiently large that for $n > N$, we have

$$|F_n(z) - F(z)| \leq \min(y - F(x - \varepsilon), F(x + \varepsilon) - y)$$

for the two points $z = x - \varepsilon$ and $z = x + \varepsilon$, then it is easy to show that

$$|F_n^{-1}(y) - F^{-1}(y)| \leq \varepsilon.$$

It follows that $F_n^{-1}(U)$ converges almost surely to $F^{-1}(U)$ ■

Theorem 93 *Suppose $X_n \Rightarrow X$ and g is a Borel measurable function. Define $D_g = \{x; g$ is discontinuous at $x\}$. If $P[X \in D_g] = 0$, then $g(X_n) \Rightarrow g(X)$.*

Proof. We prove this result assuming the last lemma which states that we can find a sequence of random variables Y_n and a random variable Y which have the same distribution as X_n, X respectively but such that Y_n converges almost surely (i.e. with probability one) to Y. Note that in this case

$$g(Y_n(\omega)) \to g(Y(\omega))$$

provided that the function g is continuous at the point $Y(\omega)$ or in other words, provided that $Y(\omega) \notin D_g$. Since $P[Y(\omega) \notin D_g] = 1$, we have that

$$g(Y_n) \to g(Y) \text{ a.s.}$$

and therefore convergence holds also in distribution (you may either use Theorems 97 and 98 or prove this fact seperately). But since Y_n and X_n have the same distribution, so too do $g(Y_n)$ and $g(X_n)$ implying that $g(X_n)$ converges in distribution to $g(X)$. ■

In many applications of probability, we wish to consider stochastic processes $X_n(t)$ and their convergence to a possible limit. For example, suppose $X_n(t)$ is defined to be a random walk on discrete time, with time steps $1/n$ and we wish to consider a limiting distribution of this process as $n \to \infty$. Since X_n is a stochastic process, not a random variable, it does not have a cumulative distribution function, and any notion of weak convergence must not rely on the c.d.f. In this case, the following theorem is used as a basis for defining weak convergence. In general, we say that X_n *converges weakly to* X *if* $E[f(X_n)] \to E[f(X)]$ *for all bounded continuous functions* f. This is a more general definition of weak convergence.

Definition 94 *(general definition of weak convergence) A sequence of random elements X_n of a metric space $\mathcal{M}$ is said to "converge weakly to a random element X i.e. $X_n \Rightarrow X$ if and only if $E[f(X_n)] \to E[f(X)]$ for all bounded continuous functions f from $\mathcal{M}$ to $\mathcal{R}$.*

Theorem 95 *If X_n and X are random variables, X_n converges weakly to X in the sense of definition 94 if and only if $F_n(x) \to F(x)$ for all $x \notin D_F$.*

Proof. The proof is based on Lemma 92. Consider a sequence Y_n such that Y_n and X_n have the same distribution but $Y_n \to Y$ almost surely. Since $f(Y_n)$ is bounded above by a constant (and the expected value of a constant is finite), we have by the Dominated Convergence Theorem $Ef(Y_n) \to Ef(Y)$. (We have used here a slightly more general version of the dominated convergence theorem in which convergence holds almost surely rather than pointwise at all ω.) For the converse direction, assume $E[f(X_n)] \to E[f(X)]$ for all bounded continuous functions f. Suppose we take the function

$$f_\epsilon(t) = \begin{cases} 1, & t \le x \\ 0, & t > x + \epsilon \\ \frac{x+\epsilon-t}{\epsilon}, & x < t < x + \epsilon \end{cases}$$

Assume that x is a continuity point of the c.d.f. of X. Then $E[f_\epsilon(X_n)] \to E[f_\epsilon(X)]$. We may now take $\epsilon \to 0$ to get the convergence of the c.d.f. ■

5.4 Convergence in Probability

Definition 96 *We say a sequence of random variables $X_n \to X$* in probability *if for all $\epsilon > 0$, $P[|X_n - X| > \epsilon] \to 0$ as $n \to \infty$.*

Convergence in probability is in general a somewhat more demanding concept than weak convergence, but less demanding than almost sure convergence. In other words, convergence almost surely implies convergence in probability and convergence in probability implies weak convergence.

Theorem 97 *If* $X_n \to X$ almost surely *then* $X_n \to X$ in probability.

Proof. Because we can replace X_n by $X_n - X$, we may assume without any loss of generality that $X = 0$. Then the set on which X_n converges almost surely to zero is

$$\{\omega; X_n(\omega) \to 0\} = \cap_{m=1}^{\infty}([|X_n| \leq 1/m] a.b.f.o.)$$

and so for each $\epsilon = 1/m > 0$, we have, since X_n converges almost surely,

$$P([|X_n| \leq \epsilon] a.b.f.o.) = 1.$$

or

$$1 = P(\cup_{j=1}^{\infty} \cap_{n=j}^{\infty} [|X_n| \leq \epsilon]) = lim_{j\to\infty} P(\cap_{n=j}^{\infty}[|X_n| \leq \epsilon]).$$

Since $P(\cap_{n=j}^{\infty}[|X_n| \leq \epsilon]) \leq P[|X_j| \leq \epsilon]$ and the sets $\cap_{n=j}^{\infty}[|X_n| \leq \epsilon]$ are increasing in j, it must follow that

$$P[|X_j| \leq \epsilon] \to 1 \text{ as } j \to \infty.$$

■

Convergence in probability does not imply convergence almost surely. For example let $\Omega = [0, 1]$ and for each n write it uniquely in the form $n = 2^m + j$ for $0 \leq j < 2^m$. Define $X_n(\omega)$ to be the indicator of the interval $[j/2^m, (j + 1)/2^m)$. Then X_n converges in probability to 0 but $P[X_n \to 0] = 0$.

Theorem 98 *If* $X_n \to X$ in probability*, then* $X_n \Rightarrow X$.

Proof. Assuming convergence in probability, we need to show that $P[X_n \leq x] \to P[X \leq x]$ whenever x is a continuity point of the function on the right hand side. Note that

$$P[X_n \leq x] \leq P[X \leq x + \epsilon] + P[|X_n - X| > \epsilon]$$

for any $\epsilon > 0$. Taking limits on both sides as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \sup P[X_n \leq x] \leq P[X \leq x + \epsilon].$$

By a similar argument

$$\lim_{n\to\infty} \inf P[X_n \leq x] \geq P[X \leq x - \epsilon].$$

Now since $\epsilon > 0$ was arbitrary, we may take it as close as we wish to 0. and since the function $F(x) = P[X \leq x]$ is continuous at the point x, the limit as $\epsilon \to 0$ of both $P[X \leq x + \epsilon]$ and $P[X \leq x - \epsilon]$ is $F(x)$. It follows that

$$F(x) \leq liminf P[X_n \leq x] \leq limsup P[X_n \leq x] \leq F(x)$$

and therefore $P[X_n \leq x] \to F(x)$ as $n \to \infty$. ■

Theorem 99 *If $X_n \Rightarrow c$* i.e. in distribution *for some constant c then $X_n \to c$ in probability.*

Proof. Since the c.d.f. of the constant c is $F(x) = 0$,if $x < c$, and $F(x) = 1$,if $x \geq c$, and since this function is continuous at all points except the point $x = c$, we have, by the convergence in distribution,

$$P[X_n \leq c + \epsilon] \to 1 \quad \text{and} \quad P[X_n \leq c - \epsilon] \to 0$$

for all $\epsilon > 0$. Therefore,

$$P[|X_n - c| > \epsilon] \leq (1 - P[X_n \leq c + \epsilon]) + P[X_n \leq c - \epsilon] \to 0.$$

■

Theorem 100 *If X_n and Y_n, $n = 1, 2, \ldots$ are two sequences of random variables such that $X_n \Rightarrow X$ and $Y_n - X_n \Rightarrow 0$, then $Y_n \Rightarrow X$.*

Proof. Assume that $F(x)$, the c.d.f of X is continuous at a given point x. Then for $\epsilon > 0$,

$$P[Y_n \leq x - \epsilon] \leq P[X_n \leq x] + P[|X_n - Y_n| > \epsilon].$$

Now take limit supremum as $n \to \infty$ to obtain

$$\limsup P[Y_n \leq x - \epsilon] \leq F(x).$$

A similar argument gives

$$\liminf P[Y_n \leq x + \epsilon] \geq F(x).$$

Since this is true for ϵ arbitrarily close to 0, $P[Y_n \leq x] \to F(x)$ as $n \to \infty$. ■

Theorem 101 *(A Weak Law of Large Numbers) If X_n, $n = 1, 2, \ldots$ is a sequence of independent random variables all with the same expected value $E(X_n) = \mu$, and if their variances satisfy $\frac{1}{n^2}\sum_{i=1}^n var(X_i) \to 0$, then $\frac{1}{n}\sum_{i=1}^n X_i \to \mu$ in probability.*

Proof. By Chebyschev's inequality,

$$P[|\frac{\sum_{i=1}^n X_i}{n} - \mu| \geq \epsilon] \leq \frac{\sum_{i=1}^n var(X_i)}{\epsilon^2 n^2}$$

and this converges to 0 by the assumptions. ■

5.5 Fubini's Theorem and Convolutions.

Theorem 102 *(Fubini's Theorem) Suppose $g(x,y)$ is integrable with respect to a product measure $\pi = \mu \times \nu$ on $M \times N$, then*

$$\int_{M\times N} g(x,y)d\pi = \int_M [\int_N g(x,y)d\nu]d\mu \;=\; \int_N [\int_M g(x,y)d\mu]d\nu.$$

We can dispense with the assumption that the function $g(x,y)$ is integrable in Fubini's theorem (permiting infinite integrals) if we assume instead that $g(x,y) \geq 0$.

Proof. First we need to identify some measurability requirements. Suppose E is a set measurable with respect to the product sigma-algebra on $M \times N$. We need to first show that the set $E_y = \{x \in M; (x,y) \in E\}$ is a measurable set in M. Consider the class of sets

$$\mathcal{C} = \{E; \{x \in M; (x,y) \in E\} \text{ is measurable in the product sigma algebra}\}$$

It is easy to see that $\mathcal{C}$ contains all product sets of the form $A \times B$ and that it satisfies the properties of a sigma-algebra. Therefore, since the product sigma algebra is generated by $\{A \times B; A \in \mathcal{N}, B \in \mathcal{M}\}$, it is contained in $\mathcal{C}$. This shows that sets of the form E_y are measurable. Now define the measure of these sets $h(y) = \mu(E_y)$. The function $h(y)$ is a measurable function defined on N (see Problem 23).

Now consider a function $g(x,y) = I_E$ where $E \in \mathcal{F}$. The above argument is needed to show that the function $h(y) = \int_M g(x,y)d\mu$ is measurable so that the integral $\int_N h(y)d\nu$ potentially makes sense. Finally note that for a set E of the form $A \times B$,

$$\int I_E d\pi = \pi(A \times B) = \mu(A)\nu(B) = \int_N (\int_M I_E(x,y)d\mu)d\nu$$

and so the condition $\int I_E d\pi = \int_N (\int_M I_E(x,y)d\mu)d\nu$ holds for sets E that are product sets. It follows that this equality holds for all sets $E \in \mathcal{F}$(see problem 24). Therefore this holds also when I_E is replaced by a simple function. Finally we can extend this result to an arbitrary non-negative function g by using the fact that it holds for simple functions and defining a sequence of simple functions $g_n \uparrow g$ and using monotone convergence. ■

Example 103 *The formula for integration by parts is*

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

Does this formula apply if $F(x)$ is the cumulative distribution function of a of a constant z in the interval $(a,b]$ and the function G has a discontinuity at the point z?

Lemma 104 *(Integration by Parts) If F, G are two monotone right continuous functions on the real line having* no common discontinuities*, then*

$$\int_{(a,b]} G(x)dF(x) \;=\; G(b)F(b) - G(a)F(a) - \int_{(a,b]} F(x)dG(x)$$

5.5.1 Convolutions

Consider two independent random variables X, Y, both having a discrete distribution. Suppose we wish to find the probability function of the sum $Z = X + Y$. Then

$$P[Z = z] = \sum_x P[X = x]P[Y = z - x] = \sum_x f_X(x)f_Y(z - x).$$

Similarly, if X, Y are independent absolutely continuous distributions with probability density functions f_X, f_Y respectively, then we find the probability density function of the sum $Z = X + Y$ by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx$$

In both the discrete and continuous case, we can rewrite the above in terms of the cumulative distribution function F_Z of Z. In either case,

$$F_Z(z) = E[F_Y(z - X)] = \int_{\Re} F_Y(z - x)dF_X(x)$$

We use the last form as a more general definition of a *convolution* between two cumulative distribution functions F, G. We define the *convolution* of two functions F and G to be

$$F * G(x) = \int_{-\infty}^{\infty} F(x - y)dG(y).$$

5.5.2 Properties.

(a) If F, G are cumulative distributions functions, then so is $F*G$ (see Problem 5.25)

(b) If F, G are cumulative distributions functions, $F*G = G*F$ (see Problem 6.3)

(c) If either F or G are absolutely continuous with respect to Lebesgue measure, then $F * G$ is absolutely continuous with respect to Lebesgue measure.

The convolution of two cumulative distribution functions $F * G$ represents the c.d.f of the sum of two independent random variables, one with c.d.f. F and the other with c.d.f. G. The next theorem says that if we have two independent

sequences X_n independent of Y_n and $X_n \Longrightarrow X$, $Y_n \Longrightarrow Y$, then the pair (X_n, Y_n) converge weakly to the joint distribution of two random variables (X, Y) where X and Y are independent. There is an easier proof available using the characteristic functions in Chapter 6.

Theorem 105 *If $F_n \Rightarrow F$ and $G_n \Rightarrow G$, then $F_n * G_n \Rightarrow F * G$.*

Proof.

First suppose that X, X_n, Y, Y_n have cumulative distribution functions given by F, F_n, G, G_n respectively and denote the set of points at which a function such as F is discontinuous by D_F. Recall that by Lemma 92, we may redefine the random variables Y_n and Y so that $Y_n \to Y$ almost surely. Now choose a point $z \notin D_{F*G}$. We wish to show that

$$F_n * G_n(z) \to F * G(z)$$

for all such z. Note that since $F * G$ is the cumulative distribution function of $X + Y$, $z \notin D_{F*G}$ implies that

$$0 = P[X + Y = z] \geq \sum_{x \epsilon D_F} P[Y = z - x] P[X = x].$$

so $P[Y = z - x] = 0$ whenever $P[X = x] > 0$, implying $P[Y \in z - D_F] = 0$. Therefore the set $[Y \notin z - D_F]$ has probability one, and on this set, since $z - Y_n \to z - Y$ almost surely, we also have $F_n(z - Y_n) \to F(z - Y)$ almost surely. It follows from the dominated convergence theorem (since $F_n(z - Y_n)$ is bounded above by 1) that

$$F_n * G_n(z) = E(F_n(z - Y_n)) \to E(F(z - Y)) = F * G(z)$$

■

5.6 Problems

1. Prove that if X, Y are independent random variables, $E(XY) = E(X)E(Y)$(Lemma 28). Are there are random variables X, Y such that $E(XY) = E(X)E(Y)$ but X, Y are not independent? What if X and Y only take two possible values?

2. Find two absolutely continuous random variables such that the joint distribution (X, Y) is not absolutely continuous.

3. If two random variables X, Y has joint probability density function $f(x, y)$ show that the joint density function of $U = X + Y$ and $V = X - Y$ is

$$f_{U,V}(u, v) = \frac{1}{2} f_{X,Y}(\frac{u + v}{2}, \frac{u - v}{2}).$$

4. If X_n is a sequence of non-negative random variables, show that the set of

$$\{\omega; X_n(\omega) \text{ converges}\} = \cap_{m=1}^{\infty} \cup_{N=1}^{\infty} \cap_{j=N}^{\infty} \cap_{n=N}^{\infty} [|X_n - X_j| \leq \frac{1}{m}]$$

5. Give an example of a sequence of random variables X_n defined on $\Omega = [0,1]$ which converges in probability but does not converge almost surely. Is there an example of the reverse (i.e. the sequence converges almost surely but not in probability)? If X_n is a Binomial (n,p) random variable for each n, in what sense does $n^{-1}X_n$ converge to p as $n \to \infty$?

6. Suppose that F_n is a sequence of c.d.f.'s converging to a right continuous function F at all continuity points of F. Prove that if the sequence has the property that for every $\epsilon > 0$ there exists $M < \infty$ such that $F_n(-M) + 1 - F_n(M) < \epsilon$ for all n, then F must be a proper cumulative distribution function.

7. Prove directly (using only the definitions of almost sure and weak convergence) that if X_n is a sequence of random variables such that $X_n \to X$ almost surely, then $X_n \Longrightarrow X$ (convergence holds weakly).

8. Prove that if X_n converges in distribution (weakly) to a constant $c > 0$ and $Y_n \Rightarrow Y$ for a random variable Y, then $Y_n/X_n \Rightarrow Y/c$. Show also that if $g(x,y)$ is a continuous function of (x,y), then $g(X_n, Y_n) \Rightarrow g(c, Y)$.

9. Prove that if $X_n \Rightarrow X$ then there exist random variables Y_n, Y with the same distribution as X_n, X respectively such that $Y_n \to Y$ a.s. (Lemma 32).

10. Prove that if X_n converges with probability 1 to a random variable X then it converges in distribution to X (Theorem 36).

11. Suppose $X_i, i = 1, 2, ...$ are independent identically distributed random variables with finite mean and variance $var(X_i) = \sigma^2$.Let $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Prove that

$$\frac{1}{n-1}\sum (X_i - \overline{X_n})^2 \to \sigma^2 \text{ almost surely as } n \to \infty.$$

12. A multivariate c.d.f. $F(\mathbf{x})$ of a random vector $\mathbf{X} = (\mathbf{X_1}, ..., \mathbf{X_n})$ is *discrete* if there are countably many points $\mathbf{y_j}$ such that

$$\sum_j P[\mathbf{X} = \mathbf{y_j}] = \mathbf{1}.$$

Prove that a multivariate distribution function is discrete if and only if its marginal distribution functions are all discrete.

13. Let $X_n, \quad n = 1, 2, \ldots$ be independent positive random variables all having a distribution with probability density function $f(x), \quad x > 0$. Suppose $f(x) \to c > 0$ as $x \to 0$. Define the random variable

$$Y_n = min(X_1, X_2, \ldots X_n).$$

   (a) Show that $Y_n \to 0$ in probability.

   (b) Show that nY_n converges in distribution to an exponential distribution with mean $1/c$.

14. *Continuity* Suppose X_t is, for each $t \in [a, b]$, a random variable defined on Ω. Suppose for each $\omega \in \Omega$, $X_t(\omega)$ is continuous as a function of t for $t \in [a, b]$.

    If for all $t \in [a, b]$, $|X_t(\omega)| \leq Y(\omega)$ for all $\omega \in \Omega$, where Y is some integrable random variable, prove that $g(t) = E(X_t)$ is a continuous function of t in the interval $[a, b]$.

15. *Differentiation under Integral.* Suppose for each $\omega \in \Omega$ that the derivative $\frac{d}{dt}X_t(\omega)$ exists and $|\frac{d}{dt}X_t(\omega)| \leq Y(\omega)$ for all $t \in [a, b]$, where Y is an integrable random variable. Then show that

$$\frac{d}{dt}E(X_t) = E[\frac{d}{dt}X_t]$$

16. Find the moment generating function of the *Gamma distribution* having probability density function

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, \ x > 0$$

    and show that the sum of n independent identically distributed Gamma (α, β) random variables is Gamma $(n\alpha, \beta)$. Use this fact to show that the moment generating function of the random variable

$$Z^* = \frac{\sum_{i=1}^n X_i - n\alpha\beta}{\sqrt{n\alpha\beta^2}}$$

    approaches the moment generating function of the standard normal distribution as $n \to \infty$ and thus that $Z^* \Rightarrow Z \sim N(0, 1)$.

17. Let $X_1, \ldots X_n$ be independent identically distributed random variables with the uniform distribution on the interval $[0, b]$. Show convergence in distribution of the random variable

$$Y_n = n \ \min(X_1, X_2, \ldots, X_n)$$

    and identify the limiting distribution.

18. Assume that the value of a stock at time n is given by

$$S_n = c(n)exp\{2X_n\}$$

where X_n has a binomial distribution with parameters (n, p) and $c(n)$ is a sequence of constants. Find $c(n)$ so that the expected value of the stock at time n is the risk-free rate of return e^{rn} . Consider the present value of a call option on this stock which has exercise price K.

$$V = e^{-rn}E\{max(S_n - K, 0)\}.$$

Show, using the weak convergence of the binomial distribution to the normal, that this expectation approaches a similar quantity for a normal random variable.

19. The usual student t-statistic is given by a form

$$t_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n}$$

where $\bar{X}_n$, s_n are the sample mean and standard deviation respectively. It is known that

$$z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

converges in distribution to a standard normal (N(0,1)) and that $s_n \to \sigma$ in probability. Show that t_n converges in distribution to the standard normal.

20. Let $X_1, X_2, \ldots, X_{2n+1}$ be independent identically distributed $U[0, 1]$ random variables . Define $M_n = median\ (X_1, X_2, \ldots, X_{2n+1})$. Show that $M_n \to \frac{1}{2}$ in probability and almost surely as $n \to \infty$.

21. We say that $X_n \to X$ in L_p for some $p \geq 1$ if

$$E(|X_n - X|^p) \to 0$$

as $n \to \infty$. Show that if $X_n \to X$ in L_p then $X_n \to X$ in probability. Is the converse true?

22. If $X_n \to 0$ in probability, show that there exists a subsequence n_k such that $X_{n_k} \to 0$ almost surely as $k \to \infty$.

23. Consider the product space $\{M \times N, \mathcal{F}, \pi)$ of two measure spaces $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$. Consider a set $E \in \mathcal{F}$ and define $E_y = \{x \in M; (x, y) \in E\}$.This is a measurable set in M. Now define the measure of these sets $g(y) = \mu(E_y)$. Show that the function $g(y)$ is a measurable function defined on N.

24. Consider the product space $\{M \times N, \mathcal{F}, \pi)$ of two measure spaces $(M, \mathcal{M}, \mu)$ and $(N, \mathcal{N}, \nu)$. Suppose we verify that for all $E = A \times B$,

$$\pi(E) = \int_N (\int_M I_E(x, y) d\mu) d\nu. \tag{5.2}$$

Prove that (5.2) holds for all $E \in \mathcal{F}$.

25. Prove that if F, G are cumulative distributions functions, then so is $F * G$.

26. Prove: If either F or G are absolutely continuous with respect to Lebesgue measure, then $F * G$ is absolutely continuous with respect to Lebesgue measure.

27. Prove

$$\begin{aligned} &P[a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n] \\ &\quad = F(b_1, b_2, \ldots b_n) - \sum_j F(b_1, ..., a_j, b_{j+1}, ...b_n) \\ &\quad + \sum_{i<j} F(b_1, ..., a_i, b_{i+1} ... a_j, b_{j+1}, ...b_n) - ... \end{aligned} \tag{5.3}$$

where $F(x_1, x_2, ...x_n)$ is the joint cumulative distribution function of random variables $X_1, ..., , X_n$.

28. Find functions F and G so that the integration by parts formula fails, i.e. so that

$$\int_{(a,b]} G(x) dF(x) \neq G(b)F(b) - G(a)F(a) - \int_{(a,b]} F(x) dG(x).$$