

Lecture Notes for Chapter 4

1 Averages

Two of the most important questions that arise in situations of uncertainty are (a) what is the likelihood of the occurrence of the possible outcomes, and (b) what are the anticipated consequences. Although related, these are two distinct questions. The probability space provides the mathematical structure to address the first question: the sample space defines the possible outcomes, the sigma field describes the class of distinguishable outcomes, and the probability measure describes the degree of belief or evidential support for the distinguishable outcomes. The probability space, however, does not itself provide an assessment of the consequences of the outcomes.

Consequences are addressed through random variables, which are measurable functions over the sample space. Although random variables provide a numerical measure of all possible consequences, but by themselves do not inform us of the *anticipated* consequences. Consider Pascal's wager, which is one of the earliest examples of this kind of issue: no matter how small we make the odds of God's existence, the payoff is infinite; infinite bliss for the saved and infinite misery for the damned. Under such conditions, rational self interest dictates that we sacrifice our certain but merely finite worldly pleasures to the uncertain but infinite prospect of salvation. What is interesting about Pascal's wager is that it is about neither the probability of God's existence nor the infinite bliss or misery that awaits the sinner, but it is about *the simultaneous consideration of the two!* It is apparent that, in order for one to take action, one must somehow combine the likelihood of outcomes with the consequences to form an assessment of the *anticipated* outcome.

One way to get a handle on anticipated outcomes is with an empirical argument. For example, suppose you are offered the following gamble. A coin is tossed, and you win two dollars if the coin lands heads and lose one dollar if it lands tails. Consider the following thought experiment. Suppose you were to play the game n times. You anticipate winning two dollars half of the time and losing one dollar half of the time. Your net anticipated outcome would then be $\frac{2n}{2} - \frac{n}{2} = \frac{n}{2}$, so you could reasonably conclude that, on average,

you would win $\frac{1}{2}$ dollars per coin toss. After taking into consideration your potential wins and losses, along with the chances for both, this is your *anticipated* winnings. Of course, you are not guaranteed to win 50 cents, or to even win anything. But if you choose to play such a game, you are willing to act *as if* the outcome of winning 50 cents were certain. This anticipation is not based on a watertight argument, but it might provide a degree of conviction sufficient to impel a prudent person such as yourself to take action.

Another way to address this problem is as follows. Assuming that the coin is fair, the probability of winning two dollars is $\frac{1}{2}$, which is also the probability of losing one dollar. You could simply weight each possible winning by the probability of winning, sum them up, and you would arrive at exactly the same conclusion: an expected 50-cent profit. As with the empirical analysis, the conclusion reached by this “one-shot” analysis might also be sufficient to impel you to take the gamble, in which case, as before, you would be acting *as if* you were certain to win 50 cents.

Although these two approaches yield the same result for this case, they are quite different. With the empirical approach, you would base your decision on the result of many trials (real or virtual, it doesn’t really matter), where as with the probabilistic approach you would base your decision on a formal mathematical analysis.

Are both points of view equivalent? In other words, are long-term empirical averages consistent with one-shot mathematical analyses? Consider another example. Given a discrete random process $\{X_i, i = 1, 2, \dots\}$, suppose you observe a particular realized discrete waveform x_i , and suppose you compute the empirical average

$$s_n = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

for $n = 1, 2, \dots$. Under what conditions, if any, would such a running average be consistent with a formal mathematical evaluation of the anticipated behavior of the process if you were required to predict the future outcome of the process? This is the topic or *ergodic theory*, of which the commonly known law of large numbers is only one example.

Let’s take a closer look at the above averaging exercise, but instead of working with realizations, let work with the random variables themselves. Let us first consider a special case where the random variables are all independent and identically distributed, and introduce the following statistical parlance.

Definition 1 A **population**, or **parent population**, is a derived probability space associ-

ated with a random variable X called the **population random variable**. The distribution function F_X of X is called the **distribution of the population**. The population is continuous or discrete according as X is continuous or discrete. \square

Definition 2 By **statistical sampling** we mean that we repeat a given experiment a number of times; the i th repetition involves the creation, mathematically, of a replica, or copy, of the population on which the random variable X_i is defined. The distribution of X_i is the same as the distribution of the population random variable X . The random variables X_1, X_2, \dots are called **sample random variables** or, sometimes, the **sample values**, or just **samples** of X . \square

The act of sampling can take many forms. Perhaps the simplest sampling procedure is that of *sampling with replacement*, where the distribution of the population is unchanged by the sampling. Unless stated otherwise, the sampling process will be done in a way such that the samples do not influence each other, in which case the sample random variables will be *i.i.d.*.

Definition 3 A function of the sample values of a population random variable X is called a **statistic**. \square

Note that statistics are random variables; they are not “data.” Conversely, numbers are not “statistics.” It is common to confuse the random variable with the values they assume. Just as $\sin x$ is a function but $\sin 2$ is a number, X is a random variable but $x = X(\omega)$ is a number. The former is a statistic, the latter is a realization.

Definition 4 Let X_1, X_2, \dots, X_n be sample values of a population random variable X with distribution function F_X . The statistic S_n defined by

$$S_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

is called the **sample mean**. Since it is the sum of n random variables, the sample mean is actually defined over the product sample space $\underbrace{\Omega \times \Omega \times \dots \times \Omega}_{n \text{ times}} = \Omega^n$. Each X_i is a mapping of a particular $\omega_i \in \Omega$ to \mathfrak{R} , thus S_n is a mapping of the vector $\omega = (\omega_1, \omega_2, \dots, \omega_n)$ to \mathfrak{R} . In order to keep notation simple, however, we will interpret each X_i as a function of the i th coordinate of the vector ω , i.e., $X_i(\omega) = X_i(\omega_i)$. Then we can write (2), more explicitly and

compactly, as

$$S_n(\omega) = \frac{1}{n} \sum_{i=1}^n X_i(\omega). \quad (3)$$

□

We emphasize that the sample mean is a random variable, in contrast to an empirical average, as given by (1). More specifically, the empirical average corresponds to *one realization* of the process (a waveform), while the sample mean represents the entire ensemble of possible waveforms.

Notice that the sample mean $\{S_n, n = 1, 2, \dots\}$ is also a random process. One question we may address is whether or not, as $n \rightarrow \infty$, S_n “converges” to anything in any sense. For example, is it true that $\lim_{n \rightarrow \infty} S_n = \lim_{n \rightarrow \infty} s_n$? If so, does it converge for all ω ? Another question we might ask is to consider the power in the error. That is, suppose we form the quantity $E_n = (S_n - s_n)^2$. Does E_n get small as n gets large? If so, in what sense?

The reason these questions are important is that in practice we may be limited to observing only a few realizations of a process, or perhaps only one realization. In what sense is this one waveform representative of the entire ensemble? What information about the ensemble can we infer from this one realization? For example, consider the coin-tossing experiment, and we wish to determine whether or not the coin is fair. If we limit our experimentation to one toss, we can tell nothing about fairness, but if we are allowed multiple tosses, we may suppose we eventually will be able to make some defensible inferences. But even so, that one sequence of tosses is only one waveform drawn from the ensemble of all possible waveforms. Do we really have a solid basis to suppose that inferences obtained from just one random waveform will be a reliable indicator of the behavior of the entire ensemble?

We require precise, mathematically defensible answers to these questions. It may be intuitively obvious to you that if you toss a fair coin a thousand times, you will get approximately 500 heads and 500 tails, but intuition is just not good enough, even in this simple case. To proceed, we must first formalize a precise definition of mathematical expectation, and then reconcile that definition with empirical averages.

2 Expectation

Definition 5 Let X be a discrete random variable with probability mass function $p_X(x)$. The **expected value** or **expectation** or **mean value** or **mean** of X is the quantity

$$\begin{aligned} EX &= \sum_{x:p_X(x)>0} xp_X(x) \\ &= \sum_i x_i p_X(x_i) \end{aligned}$$

Let X be a continuous random variable with density function f_X . The expected value of X is

$$EX = \int_{-\infty}^{\infty} xf_X(x)dx.$$

More generally, for *any* distribution, the expected value of a random variable is the Stieltjes integral

$$EX = \int_{-\infty}^{\infty} x dF_X(x).$$

This last definition, although more general and therefore more useful for theoretical development, is actually less useful in applications, because of the cumbersomeness of computing Stieltjes integrals. \square

Definition 6 Let X be a random variable; the n th **moment** is the expectation of the n th power of X :

$$EX^n = \int_{-\infty}^{\infty} x^n f_X(x)dx.$$

Of particular importance are the second moment

$$EX^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx$$

and the **second central moment**, or **variance**

$$\text{Var}(X) = E(X - EX)^2 = \int_{-\infty}^{\infty} (x - EX)^2 f_X(x)dx.$$

\square

Theorem 1 Let X be a random variable with finite first and second moments, and let c be an arbitrary constant. Then

1. $\text{Var}(X) = EX^2 - (EX)^2$

2. $\text{Var}(X) \geq 0$
3. $\text{Var}(c) = 0$
4. $\text{Var}(X + c) = \text{Var}(X)$
5. $\text{Var}(cX) = c^2 \text{Var}(X)$

Proof

1. By direct evaluation, we have

$$E(X - EX)^2 = EX^2 - 2EX \cdot EX + (EX)^2 = EX^2 - (EX)^2.$$

2. That $\text{Var}(X) \geq 0$ is obvious
3. This follows from the fact that $Ec = c$.
4. This follows from the fact that $X + c - E(X + c) = X - EX$.
5. By direct evaluation, we have

$$E(cX - E(cX))^2 = E(c^2(X - E(X))^2) = c^2 \int_{-\infty}^{\infty} (x - EX)^2 f_X(x) dx = c^2 \text{Var}(X).$$

□

2.1 Some Examples

Example 1 The St. Petersburg Paradox. *Pierre and Paul play a coin toss game. If the coin comes up heads on the first toss, Pierre agrees to pay Paul \$2; if heads does not turn up until the second toss, Paul receives \$4; if not until the third toss, \$8, and so on, so that if heads does not turn up until the n th toss, Pierre pays Paul $\$2^n$. Assuming that the coin is fair and that the tosses are independent, then, according to the geometric distribution, the probability of the first head turning up on the n th toss is $\frac{1}{2^n}$. Let $X = 2^n$ represent the payoff if the first head turns up on the n th thoss. Then the expected value is*

$$\begin{aligned} EX &= \left(\frac{1}{2} \times \$2\right) + \left(\frac{1}{4} \times \$4\right) + \left(\frac{1}{8} \times \$8\right) + \cdots + \left(\frac{1}{2^n} \times \$2^n\right) + \cdots \\ &= 1 + 1 + 1 + \cdots = \infty. \end{aligned}$$

Since there is a small but finite chance that even a fair coin will produce an unbroken run of tails of arbitrary finite length, and since the payoffs increase in proportion to the decreasing probabilities of such an event, the expectation is infinite.

Example 2 The Rotating Flashlight. Suppose a narrow beam flashlight is spun around its center, which is located a unit distance from the x -axis. When the flashlight stops spinning, consider the point X at which the beam intersects the x -axis (if the beam is not pointing toward the x -axis, repeat the experiment). The point X is determined by the angle Θ between the flashlight and the y -axis. We assume that Θ is a random variable that is uniformly distributed over the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Thus,

$$F_{\Theta}(\theta) = \frac{\theta + \frac{\pi}{2}}{\pi}$$

This situation is illustrated in Figure 1

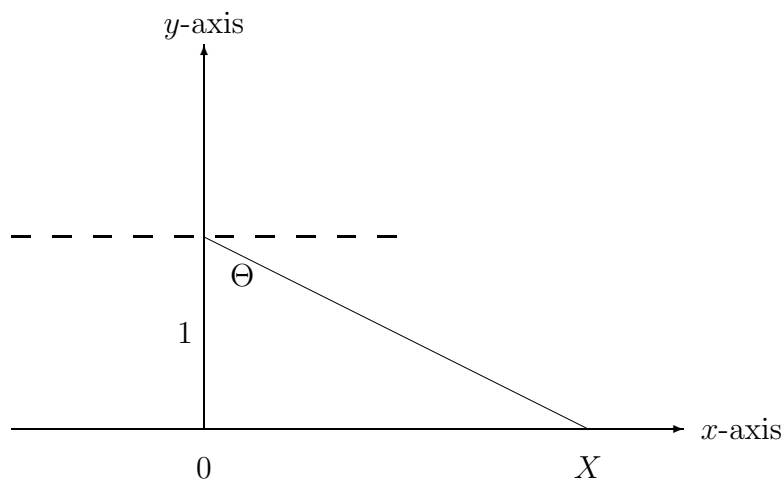


Figure 1: An experiment characterized by the Cauchy distribution.

The distribution function of $X = \tan \Theta$ is thus given by

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(\tan \Theta \leq x) \\ &= P(\Theta \leq \tan^{-1} x) \\ &= F_{\Theta}(\tan^{-1} x) \\ &= \frac{1}{2} + \frac{1}{\pi} \tan^{-1} x \end{aligned}$$

Differentiating,

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

This is called the **Cauchy** distribution. It is straightforward to see that this is a legitimate density, since it is non-negative and

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx = \frac{1}{\pi} \tan^{-1}(x) \Big|_{-\infty}^{\infty} = 1.$$

However,

$$EX = \int_{-\infty}^0 \frac{1}{\pi} \frac{x}{1+x^2} + \int_0^{\infty} \frac{1}{\pi} \frac{x}{1+x^2} = \frac{1}{2} \log(1+x^2) \Big|_{-\infty}^0 + \frac{1}{2} \log(1+x^2) \Big|_0^{\infty} = -\infty + \infty$$

which is not well-defined. Thus, the EX does not exist.

2.2 Some Properties of Expectation

Theorem 2 Let $X = c$, with probability one, where c is a constant. Then it is immediate that

$$EX = cP(X = c) = c.$$

If $P(X \geq 0) = 1$, then $EX \geq 0$. To see, simply observe that

$$EX = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x f_X(x) dx \geq 0.$$

If $P(X > Y) = 1$, then $EX \geq EY$. To see, we invoke sign preservation and superposition to assert that $0 \leq E(X - Y) = EX - EY$.

2.3 A Brief Primer On Lebesgue Theory

Although a detailed development of Lebesgue theory is properly the subject of a course in real analysis, it may be helpful to at least indicate how such an integral is constructed in contrast to the way the more familiar Riemann integral is constructed.

2.3.1 The Riemann Integral

Let $f(x)$ be a function defined over the interval (a, b) . To construct the Riemann integral, we form a partition of (a, b) as

$$a = x_0 < x_1 < \cdots < x_n = b$$

and form the sum

$$\sum_{i=1}^n f(x'_i)(x_i - x_{i-1}),$$

where x'_i is an arbitrary point in $[x_{i-1}, x_i]$. We define the *mesh size* ρ as $\rho = \max_i(x_i - x_{i-1})$.

We then take the limit as, simultaneously, $n \rightarrow \infty$ and $\rho \rightarrow 0$, and define the Riemann integral as

$$\int_a^b f(x) dx = \lim_{\rho \rightarrow 0, n \rightarrow \infty} \sum_{i=1}^n f(x'_i)(x_i - x_{i-1}). \quad (4)$$

Important to the well-formedness of this definition is that this limit does not depend on the particular sequence $\{x'_i\}$. If it did, then we could have a different value for each sequence, which would render the limit undefined. It is easy to construct examples of functions for which the Riemann integral is not well-defined. Consider the so-called salt-and-pepper function defined over the unit interval as

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases} .$$

It is easy to see that if we take x'_i as rational numbers, we get a value of unity for the integral, but if we take x'_i as irrational numbers, we get zero for the integral. Although this is a somewhat pathological example, problems such as this make it difficult to take limiting operations with Riemann integrals.

2.3.2 The Lebesgue Integral

The development of the general theory of Lebesgue integration requires some additional definitions and notation, but if we restrict our attention to the unit interval case, we have nearly enough mathematical machinery to define the Lebesgue integral. To begin, let us consider the probability space $([0, 1], \mathcal{B}([0, 1]), \mu)$, where $\mathcal{B}([0, 1])$ is the Borel field over the unit interval and μ is a probability over this field. such that, for every interval (a, b) the probability is the length of the interval (this corresponds to the uniform distribution); Thus,

$$\mu((a, b)) = b - a.$$

Since μ is a probability, we also have that, for any disjoint countable collection of Borel sets $\{A_i, i = 1, 2, \dots\}$,

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i).$$

This particular probability measure can be viewed as a *generalized length*; that is, for any disjoint set of intervals, the measure of their union is the sum of the lengths of the individual intervals. Clearly, the length of a singleton set $\{x\}$ is zero. This particular set function is called *Lebesgue measure*. (Lebesgue measure is actually defined on a larger class of sets than the Borel field. The Lebesgue measurable sets include, in addition to the Borel sets, all subsets of Borel sets that have Lebesgue measure zero. This technical note need not overly concern us in this discussion.)

We construct the Lebesgue integral as follows. Rather than partition the domain space (the x axis), as is done with the Riemann definition, let us partition the range space (the

y axis). In the interest of brevity, let us assume that the function f is bounded, that is, there exists a real number M such that $|f(x)| \leq M$ for all $x \in [0, 1]$ (this restriction can be relaxed). Now let us define a sequence $\{y_i\}$ such that

$$-M = y_0 < y_1 < \cdots < y_n = M.$$

Now let $I_i = (y_{i-1}, y_i)$ denote the interval of length $y_i - y_{i-1}$, and form the sum

$$\sum_{i=1}^n y_i \mu(f^{-1}(I_i)),$$

that is, we compute the probability of the inverse images of each interval of the partition, weight it by the upper value of the interval, and sum over all partitions. We then pass to the limit as the mesh size goes to zero. The resulting quantity is the Lebesgue integral:

$$\int_0^1 f(x) d\mu(x) = \lim_{\rho \rightarrow 0, n \rightarrow \infty} \sum_{i=1}^n y_i \mu(f^{-1}(I_i)).$$

We have ignored the technical details that are necessary to form a rigorous proof, but the essentials of the construction of the Lebesgue integral are in place. It can be shown that if a function is Riemann integrable, then it is also Lebesgue integrable, and that the values of the two integrals coincide. The converse is not true, however: not all functions that are Lebesgue integrable are Riemann integrable. For example, it can be shown that the set of rationals has probability zero (this can perhaps be most easily seen by noting that the Lebesgue measure of a singleton set is zero, and the rational numbers can be expressed as the disjoint union of a countable number of singleton sets, and that the sum of a countable number of zeros is still zero), and, consequently, the set of irrational numbers on the unit interval has unit generalized length. Thus, the Lebesgue integral of the salt-and-pepper function is unambiguously defined as zero.

One of the main advantages of the Lebesgue integral over the Riemann integral is that with the former it is possible to define precise conditions for when it is permissible to commute the operations of taking expectation and passing to the limit. The two theorems presented below (without formal proof) are among the most important results of modern integration theory.

Theorem 3 Lebesgue's Monotone Convergence Theorem. *Let $\{X_n, n = 0, 1, \dots\}$ be a sequence of non-negative random variables all defined over the same probability space*

(Ω, \mathcal{A}, P) such that $X_n(\omega) \geq X_{n-1}(\omega)$ for all n and for all $\omega \in \Omega$ except possibly for a set of probability zero, and suppose that there exists a function X such that

$$X(\omega) = \lim_{n \rightarrow \infty} X_n(\omega)$$

for all ω except possibly in a set of probability zero. Then X is a random variable (i.e., the inverse images of Borel sets are elements of \mathcal{A} and

$$E(\lim_{n \rightarrow \infty} X_n) = \lim_{n \rightarrow \infty} EX_n.$$

Theorem 4 Lebesgue's Dominated Convergence Theorem. Let $\{X_n, n = 0, 1, \dots\}$ be a sequence of random variables all defined over the same probability space (Ω, \mathcal{A}, P) , suppose there exists a random variable X such that $\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega$ except possibly for a set of probability zero, and suppose that there exists a random variable Y such that $|X(\omega)| \leq Y(\omega)$ for all $\omega \in \Omega$ except possibly for a set of probability zero and such that EY exists. Then

$$\lim_{n \rightarrow \infty} EX_n = EX.$$

These theorems provide justification for interchanging the operations of limit and integration; i.e., when the limit of the integral equals the integral of the limit. The proofs of these theorems are not difficult, but they require more mathematical development than is convenient to offer in this class. These theorems are extremely useful for establishing such results as the central limit theorem and the law of large numbers.

2.4 The Fundamental Theorem of Expectation

Let $Y = g(X)$ for some Borel measurable function g , and suppose we wish to compute the expectation of Y .

Theorem 5 The fundamental theorem of expectation (discrete case). Let X be a discrete random variable with probability mass function p_X . Given a function $g : \mathbb{R} \rightarrow \mathbb{R}$, the resulting random variable $Y = g(X)$ has expectation

$$EY = Eg(X) = \sum_x g(x)p_X(x).$$

Proof Let $\{x_1, x_2, \dots\}$ denote the discrete values that X assumes, and define the sequence $\{z_1, z_2, \dots\}$ with $z_i = g(x_i)$ being the discrete values assumed by Y . Note that, unless the

function g is one-to-one, not all of the z_i 's will be distinct (this would be the case, for example, if $g(x) = x^2$). To account for this possibility, we may have to eliminate duplications in the sequence $\{z_1, z_2, \dots\}$, yielding the sequence $\{y_1, y_2, \dots\}$, where all of the y_i 's are distinct.

For each y_j let $A_j = \{x_{j1}, \dots, x_{jn_j}\}$ denote the subset of the x_i 's that map to y_j , that is,

$$A_j = \{x : g(x) = y_j\}.$$

If g is one-to-one, then $A_j = \{x_j\}$. Then

$$p_Y(y_j) = P(Y = y_j) = \sum_{x \in A_j} p_X(x). \quad (5)$$

Once we have the probability mass function for Y available to us, we may compute the expectation of Y as

$$EY = \sum_j y_j p_Y(y_j). \quad (6)$$

Substituting (5) into (6), we get

$$\begin{aligned} EY &= \sum_j y_j p_Y(y_j) \\ &= \sum_j y_j \sum_{x \in A_j} p_X(x) \\ &= \sum_j \sum_{x \in A_j} y_j p_X(x) \\ &= \sum_j \sum_{x \in A_j} g(x) p_X(x) \quad (\text{since } x \in A_j \text{ implies } g(x) = y_j) \\ &= \sum_x g(x) p_X(x) \quad (\text{since the } A_j\text{'s are disjoint}). \end{aligned}$$

□

To prove the continuous case we need the following lemma.

Lemma 1 *Let X be a random variable. Then*

$$EX = \int_0^\infty P(X > x) dx - \int_0^\infty P(X < -x) dx \quad (7)$$

$$= \int_0^\infty (1 - F_X(x)) dx - \int_{-\infty}^0 F_X(x) dx \quad (8)$$

Proof Let us examine the first term on the right hand side of (8).

$$\int_0^\infty P(x > x) dx = \int_0^\infty \left[\int_x^\infty f_X(y) dy \right] dx \quad (9)$$

The obvious thing to do when one has a double integral such as this is to reverse the order of integration. This operation is complicated, however, by the fact that the lower limit of the inner integral is a function of the variable of integration of the inner integral. Things might be a little more clear if we were to introduce some new notation that more clearly demonstrates what is going on.

Recall the **indicator** function, i.e.,

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}.$$

Let us use this function to rewrite the right hand side of (9), namely,

$$\int_0^\infty \left[\int_x^\infty f_X(y) dy \right] dx = \int_0^\infty \left[\int_0^\infty f_X(y) I_{\{x:x < y\}}(x) dy \right] dx$$

Changing the order of integration obtains

$$\begin{aligned} \int_0^\infty \left[\int_0^\infty f_X(y) I_{\{x:x < y\}}(x) dy \right] dx &= \int_0^\infty \left[\int_0^\infty f_X(y) I_{\{x:x < y\}}(x) dx \right] dy \\ &= \int_0^\infty f_X(y) \left[\int_0^\infty I_{\{x:x < y\}}(x) dx \right] dy \\ &= \int_0^\infty f_X(y) \left[\int_0^y dx \right] dy \\ &= \int_0^\infty y f_X(y) dy. \end{aligned}$$

Now let us apply a similar analysis to the second term on the right hand side of (8).

$$\begin{aligned} \int_0^\infty P(X < -x) dx &= \int_0^\infty \left[\int_{-\infty}^{-x} f_X(y) dy \right] dx \\ &= \int_0^\infty \left[\int_{-\infty}^0 f_X(y) I_{\{x:-x > y\}}(x) dy \right] dx \\ &= \int_{-\infty}^0 \left[\int_0^\infty f_X(y) I_{\{x:-x > y\}}(x) dx \right] dy \\ &= \int_{-\infty}^0 f_X(y) \left[\int_0^\infty I_{\{x:-x > y\}}(x) dx \right] dy \\ &= \int_{-\infty}^0 f_X(y) \left[\int_0^{-y} dx \right] dy \\ &= - \int_{-\infty}^0 y f_X(y) dy. \end{aligned}$$

□

Theorem 6 The fundamental theorem of expectation (*continuous case*). Let X be a continuous random variable with probability density function f_X . Given a function $g : \mathfrak{R} \rightarrow \mathfrak{R}$, the resulting random variable $Y = g(X)$ has expectation

$$EY = Eg(X) = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

Proof Applying Lemma 1, we obtain

$$Eg(X) = \int_0^{\infty} P(g(X) > y)dy - \int_0^{\infty} P(g(X) < -y)dy, \quad (10)$$

Examination of the first term on the right hand side of (10) yields

$$\begin{aligned} \int_0^{\infty} P(g(X) > y)dy &= \int_0^{\infty} \left[\int_{\{x:g(x)>y\}} f_X(x)dx \right] dy \\ &= \int_0^{\infty} \left[\int_{-\infty}^{\infty} f_X(x)I_{\{y:0 \leq y < g(x)\}}(y)dx \right] dy \\ &= \int_0^{\infty} \left[\int_{-\infty}^{\infty} f_X(x)I_{\{y:y < g(x)\}}(y)I_{\{x:g(x)>0\}}(x)dx \right] dy \\ &= \int_{-\infty}^{\infty} f_X(x)I_{\{x:g(x)>0\}}(x) \left[\int_0^{\infty} I_{\{y:y < g(x)\}}(y)dy \right] dx \\ &= \int_{\{x:g(x)>0\}} f_X(x) \left[\int_0^{g(x)} dy \right] dx \\ &= \int_{\{x:g(x)>0\}} g(x)f_X(x)dx \end{aligned}$$

Examination of the second term on the right hand side of (10) yields

$$\begin{aligned} \int_0^{\infty} P(g(X) < -y)dy &= \int_0^{\infty} \left[\int_{\{x:g(x)<-y \leq 0\}} f_X(x)dx \right] dy \\ &= \int_0^{\infty} \left[\int_{-\infty}^{\infty} f_X(x)I_{\{y:y < -g(x)\}}(y)I_{\{x:g(x) \leq 0\}}(x)dx \right] dy \\ &= \int_{-\infty}^{\infty} f_X(x)I_{\{x:g(x) \leq 0\}}(x) \left[\int_0^{\infty} I_{\{y:y < -g(x)\}}(y)dy \right] dx \\ &= \int_{\{x:g(x) \leq 0\}} f_X(x) \left[\int_0^{-g(x)} dy \right] dx \\ &= - \int_{\{x:g(x) \leq 0\}} g(x)f_X(x)dx \end{aligned}$$

Putting things together, we have that

$$\begin{aligned}
 Eg(X) &= \int_0^\infty P(g(X) > y)dy - \int_0^\infty P(g(X) < -y)dy \\
 &= \int_{\{x:g(x)>0\}} g(x)f_X(x)dx + \int_{\{x:g(x)\leq 0\}} g(x)f_X(x)dx \\
 &= \int_{\{x:-\infty < g(x) < \infty\}} g(x)f_X(x)dx \\
 &= \int_{-\infty}^\infty g(x)f_X(x)dx
 \end{aligned}$$

□

We may extend these results to the multivariate case. For example, with two random variables, we have

Theorem 7 *Let X and Y be discrete random variables with joint probability mass function $p_{XY}(x, y)$ and let $g : \mathfrak{R}^2 \rightarrow \mathfrak{R}$. Then*

$$Eg(X, Y) = \sum_x \sum_y g(x, y)f_{XY}(x, y).$$

Let X and Y be continuous random variables with joint probability density function $f_{XY}(x, y)$ and let $g : \mathfrak{R}^2 \rightarrow \mathfrak{R}$. Then

$$Eg(X, Y) = \int_{-\infty}^\infty \int_{-\infty}^\infty g(x, y)f_{XY}(x, y)dxdy.$$

Proof The proof follows essentially the same lines as the scalar case but is more notationally complicated. We will omit the details. □

With this theorem in place we are in a position to present an alternate proof for the superposition property of expectation.

Theorem 8 *Let X and Y be random variables. Then*

$$E(aX + bY) = aEX + bEY.$$

Proof We demonstrate only the continuous case; the discrete case is virtually identical.

$$\begin{aligned}
 E(aX + bY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f_{XY}(x, y) dx dy \\
 &= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) dx dy + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x, y) dx dy \\
 &= a \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x, y) dy dx + b \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy \\
 &= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} y f_Y(y) dy \\
 &= aEX + bEY.
 \end{aligned}$$

□

Theorem 9 *The expectation of a finite sum of random variables is the sum of their expectations; that is, let X_1, X_2, \dots, X_n be jointly distributed random variables. Then*

$$E(X_1 + X_2 + \dots + X_n) = EX_1 + EX_2 + \dots + EX_n.$$

Proof The result follows by induction. □

3 Correlation and Covariance

Definition 7 Let X and Y be random variables. The **correlation** of X with Y , denoted $\text{Cor}(X, Y)$, is the expectation of their product; that is,

$$\text{Cor}(X, Y) = E(XY).$$

We say that X and Y are **uncorrelated** if the expectation of the product is the product of the expectations; or

$$E(XY) = EXEY.$$

□

Theorem 10 *If X and Y are independent random variables with expectations EX and EY , respectively, then X and Y are uncorrelated.*

$$E(XY) = EXEY$$

Proof By the fundamental theorem and the fact that the joint density factors into the product of the marginals,

$$\begin{aligned}
 E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy \\
 &= \int_{-\infty}^{\infty} x f_X(x) dx \int_{-\infty}^{\infty} y f_Y(y) dy \\
 &= EXEY.
 \end{aligned}$$

□

Theorem 11 *The expectation of a finite product of independent random variables is the product of their expectations; that is, let X_1, X_2, \dots, X_n be independent random variables. Then*

$$E(X_1 X_2 \dots X_n) = EX_1 EX_2 \dots EX_n.$$

Proof The result follows by induction. □

Let us now refine the notion of correlation by centering the random variables with respect to their mean values and then computing the correlation of the result.

Definition 8 Let X and Y be random variables. The **covariance** between X and Y , denoted $\text{Cov}(X, Y)$, is given by

$$\text{Cov}(X, Y) = E(X - EX)(Y - EY).$$

□

Theorem 12

$$\text{Cov}(X, Y) = \text{Cor}(X, Y) - EXEY.$$

Proof Expanding the defining expression, we obtain

$$\begin{aligned}
 \text{Cov}(X, Y) &= E(XY - YEX - XEY + EXEY) \\
 &= E(XY) - EYEX - EXEY + EXEY \\
 &= E(XY) - EXEY.
 \end{aligned}$$

□

Theorem 13

- (i) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
- (ii) $\text{Cov}(X, X) = \text{Var}(X)$.
- (iii) $\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$.
- (iv) $\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$.
- (v) If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

Proof Parts (i) and (ii) are obvious from the definition. Part (iii) follows from

$$\text{Cov}(aX, Y) = E(aXY) - E(aX)EY = a(E(XY) - EXEY).$$

We establish (iv) as follows, where $\mu_i = EX_i$ and $\nu_j = EY_j$.

$$\begin{aligned} \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) &= E\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)\left(\sum_{j=1}^m Y_j - \sum_{j=1}^m \nu_j\right)\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu_i) \sum_{j=1}^m (Y_j - \nu_j)\right] \\ &= E\left[\sum_{i=1}^n \sum_{j=1}^m (X_i - \mu_i)(Y_j - \nu_j)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^m E[(X_i - \mu_i)(Y_j - \nu_j)] \end{aligned}$$

To establish (v), we simply apply the fact that independence implies uncorrelation. \square

Theorem 14

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j)$$

Proof

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^n X_i\right) &= \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{Cov}(X_i, X_j). \end{aligned}$$

Since each pair of indices i, j , $j \neq i$ appears twice in the double summation, the conclusion follows. \square

Definition 9 Let X and Y be random variables. The **correlation coefficient**, denoted $\rho(X, Y)$ is given by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.$$

\square

Theorem 15 Let X and Y be random variables with correlation coefficient $\rho(X, Y)$. Then

$$-1 \leq \rho(X, Y) \leq 1.$$

Proof Let $\text{Var}(X) = \sigma_x^2$ and $\text{Var}(Y) = \sigma_y^2$. Then

$$\begin{aligned} 0 &\leq \text{Var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}\right) \\ &= \frac{\text{Var}(X)}{\sigma_x^2} + \frac{\text{Var}(Y)}{\sigma_y^2} + 2\frac{\text{Cov}(X, Y)}{\sigma_x\sigma_y} \\ &= 2[1 + \rho(X, Y)], \end{aligned}$$

implying that $-1 \leq \rho(X, Y)$. Also,

$$\begin{aligned} 0 &\leq \text{Var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}\right) \\ &= \frac{\text{Var}(X)}{\sigma_x^2} + \frac{\text{Var}(Y)}{\sigma_y^2} - 2\frac{\text{Cov}(X, Y)}{\sigma_x\sigma_y} \\ &= 2[1 - \rho(X, Y)], \end{aligned}$$

implying that $\rho(X, Y) \leq 1$. \square

Let us now extend the notion of uncorrelation and establish that arbitrary functions of independent random variables are also uncorrelated.

Theorem 16 Let X and Y be independent random variables. Then for any Borel functions g and h ,

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Proof We prove only the continuous case; the proof in discrete case is similar.

$$\begin{aligned}
 E[g(X)h(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f_{XY}(x,y)dxdy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\
 &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy \\
 &= E[g(X)]E[h(Y)].
 \end{aligned}$$

□

It is not true, however, that if X and Y are uncorrelated, then they are independent, as the following counterexample establishes.

Example 3 Let the random variable Θ be uniformly distributed over the interval $[0, 2\pi]$, and define the random variables

$$\begin{aligned}
 X &= \cos \Theta \\
 Y &= \sin \Theta.
 \end{aligned}$$

We have

$$\begin{aligned}
 EX &= \frac{1}{2\pi} \int_0^{2\pi} \cos \theta d\theta = 0 \\
 EY &= \frac{1}{2\pi} \int_0^{2\pi} \sin \theta d\theta = 0,
 \end{aligned}$$

and

$$E(XY) = \frac{1}{2\pi} \int_0^{2\pi} \sin \theta \cos \theta d\theta = 0,$$

Thus X and Y are uncorrelated, but they are clearly not independent, since they are both functions of Θ .

We do, however, have the following fact about independence and uncorrelation.

Example 4 The bivariate normal population Let X and Y be bivariate normal random variables with joint density function

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-m_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-m_X}{\sigma_X}\right)\left(\frac{y-m_Y}{\sigma_Y}\right) + \left(\frac{y-m_Y}{\sigma_Y}\right)^2\right]}, \quad (11)$$

where $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$.

Now let us compute the expectation

$$E \left[\left(\frac{X - m_X}{\sigma_X} \right) \left(\frac{Y - m_Y}{\sigma_Y} \right) \right].$$

Letting $Z = \frac{X - m_X}{\sigma_X}$ and $W = \frac{Y - m_Y}{\sigma_Y}$ and using the identity $z^2 - 2\rho zw + w^2 = (z - \rho w)^2 + (1 - \rho^2)w^2$, we obtain

$$\begin{aligned} E(ZW) &= \frac{1}{2\pi(1 - \rho^2)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z w e^{-\frac{1}{2}(z - \rho w)^2 + (1 - \rho^2)w^2} dz dw \\ &= \frac{1}{2\pi(1 - \rho^2)} \int_{-\infty}^{\infty} w e^{-\frac{1}{2}(1 - \rho^2)w^2} \underbrace{\left[\int_{-\infty}^{\infty} z e^{-\frac{1}{2}(z - \rho w)^2} dz \right]}_{\sqrt{2\pi}\rho w} dw \\ &= \frac{\sqrt{2\pi}\rho}{2\pi(1 - \rho^2)} \int_{-\infty}^{\infty} w^2 e^{-\frac{1}{2}(1 - \rho^2)w^2} dw \end{aligned}$$

Now make the change of variable $Q = (\sqrt{1 - \rho^2})W$ to obtain

$$E(ZW) = \frac{\rho}{\sqrt{2\pi}} \int_{-\infty}^{\infty} q^2 e^{-\frac{1}{2}q^2} dq = \rho.$$

since $Q \sim \mathcal{N}(0, 1)$.

We thus have the following result:

Theorem 17 *If X and Y are normally distributed random variables they are uncorrelated if and only if they are independent.*

3.1 Interpreting Correlation

We have seen that, for the bivariate normal distribution, the correlation coefficient can be viewed as a measure of the degree of dependency that exists between the two random variables. It is tempting to infer, therefore, that this interpretation should extend to other distributions as well. But as example 3 illustrates, highly dependent random variables may be uncorrelated. Thus, we must look deeper to gain appropriate insight as to how to interpret the correlation coefficient. To gain this insight, let us return to the proof of Theorem 15, specifically, the results

$$0 \leq \text{Var} \left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y} \right) = 2[1 + \rho(X, Y)] \quad (12)$$

and

$$0 \leq \text{Var} \left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y} \right) = 2[1 - \rho(X, Y)]. \quad (13)$$

Let us examine (13), and suppose that $\rho(X, Y) = 1$. In this case, the random variable $\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}$ has zero variance. As will be shown in the next chapter, a random variable that has zero variance is constant with probability one. In other words, if $\rho(X, Y) = 1$, we must have

$$\frac{X}{\sigma_x} - \frac{Y}{\sigma_y} = \text{constant},$$

or

$$Y = \frac{\sigma_y}{\sigma_x} X + a$$

for some constant a . By a similar argument using (12), if $\rho(X, Y) = -1$, then we must have

$$Y = -\frac{\sigma_y}{\sigma_x} X + a.$$

Theorem 18 $|\rho(X, Y)| = 1$ if and only if X and Y are related by an affine transformation.

Proof We have already established the “only if” portion of this theorem. To establish the converse, let $Y = bX + a$. Then

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - \mu_x)(bX + a - b\mu_x - a)] \\ &= bEX^2 - 2b\mu_x^2 + b\mu_x^2 \\ &= bEX^2 - b\mu_x^2 \\ &= b \text{Var}(X). \end{aligned}$$

But $\text{Var}(Y) = b^2 \text{Var}(X)$, so

$$\rho(X, Y) = \frac{b \text{Var}(X)}{\sqrt{\text{Var}(X)} \sqrt{b^2 \text{Var}(X)}} = \pm 1,$$

where the coefficient is $+1$ if $b > 0$ and -1 if $b < 0$. □

Thus, an interpretation of the correlation coefficient is the *degree of linearity* between X and Y . The closer $|\rho(X, Y)|$ is to unity, the closer X and Y are to being related by an affine transformation. To relate this interpretation to the bivariate normal distribution, we note, in (11), that when $\rho \rightarrow \pm 1$, the denominator tends to zero and the exponent tends to $-\infty$, yielding an indeterminate form. The distribution is said to be **singular**. Let's rewrite the bivariate normal density in matrix form. Define the vector $\mathbf{x} = [x, y]^T$, $\mathbf{m} = [\mu_x, \mu_y]^T$, and

the matrix

$$\begin{aligned}
 A &= E \left[\begin{bmatrix} X - \mu_x \\ Y - \mu_y \end{bmatrix} \begin{bmatrix} X - \mu_x, Y - \mu_y \end{bmatrix} \right] \\
 &= \begin{bmatrix} E(X - \mu_x)^2 & E(X - \mu_x)(Y - \mu_y) \\ E(Y - \mu_y)(X - \mu_x) & E(Y - \mu_y)^2 \end{bmatrix} \\
 &= \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.
 \end{aligned}$$

Then we may rewrite (11) as

$$f_{XY}(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det A}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T A^{-1}(\mathbf{x}-\mathbf{m})},$$

where

$$A^{-1} = \begin{bmatrix} \frac{\sigma_y^2}{\sigma_x^2\sigma_y^2(1-\rho^2)} & \frac{-\rho\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2(1-\rho^2)} \\ \frac{-\rho\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2(1-\rho^2)} & \frac{\sigma_x^2}{\sigma_x^2\sigma_y^2(1-\rho^2)} \end{bmatrix}$$

Clearly, when $\rho = 1$, the matrix A becomes singular. The singularity of A means that linear dependency exists between the columns of A .

We see that uncorrelation is a substantially different property than independence. We have shown that if X and Y are independent, then any functions g and h of them are uncorrelated. We now address the question of whether there exist conditions such that uncorrelation implies independence. The following theorem answers that question.

Theorem 19 *Let X and Y be random variables. Suppose that X and Y are uncorrelated for all functions g and h with finite expectations, that is,*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

for all g and all h , assuming that all such expectations exist. Then X and Y are independent.

Proof We provide a rigorous proof for the discrete case and a rather arm-waving demonstration for the continuous case.

Discrete case Let a be any number in the range space of X and b any number in the range space of Y , and define

$$\begin{aligned}
 g(x) &= I_a(x) \\
 h(y) &= I_b(y).
 \end{aligned}$$

where I_a and I_b are **indicator functions** that is,

$$I_a(x) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} E[g(X)h(Y)] &= \sum_x \sum_y I_a(x)I_b(y)p_{XY}(x, y) \\ &= I_a(a)I_b(b)p_{XY}(a, b) \\ &= p_{XY}(a, b). \end{aligned}$$

But, since $g(X)$ and $h(Y)$ are uncorrelated,

$$\begin{aligned} E[g(X)h(Y)] &= E[g(X)]E[h(Y)] \\ &= \sum_x I_a(x)p_X(x) \sum_y I_b(y)p_Y(y) \\ &= p_X(a)p_Y(b). \end{aligned}$$

Thus

$$p_{XY}(a, b) = p_X(a)p_Y(b)$$

for all a in the range space of X and all b in the range space of Y . This proves independence.

Demonstration for Continuous case Let a be any number in the range space of X and b any number in the range space of Y , and define

$$\begin{aligned} g(x) &= \delta(x - a) \\ h(x) &= \delta(x - b). \end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function. The reason I term this a “demonstration” is that Dirac delta functions are not allowed with rigorous proofs; but as engineers, we know their value and can trust them most of the time.

Continuing, we have

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - a)\delta(y - b)f_{XY}(x, y)dx dy \\ &= f_{XY}(a, b) \end{aligned}$$

for all a and b . But, since $g(X)$ and $h(Y)$ are uncorrelated,

$$\begin{aligned} E[g(X)h(Y)] &= E[g(X)]E[h(Y)] \\ &= \int_{-\infty}^{\infty} \delta(x-a)f_X(x)dx \int_{-\infty}^{\infty} \delta(y-b)f_Y(y)dy \\ &= f_X(a)f_Y(b). \end{aligned}$$

Thus $f_{XY}(a,b) = f_X(a)f_Y(b)$ for all a in the range space of X and all b in the range space of Y . \square

4 Inner Products and Orthogonality

We have seen that two random variables are uncorrelated if the expectation of the product is the product of the expectations. Let us now restrict attention to zero-mean random variables, in which case they are uncorrelated if $E(XY) = 0$. This observation prompts us to consider another mathematical concept: the inner product.

Definition 10 Let \mathcal{V} be a collection of random variables defined over the same probability space. The set Vc is called a **vector space** if

1. Addition is commutative: $X + Y = Y + X$.
2. Addition is associative: $(X + Y) + Z = X + (Y + Z)$.
3. There exists a **zero vector**, denoted 0 , such that $X + 0 = X$.
4. For every $X \in Vc$ there exists an **additive inverse**, $Y = -X \in \mathcal{V}$, such that $X + Y = 0$.
5. The distributive law holds: for every scalar α and every $X, Y \in \mathcal{V}$, $\alpha(X + Y) = \alpha X + \alpha Y$.
6. The scalar associative law holds: for all scalars α, β , $\alpha(\beta X) = \beta(\alpha X)$.
7. The scalar distributive law holds: for all scalars α, β , $(\alpha + \beta)X = \alpha X + \beta X$.
8. There is a unity scalar, denoted 1 , such that, $\forall X \in \mathcal{V}$, $1X = X$.

□

Definition 11 Let \mathcal{V} be a vector space of real random variables. An **inner product** is a function $\langle \cdot, \cdot \rangle: \mathcal{V} \times \mathcal{V} \rightarrow \Re$ such that

- $\langle X, Y \rangle = \langle Y, X \rangle$
- For any scalar α , $\langle \alpha X, Y \rangle = \alpha \langle X, Y \rangle$
- $\langle X + Y, Z \rangle = \langle X, Z \rangle + \langle Y, Z \rangle$.
- $\langle X, X \rangle \geq 0$ and $\langle X, X \rangle = 0$ iff $X = 0$.

□

Theorem 20 Let \mathcal{V} be a vector space of real random variables. Then

$$\langle X, Y \rangle = E(XY)$$

is an inner product.

Proof It is sufficient to show that this expression satisfies the definition. We prove the continuous case and leave the discrete case as an exercise. By inspection, it is easily seen that the expression

$$E(XY) = \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy$$

satisfies the first three parts of the definition. Also, it is easily seen that $E(XX) \geq 0$. Thus, all that remains is to consider the claim that $\langle X, X \rangle = 0$ implies $X = 0$. This may appear obvious at first glance, but we must remember that there do exist functions that, while not identically zero, nevertheless have zero power. Consider, for example, the function $X: \Re \rightarrow [0, 1]$ that is uniformly distributed, and suppose $\int_0^1 x^2 dx = 0$. Does this mean that $X \equiv 0$; that is $X(\omega) = 0 \forall \omega \in [0, 1]$? Not necessarily. Suppose

$$X'(\omega) = \begin{cases} 0 & x < 1 \\ 1 & x = 1 \end{cases} .$$

Clearly, $\int_0^1 x' dx' = 0$, but $X' \not\equiv 0$. What is true, however, is that $X' = 0$ on a set of probability measure unit. Thus, we must modify the last part of the definition to become $\langle X, X \rangle = 0$ iff $P(X = 0) = 1$. This generalization of the definition will have no effect on us

since, from an engineering perspective, two functions that possess the property that there is no energy in their difference cannot affect anything physical and cannot ever be detected by instrumentation. \square

Definition 12 Let \mathcal{V} be a vector space of real random variables. The **norm**, or **energy** of X , denoted $\|X\|$, is given by the square root of the inner product of the vector with itself.

$$\|X\| = \langle X, X \rangle^{\frac{1}{2}} = E(XX)^{\frac{1}{2}}.$$

\square

Definition 13 Let \mathcal{V} be a vector space of real random variables. X and Y are **orthogonal** if their inner product is zero:

$$\langle X, Y \rangle = E(XY) = 0,$$

in which case we write

$$X \perp Y.$$

\square

Two important facts are the so-called Cauchy-Schwarz inequality and the triangle inequality, stated without proof (you should have seen them in other contexts).

Theorem 21 Let \mathcal{V} be a vector space of real random variables. Then

$$|E(XY)| \leq \|X\| \|Y\|.$$

$$\|X + Y\| \leq \|X\| + \|Y\|.$$

5 Conditioning

The notion of “conditioning” is central to probability theory. It is the vehicle that connects the things we observe to the things we cannot directly observe but need to learn about. Suppose X and Y are two random variables such that direct observation of X is not possible, but it is possible to observe Y . Given that $Y = y$, what can this knowledge tell us about X ? One possibility is to compute the expected value of X *conditioned* on the event $Y = y$. In this section we explore this candidate and assess its attributes as an estimator of the value assumed by X .

5.1 Conditional Densities

The most obvious way to compute the conditional expectation is first to compute the conditional density function and then to compute

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx,$$

where $f_{X|Y}(x|y)$ is the conditional density of X given $Y = y$. The problem is, how to obtain this conditional density. If Y may assume a finite number of values, each with positive probability, this is not a difficult task, for then we have

$$f_{X|Y}(x|y) = \lim_{\Delta x \rightarrow 0} \frac{P(X \in [x - \Delta x, x + \Delta x] \times (Y = y))}{2\Delta x \cdot P(Y = y)}.$$

Writing this expression in terms of the joint distribution function, we obtain

$$f_{X|Y}(x|y) = \lim_{\Delta x \rightarrow 0} \frac{F_{XY}(x + \Delta x, y) - F_{XY}(x - \Delta x, y)}{2\Delta x \cdot P[Y = y]} = \frac{f_{XY}(x, y)}{f_Y(y)},$$

where f_Y is the probability mass function for Y and f_{XY} is the joint density/mass function of X and Y . As we let Δx tend to zero, this expression is well-defined.

However, what if Y assumes a continuum of values? Then the event $Y = y$ has zero probability of occurrence, and we need to be very careful in the formulation of our limit.

Perhaps the most obvious way to proceed is to define the conditional density as

$$f_{X|Y}(x|y) = \lim_{\Delta x, \Delta y \rightarrow 0} \frac{\frac{P((X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - \Delta y, y + \Delta y]))}{2\Delta x \cdot 2\Delta y}}{\frac{P([Y \in [y - \Delta y, y + \Delta y]])}{2\Delta y}} \quad (14)$$

$$= \lim_{\Delta x, \Delta y \rightarrow 0} \frac{\frac{P((X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - \Delta y, y + \Delta y]))}{2\Delta x \cdot 2\Delta y}}{\frac{P((X \in (-\infty, \infty)) \times (Y \in [y - \Delta y, y + \Delta y]))}{2\Delta y}} \quad (15)$$

Let's pay close attention to the way this limit is obtained. Note that this conditional density is defined for points (x, y) that are the limits of *rectangles* of the form

$$(X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - \Delta y, y + \Delta y]) \quad (16)$$

as Δx and Δy both approach zero independently. Without loss of generality, we assume that $\Delta x > 0$ and $\Delta y > 0$. Figure 2 illustrates a typical rectangle. To facilitate the limiting procedure it is convenient to express the probability associated with rectangles in terms

of the distribution function. We do this by means of what are called **partial difference operators**. The partial difference operator of step h_i , denoted $\Delta_{a_i}^{b_i}$, is defined by

$$\begin{aligned} \Delta_{x_i-\Delta_i}^{x_i+\Delta_i} &= F_{X_1, \dots, X_N}(x_1, \dots, x_{i-1}, x_i + \Delta_i, x_{i+1}, \dots, x_n) \\ &\quad - F_{X_1, \dots, X_N}(x_1, \dots, x_{i-1}, x_i - \Delta_i, x_{i+1}, \dots, x_n). \end{aligned}$$

Clearly, $\Delta \geq 0$. Composing Δ with itself yields, for $n = 2$,

$$\begin{aligned} \Delta_{x-\Delta x}^{x+\Delta x} \left(\Delta_{y-\Delta y}^{y+\Delta y} F_{XY}(x, y) \right) &= F_{XY}(x + \Delta x, y + \Delta y) - F_{XY}(x + \Delta x, y - \Delta y) \\ &\quad + F_{XY}(x - \Delta x, y - \Delta y) - F_{XY}(x - \Delta x, y + \Delta y). \end{aligned}$$

Using the fact that the probability associated with the rectangle $[x - \Delta x, x + \Delta x] \times [y - \Delta y, y + \Delta y]$ is expressed in terms of the distribution function as

$$P((X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - \Delta y, y + \Delta y])) = \Delta_{x-\Delta x}^{x+\Delta x} \left(\Delta_{y-\Delta y}^{y+\Delta y} F_{XY}(x, y) \right),$$

the numerator of the ratio in (15) is

$$\frac{F_{XY}(x+\Delta x, y+\Delta y) - F_{XY}(x+\Delta x, y-\Delta y) + F_{XY}(x-\Delta x, y-\Delta y) - F_{XY}(x-\Delta x, y+\Delta y)}{2\Delta x \cdot 2\Delta y}$$

which becomes, as Δx and Δy both approach zero, the joint density function $f_{XY}(x, y)$. The limit of the denominator of (15) becomes, as Δy approaches zero, the marginal density of Y , which may be expressed as $\int_{-\infty}^{\infty} f_{XY}(\alpha, y) d\alpha$. Thus, we may conclude, for this case, that

$$f_{X|Y}(x|Y = y) = \frac{f_{XY}(x, y)}{\int_{-\infty}^{\infty} f_{XY}(\alpha, y) d\alpha} = \frac{f_{XY}(x, y)}{f_Y(y)}. \quad (17)$$

The conditional density defined by 17 is what we often think of when we go about defining such things. But we must remember that we arrived at this result by a very carefully constructed limit, namely, we viewed the point (x, y) as the limit of rectangles. This is not the only way express the point (x, y) as the limit of sets. Here's another way. Consider sets of the form

$$\left\{ \frac{y}{X} - \Delta y \leq \frac{Y}{X} \leq \frac{y}{X} + \Delta y \right\},$$

or, equivalently,

$$\{y - X\Delta y \leq Y \leq y + X\Delta y\}.$$

Now consider sets of the form

$$\{X \in [x - \Delta x, x + \Delta x], Y \in [y - X\Delta y, y + X\Delta y]\}.$$

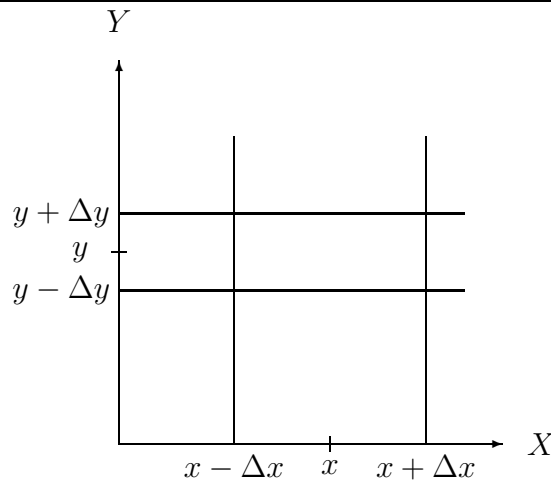


Figure 2: The family of rectangles $(X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - \Delta y, y + \Delta y])$.

These sets are trapezoids, as illustrated in Figure 3. Note that the lines defining the Y component have slope $\pm\Delta y$, but as Δx and Δy both tend to zero, the trapezoid converges to the limit point (x, y) , just as as was the case with rectangular sets. With this model, the conditional density becomes

$$f_{X|Y}(x|y) = \lim_{\Delta x, \Delta y \rightarrow 0} \frac{\frac{P((X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - X\Delta y, y + X\Delta y]))}{2\Delta x \cdot 2\Delta y}}{\frac{P((Y \in [y - X\Delta y, y + X\Delta y]))}{2\Delta y}} \quad (18)$$

$$= \lim_{\Delta x, \Delta y \rightarrow 0} \frac{\frac{P((X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - X\Delta y, y + X\Delta y]))}{2\Delta x \cdot 2\Delta y}}{\frac{P((X \in (-\infty, \infty)) \times (Y \in [y - X\Delta y, y + X\Delta y]))}{2\Delta y}} \quad (19)$$

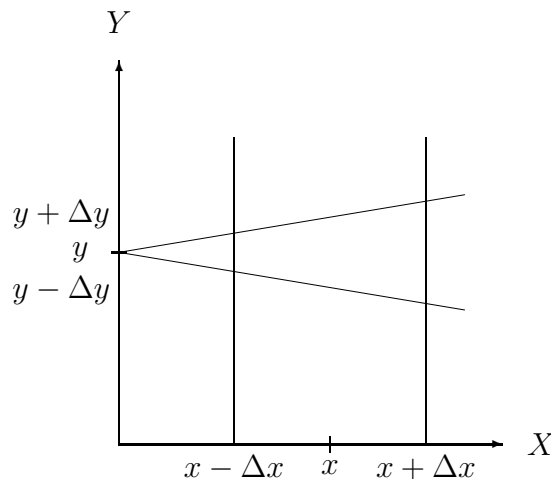


Figure 3: The family of trapezoids $(X \in [x - \Delta x, x + \Delta x]) \times (Y \in [y - X\Delta y, y + X\Delta y])$.

The numerator of the ratio in (19) may be expressed in terms of the joint distribution

function as

$$\frac{F_{XY}(x+\Delta x, y+x\Delta y) - F_{XY}(x+\Delta x, y-x\Delta y) + F_{XY}(x-\Delta x, y-x\Delta y) - F_{XY}(x-\Delta x, y+x\Delta y)}{2\Delta x \cdot 2\Delta y}$$

Now suppose we take the limit as $\Delta y \rightarrow 0$. Let us examine the quantity $F_{XY}(x + \Delta x, y + x\Delta y) - F_{XY}(x + \Delta x, y - x\Delta y)$, and note that we can re-write this expression as

$$\Delta F = F_{XY}(x + \Delta x, y + \Delta z) - F_{XY}(x + \Delta x, y - \Delta z),$$

where $\Delta z = x\Delta y$. Let us first assume that $x > 0$. We may then form the ratio

$$\frac{\Delta F}{\Delta y} = \frac{\Delta F}{\Delta z} \frac{\Delta z}{\Delta y},$$

or, since $\frac{\Delta z}{\Delta y} = x$, we have

$$\frac{\Delta F}{\Delta y} = \frac{[F_{XY}(x + \Delta x, y + \Delta z) - F_{XY}(x + \Delta x, y - \Delta z)]x}{2\Delta z}.$$

If $x < 0$, we have $\Delta z = -|\Delta z|$ and $x = -|x|$, so

$$\frac{\Delta F}{\Delta y} = \frac{[F_{XY}(x + \Delta x, y + \Delta z) - F_{XY}(x + \Delta x, y - \Delta z)](-|x|)}{-2|\Delta z|},$$

so in general, we obtain

$$\frac{\Delta F}{\Delta y} = \frac{[F_{XY}(x + \Delta x, y + \Delta z) - F_{XY}(x + \Delta x, y - \Delta z)]|x|}{|2\Delta z|}.$$

We have thus succeeded in reducing this problem to the previous case, except for the addition of the extra term $|x|$. Passing to the limit as Δx and Δz (and hence Δy) tend to zero, we obtain the conditional density function

$$f_{X|Y}(x|Y = y) = \frac{f_{XY}(x, y)|x|}{\int_{-\infty}^{\infty} f_{XY}(\alpha, y)|\alpha|d\alpha}.$$

This is a very different conditional distribution than the one obtained with the rectangle structure!

What's going on here? We have competing definitions for the conditional density. This is because there are many ways in which limiting operations can take place, and there is no mathematical reason to prefer one over the other. This suggests that we must pay very careful attention to the relationships between X and Y when computing conditional expectations. This prompts us to ask a very significant question: Is there a way to define the conditional expectation without first computing the conditional density function? To answer this question, we need to involve σ -fields.

5.2 Conditioning on a σ -field

Definition 14 Given a random variable X satisfying the condition $E|X| < \infty$ (this condition can be relaxed in various ways, but we don't need to worry about that now), the **conditional expectation** of X given the σ -field $\mathcal{F} = \sigma\{Y\}$ is defined as a random variable, written variously as $E^{\mathcal{F}}X$, $E[X|\mathcal{F}]$ or $E[X|Y]$, such that

1. $E[X|\mathcal{F}]$ is an \mathcal{F} -measurable function; that is, sets of the form

$$\{\omega \in \Omega: a < E[X|\mathcal{F}](\omega) < b\}$$

are elements of \mathcal{F} .

2. The random variable $X - E[X|\mathcal{F}]$ is orthogonal¹ to all \mathcal{F} -measurable functions; that is,

$$E[(X - E[X|\mathcal{F}])Z] = 0 \quad \forall Z \in \mathcal{F}.$$

□

This second property is the one that makes conditional expectations useful, and we will have quite a bit to say about this as we progress through the course.

Let's pause a moment and examine some differences between this and the definition of conditional expectation defined in terms of conditional distributions.

- The definition in terms of a conditional distribution is constructive, in that one is able actually to compute the conditional expectation with the conditional distribution.
- The definition in terms of σ -fields is not constructive. The definition is provided in terms of properties that the conditional expectation must possess, but does not point to a way to compute the conditional expectation.

This situation is somewhat similar, at least in spirit, to the situation with differential equations. You may recall that, when considering equations of the form $\dot{x} = f(x, u)$, all the theory provides is theorems regarding existence and uniqueness; it does not tell us how to find the solution. This is not to say, however, that the properties of conditional expectations cannot be used to identify solutions—it just can't generally be used to construct them.

¹Recall that orthogonality is defined in terms of the inner product of two random variables as $\langle X, Y \rangle = E[XY]$.

Of course, if one can construct the conditional density or mass function, one certainly may use it to compute the conditional expectation. But, by exploiting the properties of conditional expectations, one may be able to develop ways to construct the conditional expectation without first constructing the conditional density. Remember, the conditional expectation is just the first moment of the conditional density, and one may not need all of the information that the conditional density provides in order to compute the conditional expectation. Sometimes we can obtain all of the information we need by exploiting the properties of moments of distributions, rather than requiring complete knowledge of the distribution.

- The conditional expectation defined in terms of a conditional distribution is, fundamentally a number; that is, it is computed for each value event $Y = y$. It may be viewed as a function by computing its value for each possible value of y . With this extension, we can think of conditional expectation as a function of Y , and thus as a random variable.
- The conditional expectation defined in terms of a σ -field is, fundamentally, a random variable. If that σ -field is generated by a random variable Y , then the conditional expectation is a function of Y , and can be evaluated for each event $Y = y$. With this restriction, conditional expectation may be viewed as a number. (that is, it assumes the value corresponding to the inverse image of the event $Y = y$).

5.3 Some Properties of Conditional Expectation

The essential difference between an unconditional expectation and a conditional expectation is that *the latter is a function of the conditioning event $[Y = y]$, and is therefore a function of the realization, y* . Thus, before the realization is obtained, we may think of conditional expectation as a function of Y , that is, *we may view conditional expectation as a random variable*, which we denote as $E(X|Y)$. That is, $E(X|Y) : \Omega \rightarrow \mathfrak{R}$, where Ω is the sample space of Y . Now, we could proceed to define the distribution function of this random variable, that is, we could define the distribution function

$$F_{E(X|Y)}(y) = P(\{\omega \in \Omega : E(X|Y(\omega)) \leq y\}),$$

and compute density or mass functions from that distribution. Fortunately, however we generally do not need to go that route, thanks to the fundamental theorem of expectations.

Conditional expectation has proven to be of such enormous value in applications that we need to catalogue its properties.

Theorem 22 *Let X , Y , and Z be random variables, let g be any Borel function, and let c and d be arbitrary constants.*

- (a) $E(X|Y) = EX$ if X and Y are independent.
- (b) $E(g(X)) = E(E(g(X)|Y))$, which yields, as a special case $EX = E(E(X|Y))$.
- (c) $E(g(Y)X|Y) = g(Y)E(X|Y)$.
- (d) $E(g(Y)X) = E(g(Y)E(X|Y))$.
- (e) $E(c|Y) = c$.
- (f) $E(g(Y)|Y) = g(Y)$.
- (g) $E(X|g(Y)) = E(E(X|g(Y))|Y) = E(E(X|Y)|g(Y))$.
- (h) $E(cX + dZ|Y) = cE(X|Y) + dE(Z|Y)$.

Proof We will provide proofs for the continuous case only; proofs for the discrete case follow by either using delta functions for the density or by making the appropriate substitutions between integrals and summations, etc.

- (a) If X and Y are independent, then $f_{X|Y}(x|y) = f_X(x)$, so, for each event $[Y = y]$,

$$\begin{aligned} E(X|Y = y) &= \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= EX \end{aligned}$$

for all y , thus $E(X|Y) \equiv EX$.

- (b) By the fundamental theorem and the definition of $E(g(X)|Y = y)$,

$$\begin{aligned} E\left(E(g(X)|Y)\right) &= \int_{-\infty}^{\infty} E(g(X)|Y = y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx \right] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} g(x) \left[\int_{-\infty}^{\infty} f_{XY}(x, y) dy \right] dx \\ &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &= E(g(X)). \end{aligned}$$

(c) For each event $[Y = y]$, $g(y)$ is a constant, so

$$\begin{aligned} E(g(Y)X|Y = y) &= E(g(y)X|Y = y) \\ &= \int_{-\infty}^{\infty} g(y)x f_{X|Y}(x|y) dx \\ &= g(y) \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \\ &= g(y)E(X|Y = y). \end{aligned}$$

Since this relationship holds for all y , we have $E(g(Y)X|Y) = g(Y)E(X|Y)$.

(d)

$$\begin{aligned} E(g(Y)X) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y)x f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y)x f_{X|Y}(x|y) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} g(y) \left[\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \right] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} g(y)E(X|Y = y) f_Y(y) dy \\ &= E(g(Y)E(X|Y)). \end{aligned}$$

(e) This result is obvious by properties of expectation.

(f) Given the event $(Y = y)$, the distribution of $g(Y)$ is concentrated at the point $g(y)$, that is, $P(g(Y) = g(y)|Y = y) = 1$. Thus,

$$E(g(Y)|Y = y) = g(y)P(g(Y) = g(y)|Y = y) = g(y).$$

Since this relationship holds for all y , we have $E(g(Y)|Y) = g(Y)$.

(g) To show that $E(X|g(Y)) = E(E(X|g(Y))|Y)$ we note that $E(X|g(Y))$ is some function of Y , so by (c) and (e)

$$E(E(X|g(Y))|Y) = E(X|g(Y))E(1|Y) = E(X|g(Y)).$$

Showing that $E(X|g(Y)) = E(E(X|Y)|g(Y))$ is more difficult, and we will restrict our attention to the special case where g is increasing and pointwise invertible. (The case for arbitrary Borel measurable g is beyond the scope of this class, and involves what is called Radon-Nikodym theory. With that theory, however, the proof is a one-liner.)

Define the random variable $Z = g(Y)$, fix the event $(Z = z)$, and define the function $\phi(Y) = E(X|Y)$. Since g is pointwise invertible, given that the event $(Z = z)$ occurs implies that $P(Y = g^{-1}(z)) = 1$. Thus, $f_{Y|Z}(y|z) = \delta(y - g^{-1}(z))$, so

$$\begin{aligned}
 E(E(X|Y)|g(Y) = z) &= E(\phi(Y)|Z = z) \\
 &= \int_{-\infty}^{\infty} \phi(y)f_{Y|Z}(y|z)dy \\
 &= \int_{-\infty}^{\infty} \phi(y)\delta(y - g^{-1}(z))dy \\
 &= \phi(g^{-1}(z)) \\
 &= \int_{-\infty}^{\infty} xf_{X|Y}(x|g^{-1}(z))dx.
 \end{aligned} \tag{20}$$

To complete our demonstration we must evaluate $f_{X|Y}(x|g^{-1}(z))$. But

$$\begin{aligned}
 f_{X|Y}(x|g^{-1}(z))dx &= \frac{f_{XY}(x, g^{-1}(z))dxdy}{f_Y(g^{-1}(z))dy} \\
 &\approx \frac{P((x < X \leq x + dx) \cap (g^{-1}(z) < Y \leq y + dy))}{P(g^{-1}(z) < Y \leq y + dy)} \\
 &= \frac{P((x < X \leq x + dx) \cap (z < g(Y) \leq g(y + dy)))}{P(z < g(Y) \leq g(y + dy))} \\
 &\approx \frac{P((x < X \leq x + dx) \cap (z < g(Y) \leq z + dz))}{P(z < g(Y) \leq z + dz)} \\
 &= \frac{f_{XZ}(x, z)dxdz}{f_Z(z)dz} \\
 &\approx f_{X|Z}(x|z)dx.
 \end{aligned}$$

Substituting this into (20) yields

$$E(E(X|Y)|g(Y) = z) = \int_{-\infty}^{\infty} xf_{X|Z}(x|z)dx = E(X|g(Z)).$$

(h) This result follows by the property of expectations; namely, that the expectation of a sum is the sum of expectations, and the expectation of a constant times a random variable is the constant times the expectation. \square

We complete this section with a brief discussion of conditional variance.

Definition 15 Let X and Y be random variables. Then the **conditional variance of X given Y** is

$$\text{Var}(X|Y) = E((X - E(X|Y))^2|Y). \tag{21}$$

□

We note from this definition that, just as is conditional expectation, conditional variance is also a random variable, since it is a function of Y .

Theorem 23 $\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E(X|Y))$; that is, the unconditional variance of X is equal to the expectation of the conditional variance plus the variance of the conditional expectation.

Proof Expanding (21) and collecting terms yields

$$\begin{aligned}\text{Var}(X|Y) &= E\left((X - E(X|Y))^2|Y\right) \\ &= E(X^2|Y) - 2E(E(X|Y)X|Y) + (E(X|Y))^2 \\ &= E(X^2|Y) - 2E(X|Y)E(X|Y) + (E(X|Y))^2 \\ &= E(X^2|Y) - (E(X|Y))^2.\end{aligned}$$

Taking expectations of the conditional variance and using part (b) of Theorem 22 we have

$$\begin{aligned}E(\text{Var}(X|Y)) &= E(E(X^2|Y)) - E((E(X|Y))^2) \\ &= EX^2 - E\left((E(X|Y))^2\right).\end{aligned}\tag{22}$$

Next, let us compute the variance of the conditional expectation:

$$\begin{aligned}\text{Var}(E(X|Y)) &= E((E(X|Y))^2) - \left(E(E(X|Y))\right)^2 \\ &= E((E(X|Y))^2) - (EX)^2\end{aligned}\tag{23}$$

Now, adding (22) and (23) we obtain

$$\text{Var}(X) = E(\text{Var}(X|Y)) + E((E(X|Y))^2),$$

□

6 Estimation Theory

One of the most important uses of conditional expectation is in estimation theory. Let us begin this discussion by asking: What constitutes a good estimator? An obvious answer is

that the estimate be close to the true value. Suppose the random variable Y is some signal of interest, and suppose the random variable X is observed. A fundamental estimation problem is to infer the value that Y assumes by observing the value that X assumes. To proceed, we need to decide what criterion we will use to determine the quality of an estimator. Obviously, our estimator will be a function of X , call it $g(X)$, and our job is to determine the function g . The next step is to form some function $L(\cdot, \cdot)$ of Y and $g(X)$ to be minimized. The following criteria are desirable.

1. The quantity to be minimized should be a function of the difference between Y and $g(X)$.
2. The function should be symmetric, in that negative errors and positive errors are weighted the same,
3. The function should penalize large errors more than small errors.

There are many such functions, but perhaps the most appealing is the **squared error criterion**, leading to a function of the form

$$L[Y, g(X)] = E\left((Y - g(X))^2\right).$$

Definition 16 The function g that is the solution to

$$\min_g \left\{ E\left((Y - g(X))^2\right) \right\}$$

is called the **minimum mean-square estimate** (MMSE) of Y given X □

The following theorem is one of the more celebrated results of estimation theory.

Theorem 24 *The MMSE of Y given X is the conditional expectation $E(Y|X)$.*

Proof We write

$$\begin{aligned} E\left((Y - g(X))^2\right) &= E\left((Y - E(Y|X) + E(Y|X) - g(X))^2\right) \\ &= E\left((Y - E(Y|X))^2\right) + E\left((E(Y|X) - g(X))^2\right) \\ &\quad + 2E\left((Y - E(Y|X))(E(Y|X) - g(X))\right). \end{aligned}$$

Let us examine the cross term, $E[(Y - E(Y|X))(E(Y|X) - g(X))]$, and define $h(X) = E(Y|X) - g(X)$. We have

$$\begin{aligned} E[(Y - E(Y|X))h(X)] &= E\left[E((Y - E(Y|X))h(X)|X)\right] \text{ by Theorem 22(b)} \\ &= E\left[h(X)E((Y - E(Y|X))|X)\right] \text{ by Theorem 22(c)} \\ &= E\left[h(X)[(E(Y|X) - E(Y|X))]\right] \\ &= 0 \end{aligned}$$

for all functions g . Thus,

$$E(Y - g(X))^2 = E(Y - E(Y|X))^2 + E(E(Y|X) - g(X))^2. \quad (24)$$

The first term on the right of (24) is not a function of g , and the second term on the right (and hence the whole expression) is minimized by setting

$$g(X) = E(Y|X).$$

□

The MMSE is a beautiful result but, unfortunately, there is only a small handful of cases for which explicit analytical solutions can be obtained, or even computationally feasible approximation procedures are known. Therefore, sub-optimum estimates of various kinds are sought, with the most common estimation structures being constrained to be linear functions of the observations. An important exception to this limitation is when X and Y are jointly normal, in which case the MMSE is an affine function of X , as the following example illustrates.

Example 5 MMSE for Normal Distributions. *Let X and Y have the bivariate normal distribution given by (11), which we repeat here:*

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]},$$

where the univariate density of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(x-\mu_x)^2/2\sigma_x^2}.$$

The conditional density of Y given X is obtained as the ratio of these two densities, yielding

$$\begin{aligned}
 f_{Y|X}(y|x) &= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right] + \frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2} \\
 &= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\rho^2\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]} \\
 &= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{y-\mu_y}{\sigma_y} - \frac{\rho(x-\mu_x)}{\sigma_x}\right]^2} \\
 &= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_y^2}\left[y - \left(\mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2}.
 \end{aligned}$$

Clearly, this conditional distribution is normal, and we see by inspection that the mean and variance of this distribution are

$$E(Y|X = x) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x - \mu_x)$$

and

$$\text{Var}(Y|X = x) = (1 - \rho^2)\sigma_y^2.$$

Thus, we have, using the fact that $\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$, that

$$E(Y|X) = \mu_y + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - \mu_x) \quad (25)$$

and

$$\text{Var}(Y|X) = \text{Var}(Y) - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(X)}.$$

Thus, the conditional expectation is an affine transformation of the observation.

Equation (25) is easily computed and intuitively pleasing. Consequently, it is often used as an estimate of Y given X , even if the two random variables are not bivariate normal.

Definition 17 Let X and Y be random variables. The **linear least squares estimate** (LLSE) of Y given X is the function

$$g(X) = a + bX$$

where the coefficients a and b are chosen to minimize the expression

$$E(Y - a - bX)^2.$$

□

Theorem 25 *If X and Y are jointly normal, then the LLSE and the MMSE coincide.*

Proof

$$\begin{aligned} E(Y - a - bX)^2 &= E(Y^2 - 2aY - 2bXY + a^2 + 2abX + b^2X^2) \\ &= EY^2 - 2aEY - 2bE(XY) + a^2 \\ &\quad + 2abEX + b^2EX^2. \end{aligned}$$

Taking partial derivatives with respect to a and b we obtain

$$\frac{\partial}{\partial a} E(Y - a - bX)^2 = -2EY + 2a + 2bEX$$

and

$$\frac{\partial}{\partial b} E(Y - a - bX)^2 = -2E(XY) + 2aEX + 2bEX^2.$$

Equating these expressions to zero and solving for the corresponding a and b yields

$$\begin{aligned} b &= \frac{E(XY) - EXEY}{EX^2 - (EX)^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \rho \frac{\sigma_y}{\sigma_x} \\ a &= EY - bEX = EY - \frac{\rho\sigma_y}{\sigma_x} EX. \end{aligned}$$

Thus, the LLSE of Y given X is

$$g(X) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(X - \mu_x) \quad (26)$$

and the variance of this estimate is, after appropriate manipulations,

$$\sigma_y^2(1 - \rho^2),$$

which coincides with the MMSE in the bivariate normal case. \square

The LLSE is the basis of an important statistical concept called **linear regression**. Suppose you make n measurements of a quantity x , obtaining values x_1, x_2, \dots, x_n , and for each of these values of x you measure a corresponding value y_i of a quantity y , for $i = 1, \dots, n$. The result of your measurements is a set of n pairs (x_i, y_i) , $i = 1, \dots, n$. If when plotted on coordinate paper the points (x_i, y_i) appear to lie nearly on a straight line, you may wish to determine the equation of the line which best represents this relationship. Clearly, there are many possible criteria to use in choosing a “best” line. However, the criterion which generally leads to the simplest calculations is the least-squares criterion, according to which

you choose the slope and intercept so as to minimize the sum of the squares of the vertical distances from the plotted points to the line; that is, you would minimize

$$\sum_{i=1}^n (y_i - bx_i - a)^2.$$

With a little bit of work it can be shown that the coefficients a and b that minimize this sum are given by

$$a = \frac{rs_x}{s_y} k \bar{x}$$

and

$$b = r \frac{s_x}{s_y}$$

where

$$\begin{aligned} s_x^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ s_y^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{ns_x s_y}, \end{aligned}$$

with

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i. \end{aligned}$$

Thus, the fitted line is of the form

$$y = \bar{y} + \frac{rs_y}{s_x} (x - \bar{x}). \quad (27)$$

This line is called the **regression of y on x** .

We emphasize that (27) does not directly involve any considerations of probability at all, but comparing (27) and (26) reveals that the problem of finding the best regression curve is very analogous to that of minimizing the mean-square error. In fact, if we view the values x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n as being the realizations of sample random variables, we may then recognize \bar{x} and \bar{y} as estimates of the means of X and Y , respectively, s_x, s_y as estimates of their respective variances, and r as an estimate of the correlation coefficient.

7 Convergence

Often we are interested in notions of **convergence**, that is, the limiting behavior of random variables as we consider large numbers of them. For example, we have found, with independent trials drawn from a population, that the sample mean and sample variance are useful, and we observed that they can be interpreted as estimators of the true mean and variance of the population. Before we can use these quantities with confidence, however, it would behoove us to determine in what sense, if any, we are justified in assuming that, as the number of samples grows larger and larger, that the estimators tend closer and closer to the true values. In this regard, there are four types of convergence that we may consider.

Definition 18 Let X be a random variable, and let X_1, X_2, \dots be a sequence of random variables. We say that $\{X_n\}$, $n = 1, 2, \dots$, **converges pointwise** to X if

$$\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \quad \text{for every } \omega \in \Omega.$$

□

Pointwise convergence is the strongest possible notion of convergence. Unfortunately, it is too strong to be useful, because we simply cannot prove very many interesting theorems that guarantee pointwise convergence. This situation has motivated the need for the development of less stringent notions of convergence.

Definition 19 Let X be a random variable, and let X_1, X_2, \dots be a sequence of random variables. We say that $\{X_n\}$, $n = 1, 2, \dots$ **converges almost surely** to X , or X_i **converges to X with probability one** if

$$P(\{\omega \in \Omega : \lim_{n \rightarrow \infty} |X_n(\omega) - X(\omega)| = 0\}) = 1.$$

□

This notion of convergence means that, excluding a set of elementary events of probability zero, convergence is pointwise; that is, $X_n(\omega) \rightarrow X(\omega)$ for all ω except those elementary events that belong to a particular set of probability zero. Fortunately, this restriction is not very limiting, and permits great deal of mathematics to be developed. In practice, there is essentially no difference between almost sure convergence and pointwise convergence.

Although, technically speaking, almost sure convergence is weaker than pointwise convergence, it still is a very strong concept of convergence, and it is sometimes convenient to consider a yet weaker notion.

Definition 20 Let X be a random variable, and let X_1, X_2, \dots be a sequence of random variables. We say that $\{X_n\}$, $n = 1, 2, \dots$ **converges in probability** to X if, for every real number $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0.$$

□

Although they sound much the same, convergence with probability one and convergence in probability are very different concepts. Convergence with probability one applies to the individual realizations of the random variables; convergence in probability does not. This difference is illustrated most graphically by the placement of the limiting operation in the two definitions. With almost sure convergence, the limit is *inside* the probability, and is applied for each elementary event. With convergence in probability, the limit is *outside* the probability, and does not apply to each elementary event. For example, it may be that, for X_n , the subset of elementary events such that $P[|X_n - X| > \epsilon]$ is *different* from the subset of elementary events such that $P[|X_j - X| > \epsilon]$ for $j \neq n$. As long as these subsets have diminishing probability as $j \rightarrow \infty$, convergence in probability can occur without pointwise convergence. It is possible, in fact, for there to be no elementary events $\omega \in \Omega$ such that $X_n(\omega) \rightarrow X(\omega)$ as $n \rightarrow \infty$, yet X_n can still converge to X in probability!

While it therefore is not true that convergence in probability implies convergence with probability one, the converse is true.

Theorem 26 *If X_1, X_2, \dots converges to X with probability one, then X_1, X_2, \dots converges to X in probability.*

Proof Given an $\epsilon > 0$, define the sequence of sets

$$A_i(\epsilon) = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon \text{ for some } n \geq i\}.$$

Clearly, $A_j(\epsilon) \subset A_i(\epsilon)$ for $j > i$, thus the sequence $P(A_i(\epsilon))$ decreases monotonically with i . The definition of probability one convergence assures that the decrease is to zero. Therefore,

$$\lim_{i \rightarrow \infty} P(A_i(\epsilon)) = \lim_{i \rightarrow \infty} P(|X_i - X| > \epsilon) = 0.$$

□

In addition to these probability-based notions of convergence, there is a distinct notion that is motivated by considerations of the variance the sample mean of independent random samples, X_i , $i = 1, \dots, n$, of a population random variable X with variance σ^2 .

Theorem 27 Let X_1, X_2, \dots, X_n be sample values of a population random variable X with mean m and variance σ^2 . Let S_n be the sample mean. Then

$$ES_n = m \quad (28)$$

and

$$\text{Var}(S_n) = \frac{\sigma^2}{n}. \quad (29)$$

Proof

$$\begin{aligned} ES_n &= E \left[\frac{1}{n} \sum_{i=1}^n X_i \right] \\ &= \frac{1}{n} E \left[\sum_{i=1}^n X_i \right] \\ &= \frac{1}{n} \sum_{i=1}^n EX_i \\ &= m. \end{aligned}$$

Also, since the X_i 's are independent and hence $\text{COV}(X_i, X_j) = 0$, $i \neq j$, we have

$$\begin{aligned} \text{Var}(S_n) &= \frac{1}{n^2} \text{Var} \left(\sum_{i=1}^n X_i \right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

□

Recall that we computed the variance of the sample mean, $S_n = \frac{1}{n} \sum_{i=1}^n X_i$, to be $\text{Var}(S_n) = \frac{\sigma^2}{n}$. Obviously, as $n \rightarrow \infty$, we have $\text{Var}(S_n) \rightarrow 0$. At issue is how to interpret this result.

Definition 21 Let X be a random variable, and let X_1, X_2, \dots be a sequence of random variable. We say that $\{X_n\}$, $n = 1, 2, \dots$ **converges in mean square** to X if

$$\lim_{n \rightarrow \infty} E(X_n - X)^2 = 0.$$

We adopt the notation

$$\text{l. i. m.}_{n \rightarrow \infty} X_n = X$$

(read as “limit in mean square”) to denote this kind of convergence. \square

This convergence criterion is quite different from almost sure convergence, or convergence in probability, because it deals only with the second moment of the joint distributions of X_n and X . Considering the sample mean, we see that, since

$$\lim_{n \rightarrow \infty} E(S_n - m)^2 = \lim_{n \rightarrow \infty} \frac{\sigma^2}{n} = 0,$$

then S_n converges in mean square to the degenerate random variable Z , where $Z \equiv m$, that is, Z is a degenerate random variable that concentrates all of its probability mass at one point, m .

Mean-square convergence does not imply pointwise convergence, or even almost sure convergence. It does, however, admit a rather interesting interpretation from an engineering perspective. In the vernacular of engineering, the square of a signal is often viewed as “power.” For example, consider a circuit consisting of a one-ohm resistor. The power dissipated across the resistor is equal to the voltage squared. Essentially, the variance of the difference between $S_n - m$ can be viewed as the “power” in the error. Viewed from this perspective, convergence in mean-square means that, in the limit, there is no power in the error. Thus, even though the convergence may not be with probability one, the significance of the error is of no practical importance.

Convergence in mean square does not imply, nor is implied by, almost sure convergence. Mean-square convergence does, as we will shortly see, imply convergence in probability. We complete our discussion of convergence by introducing yet another notion.

Definition 22 Let X is a random variable with distribution function F_X , and let X_1, X_2, \dots with respective distribution functions F_{X_1}, F_{X_2}, \dots be a sequence of random variables, we say that $\{X_n\}$ **converges in distribution** to the random variable X if the sequence of distribution functions of X_i , $i = 1, 2, \dots$ converges to the distribution function of X ; that is, if

$$\lim_{i \rightarrow \infty} F_{X_i}(x) = F_X(x) \quad \text{for all } x \in (-\infty, \infty).$$

\square

Convergence in distribution is a very weak form of convergence, and is implied by the other forms that we have described. The convergence of X_n to X in distribution does

not imply in any sense whatsoever that $X_i(\omega)$ approach $X(\omega)$ for any values of ω ; only the sequence of distributions has to approach a limit. The most important example of convergence in distribution is provided by the central limit theorem.

To summarize, convergence in distribution is implied by convergence in probability, which in turn is implied by either convergence in mean square or convergence with probability one (neither of these latter notions of convergence implies the other).

7.1 The Central Limit Theorem

Consider the following situation. Suppose the population random variable X has finite mean m and finite variance σ^2 , but otherwise the distribution function (including the values of these parameters) is arbitrary and unknown. If we conduct independent trials, then we know that, whatever the distribution of the sum of the sample values X_1, X_2, \dots, X_n is, it is the n -fold convolution of the distribution of the population. Thus, whatever the distribution of the population is, the distribution of the sample mean seems to be a rather complicated function of it, and it would appear that all we can say with certainty is that it has mean m and variance $\frac{\sigma^2}{n}$. The essence of the central limit theorem is that, in fact, we can say much more—we can say that the distribution of the sample mean is approximately normal with mean m and variance $\frac{\sigma^2}{n}$ *regardless of the distribution of the population!* In other words, the n -fold convolution of any population distribution tends to the normal distribution. This is a rather remarkable result, and justifies the exalted status that the central limit theorem enjoys in the probability and statistics communities.

Suppose we perform the following affine transformation on the sample mean

$$Y_n = \frac{S_n - m}{\sigma/\sqrt{n}}.$$

Then $EY_n = 0$ and $\text{Var}(Y_n) = 1$. The central limit theorem states that Y_n tends to $\mathcal{N}(0, 1)$ as $n \rightarrow \infty$, and therefore, for each interval (a, b) , the probability that Y_n will assume a value between a and b , and thus, that S_n will assume values between $\frac{\sigma}{\sqrt{n}}a + m$ and $\frac{\sigma}{\sqrt{n}}b + m$, tends, as $n \rightarrow \infty$, to the probability that a random variable that is $\mathcal{N}(0, 1)$ will assume values between a and b .

The “general limit problem” of probability theory is actually a whole family of theorems dealing with sequences of sums of independent random variables. We are interested in the following version, which is generally referred to by statisticians as the central limit theorem.

Theorem 28 The central limit theorem Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables each having mean m and variance σ^2 , and let

$$S_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then for any real numbers a and b with $a < b$,

$$P\left(a < \frac{S_n - m}{\sigma/\sqrt{n}} < b\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \quad \text{as } n \rightarrow \infty,$$

or, equivalently,

$$P\left(\frac{\sigma}{\sqrt{n}}a + m < S_n < \frac{\sigma}{\sqrt{n}}b + m\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \quad \text{as } n \rightarrow \infty,$$

or, equivalently,

$$P\left(a < \frac{X_1 + \dots + X_n - nm}{\sigma\sqrt{n}} < b\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \quad \text{as } n \rightarrow \infty,$$

The content of this theorem is that, regardless of the underlying common distribution (as long as it has finite variance) of a family of random variables, the sequence of normed and centered partial sums converges in distribution to the unit normal distribution.

7.2 The Weak Law of Large Numbers

The following theorem, and particularly the two corollaries that follow, are among the most frequently used tools in probability theory.

Theorem 29 Let X be a random variable such that $E[|X|^r] < \infty$ for $r > 0$ not necessarily an integer; then

$$P(|X| \geq \epsilon) \leq \frac{E[|X|^r]}{\epsilon^r}$$

for every $\epsilon > 0$.

Proof

$$\begin{aligned} E|X|^r &= \int_{-\infty}^{\infty} |x|^r f_X(x) dx \\ &= \int_{\{|x| < \epsilon\}} |x|^r f_X(x) dx + \int_{\{|x| \geq \epsilon\}} |x|^r f_X(x) dx \\ &\geq \int_{\{|x| \geq \epsilon\}} |x|^r f_X(x) dx \\ &\geq \epsilon^r \int_{\{|x| \geq \epsilon\}} f_X(x) dx \\ &= \epsilon^r P[|X| \geq \epsilon], \end{aligned}$$

from which we obtain the desired inequality. \square

The following two corollaries are useful since they provide bounds for probability in terms of the mean and variance of a random variable.

Corollary 1 Markov's inequality. *If X is a random variable that takes only nonnegative values, then for any value $a > 0$,*

$$P(X \geq a) \leq \frac{EX}{a}.$$

Proof This is a special case of Theorem 29 for $r = 1$. \square

Corollary 2 Chebyshev's inequality. *If X is a random variable with finite mean m and variance σ^2 , then for any value $k > 0$,*

$$P[|X - m| \geq k] \leq \frac{\sigma^2}{k^2}.$$

Proof Let $Y = (X - m)^2$. By Theorem 29 with $r = 1$ and $\epsilon = k^2$,

$$P(Y \geq k^2) \leq \frac{E[|Y|]}{k^2},$$

but $(X - m)^2 \geq k^2$ if and only if $|X - m| \geq k$, so

$$P(|X - m| \geq k) \leq \frac{\sigma^2}{k^2}$$

as required. \square

We are now in a position to establish the relationship between mean-square convergence and convergence in probability.

Theorem 30 *If X_1, X_2, \dots converges in mean square to X , then X_1, X_2, \dots converges to X in probability.*

Proof From the Chebyshev inequality applied to $|X_n - X|$ we have, for any $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \leq \frac{E(X_n - X)^2}{\epsilon^2}.$$

The right side of this inequality tends to zero as $n \rightarrow \infty$ by definition of mean square, thus establishing convergence in probability. \square

The following counterexample establishes that convergence in probability does not imply mean-square convergence.

Example 6 Let X_1, X_2, \dots be a sequence of discrete random variables with probability mass functions

$$p_{X_n}(x) = \begin{cases} 1 - \frac{1}{n} & \text{if } x = 0 \\ \frac{1}{n} & \text{if } x = n \\ 0 & \text{otherwise} \end{cases}$$

We now observe that, for any $\epsilon > 0$,

$$P(|X_n - 0| > \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

thus X_n converges in probability to zero. However,

$$E(X_n - 0)^2 = 0^2(1 - 1/n) + n^2/n \not\rightarrow 0 \text{ as } n \rightarrow \infty.$$

The following theorem characterises the structure of a random variable whose variance is zero.

Theorem 31 Let X be a random variable with mean m . If $\text{Var } X = 0$, then

$$P(X = m) = 1;$$

that is, X is constant with probability one.

Proof Suppose there exists an $\epsilon > 0$ such that

$$P(|X - m| > \epsilon) = \int_{|x-m|>\epsilon} f_X(x) dx > 0.$$

But then

$$\begin{aligned} E(X - m)^2 &= \int_{-\infty}^{\infty} (x - m)^2 f_X(x) dx \\ &\geq \int_{|x-m| \geq \epsilon} (x - m)^2 f_X(x) dx \\ &\geq \epsilon^2 \int_{|x-m| \geq \epsilon} f_X(x) dx > 0, \end{aligned}$$

resulting in a contradiction to the hypothesis. \square

One of the most well known and useful results of probability theory is the following theorem.

Theorem 32 The weak law of large numbers. *Let X_1, X_2, \dots be an arbitrary sequence of random variables with expectations EX_1, EX_2, \dots . Suppose further than the random variable $\sum_{i=1}^n X_i$ has finite variance for every n . Suppose*

$$\lim_{n \rightarrow \infty} \text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = 0.$$

Then for any $\epsilon > 0$,

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n (X_i - EX_i) \right| \geq \epsilon \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Proof Let

$$Y_n = \frac{1}{n} \sum_{i=1}^n (X_i - EX_i).$$

Then

$$EY_n = \frac{1}{n} \sum_{i=1}^n E[X_i - EX_i] = 0.$$

Applying Chebyshev's inequality to this random variable states that

$$P(|Y_n| \geq \epsilon) \leq \frac{\text{Var}(Y_n)}{\epsilon^2};$$

but, by Theorem 1,

$$\text{Var}(Y_n) = \text{Var} \left(\frac{1}{n} \sum_{i=1}^n (X_i - EX_i) \right) = \text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right).$$

By hypothesis, $\text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) \rightarrow 0$ as $n \rightarrow \infty$. Therefore,

$$P(|Y_n| \geq \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

□

The following corollary is also sometimes (usually?) called the weak law of large numbers.

Corollary 3 *Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables with finite mean $EX_i = m$ and finite variance σ^2 . Then, for any $\epsilon > 0$,*

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - m \right| \geq \epsilon \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Proof From (29) we have $\text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{\sigma^2}{n}$. Hence, applying Theorem 32 yields the result. □

8 The Strong Law of Large Numbers

The law of large numbers plays a fundamental role in the theory of probability and its applications. Suppose, however, that only the weak law of large numbers (convergence in probability) held for identically distributed random variables X_i with finite expectations, i.e., that $P\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - m\right| \geq \epsilon\right] \rightarrow 0$ as $n \rightarrow \infty$. It is important to appreciate the fact, however, that this form of convergence *does not assure convergence of individual realizations*. In other words, for any given elementary event $\omega \in \Omega$, we have no assurance that $\frac{1}{n}\sum_{i=1}^n X_i(\omega) \rightarrow m$ as $n \rightarrow \infty$. In fact, it can be shown that large values of $|(X_1 + \cdots + X_n)/n - m|$ can occur for infinitely many n . Under these circumstances, it would seem doubtful that the arithmetic mean of the realizations $X_i(\omega)$ could be taken as a reliable approximation of the mean. This could be very annoying, because in many practical engineering applications we have only one realization to work with (i.e., only one ω) and we need to calculate averages that converge as determined by actual calculations.

Fortunately, however, there is a stronger version of the law of large numbers that does assure convergence for individual realizations. Convergence with probability one applies to individual realizations, while convergence in probability does not. Convergence with probability one is as close as we can get to the usual notion of convergence of a sequence of numbers since it says that the limiting sample average $\frac{1}{n}\sum_{i=1}^n X_i(\omega)$ converges to m for all elementary events $\omega \in \Omega$ in a set of probability one.

The most general form of the strong law of large numbers requires only a finite mean, but the proof of this version is quite difficult. Consequently, we state and prove a version of the theorem with the additional assumption of a finite fourth moment (this is not a very restrictive assumption, and is certainly adequate for engineering applications). Before stating and proving the strong law of large numbers, it will be helpful to establish the following preliminary lemma, which is one of the most frequently referred to results of probability theory.

Lemma 2 The Borel-Cantelli Lemma. *Let E_1, E_2, \dots denote a sequence of events. If*

$$\sum_{i=1}^{\infty} P(E_i) < \infty,$$

then The probability that an infinite number of the E_i 's occur is zero. (Notationally: $P[E_n \text{ i. o.}] = 0$, where "i. o." denotes "infinitely often.")

Proof The event $[E_n \text{ i. o.}]$ means that for every n there exists a $k \geq n$ such that E_k occurs; that is,

$$[E_n \text{ i. o.}] = \bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} E_k.$$

This follows since if an infinite number of the E_k occur, then $\bigcup_{k=m}^{\infty} E_k$ occurs for each m and thus $\bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} E_k$ occurs. On the other hand, if $\bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} E_k$ occurs, then $\bigcup_{k=m}^{\infty} E_k$ occurs for each m , and thus for each m at least one of the E_k occurs where $k \geq m$ and hence an infinite number of the E_k 's occur. We now observe that, since $\bigcup_{k=m}^{\infty} E_k$, $m \geq 1$ is a decreasing sequence of events, it follows from Theorem 4 of the Lecture Notes for Chapter 2 that

$$\begin{aligned} P\left(\bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} E_k\right) &= P\left(\lim_{m \rightarrow \infty} \bigcup_{k=m}^{\infty} E_k\right) \\ &= \lim_{m \rightarrow \infty} P\left(\bigcup_{k=m}^{\infty} E_k\right) \\ &\leq \lim_{m \rightarrow \infty} \sum_{k=m}^{\infty} P(E_k) \\ &= 0, \end{aligned}$$

by hypothesis, which establishes the lemma. \square

Theorem 33 The strong law of large numbers. *Let X_1, X_2, \dots be a sequence of independent identically distributed random variables with expectation $EX_i = m$ and finite fourth moment, that is, $E[X_i^4] < \infty$. Then*

$$P\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = m\right) = 1.$$

Proof Without loss of generality, we may assume that $m = 0$ since, once we show it is true for the zero-mean case we can extend to random variables with arbitrary mean by applying the result to the process $Y_i = X_i - m$. Let

$$S_n = \sum_{i=1}^n X_i.$$

The fourth moment of S_n is

$$\begin{aligned} ES_n^4 &= E[(X_1 + \dots + X_n)(X_1 + \dots + X_n) \\ &\quad \times (X_1 + \dots + X_n)(X_1 + \dots + X_n)]. \end{aligned}$$

The right side of this expression contains terms of the form

$$X_i^4, \quad X_i^3 X_j, \quad X_i^2 X_j^2, \quad X_i^2 X_j X_k, \quad \text{and} \quad X_i X_j X_k X_\ell,$$

where $i, j, k,$ and ℓ are all different. Applying the independence hypothesis yields

$$\begin{aligned} E(X_i^3 X_j) &= E(X_i^3) E X_j = 0 \\ E(X_i^2 X_j X_k) &= E(X_i^2) E X_j E X_k = 0 \\ E(X_i X_j X_k X_\ell) &= E X_i E X_j E X_k E X_\ell = 0, \end{aligned}$$

thus the only terms that do not vanish are terms of the form X_i^4 and $X_i^2 X_j^2$. For any fixed i and j there will be $\binom{4}{2} = 6$ terms in the expansion that equal $X_i^2 X_j^2$. Thus, expanding the above product and taking expectations yields

$$\begin{aligned} E S_n^4 &= n E X_i^4 + 6 \binom{n}{2} E(X_i^2 X_j^2) \\ &= nK + 3n(n-1) E X_i^2 E X_j^2, \end{aligned}$$

where $K = E[X_i^4]$ and we once again apply the independence hypothesis.

Since

$$0 \leq \text{Var}(X_i^2) = E X_i^4 - (E X_i^2)^2,$$

it follows that

$$(E X_i^2)^2 \leq E X_i^4 = K,$$

thus, dividing both sides by n^4 , we have

$$E \left(\frac{S_n^4}{n^4} \right) \leq \frac{K}{n^3} + \frac{3K}{n^2}.$$

Consequently,

$$E \left(\sum_{n=1}^{\infty} \frac{S_n^4}{n^4} \right) = \sum_{n=1}^{\infty} E \left(\frac{S_n^4}{n^4} \right) < \infty. \quad (30)$$

Now, for any $\epsilon > 0$, it follows from the Markov inequality that

$$P \left(\frac{S_n^4}{n^4} > \epsilon \right) \leq \frac{E(S_n^4/n^4)}{\epsilon}$$

and thus, from (30),

$$P \left(\frac{S_n^4}{n^4} > \epsilon \right) < \infty.$$

We now apply the Borel-Cantelli lemma to establish that, with probability one, $[S_n^4/n^4 > \epsilon \text{ i. o.}]$. Since this is true for any $\epsilon > 0$, we conclude that

$$\lim_{n \rightarrow \infty} \frac{S_n}{n} = 0 \quad \text{with probability one.}$$

□

9 Mean Value, Correlation and Covariance Functions

Definition 23 Let $\{X_t, -\infty < t < \infty\}$ be a stochastic process. The mean value function for this process is defined as $m_X(t) = EX_t$. Here, we allow for the possibility that the expectation of X_t may be time-varying. The **autocorrelation function** or **correlation function** $R_X : \mathfrak{R} \times \mathfrak{R} \rightarrow \mathfrak{R}$ is given by

$$R_X(t, s) = E(X_t X_s), \quad t \in \mathfrak{R}, s \in \mathfrak{R},$$

and the **autocovariance function** or **covariance function** $K_X : \mathfrak{R} \times \mathfrak{R} \rightarrow \mathfrak{R}$ is given by

$$K_X(t, s) = \text{Cov}(X_t, X_s) = E((X_t - m_X(t))(X_s - m_X(s))), \quad t \in \mathfrak{R}, s \in \mathfrak{R}.$$

It is easy to verify that

$$K_X(t, s) = R_X(t, s) - m_X(t)m_X(s).$$

□

The autocorrelation function, therefore, admits an interpretation as the inner product of a stochastic process with a time-shifted version of itself.

Example 7 Consider the process $\{X_t, t \geq 0\}$ defined by

$$X_t = Y + Zt$$

where Y and Z are random variables with specified distributions. The sample functions of this process are straight lines. This process has the mean value function

$$m_X(t) = EY + EZt,$$

correlation function

$$R_X(s, t) = EY^2 + E(YZ)(t + s) + EZ^2 ts,$$

and covariance function

$$K_X(t, s) = \text{Var}(Y) + \text{Cov}(Y, Z)(t + s) + \text{Var}(Z)ts.$$

If Y and Z are uncorrelated with zero-means and unit variances (that is, $EY = EZ = EYZ = 0$ and $EY^2 = EZ^2 = 1$), then

$$m_X(t) = 0$$

and

$$R_X(t, s) = K_X(t, s) = 1 + ts.$$

Observe that we are able to calculate the mean value, correlation, and covariance functions of $\{X_t, t \geq 0\}$ using only the means, variances, and covariances of the random variables Y and Z . Specifically, we have used only first- and second-moment properties of the distribution of the process.

10 Stationarity

Definition 24 Let $\{X_t, t \in I\}$ be a stochastic process. We will assume that the parameter set I is *linear*, in the sense that if $t, s \in I$, then $t + s \in I$. The process is said to be **stationary**, or, (to be more emphatic) **strictly stationary** if all finite joint distributions are invariant to a shift in the parameter; that is, for any finite set $\{t_i, i = 1, 2, \dots, k\} \subset I$ and any $s \in I$, the joint probability density function satisfies

$$f_{X_{t_1} X_{t_2} \dots X_{t_k}}(x_1, x_2, \dots, x_k) = f_{X_{t_1+s} X_{t_2+s} \dots X_{t_k+s}}(x_1, x_2, \dots, x_k). \quad (31)$$

□

It is useful to consider the special case of $k = 1$, for then we observe that

$$f_{X_t}(x) = f_{X_{t+s}}(x)$$

for all t and s , which means that the marginal density of the process is independent of the parameter t . Consequently, the mean value of the process is constant:

$$m_X(t) = \int_{-\infty}^{\infty} xp_{X_t}(x)dx = \int_{-\infty}^{\infty} xp_{X_{t+s}}(x)dx = \text{constant}.$$

It is also useful to examine the structure of the joint density when $k = 2$; namely

$$f_{X_{t_1}X_{t_2}}(x_1, x_2) = f_{X_{t_1+s}X_{t_2+s}}(x_1, x_2).$$

Since this condition must hold for all values of s , it must hold for $s = -t_1$, consequently

$$f_{X_{t_1}X_{t_2}}(x_1, x_2) = f_{X_0X_{t_2-t_1}}(x_1, x_2),$$

and we see that, in this case, the joint density depends only on the difference, $t_2 - t_1$. For a stationary stochastic process, therefore, the correlation function becomes

$$R_X(t, s) = E(X_t X_s) = E(X_0 X_{s-t}) = R_X(0, s - t)$$

and the covariance function becomes

$$K_X(t, s) = E(X_t X_s) - m_X(t)m_X(s) = E(X_0 X_{s-t}) - m_X^2 = K_X(0, s - t).$$

Since the correlation and covariances functions of stationary processes are functions only of the difference $s - t$, we usually abuse the notation by writing these functions as $R_X(\tau)$ and $K_X(\tau)$, where $\tau = s - t$.

Stationary is a very strong property for a process to possess, and it may be difficult to ensure that a given process is stationary. The problem is, we must ensure that **all** finite joint probability density functions satisfy (31). There is, however, a weaker notion of stationary that is often useful.

Definition 25 A random process $\{X_t, t \in I\}$, be random process with I linear, is **weakly stationary**, or **wide-sense stationary**, or **covariance stationary** if it has finite second moments (that is, $E|X_t|^2 < \infty$ for all $t \in I$), has constant mean ($EX_t = m_X$, a constant) and $R_X(t, s)$ and $K_X(t, s)$ depend only on the difference $s - t$. \square

Example 8 The Ornstein-Uhlenbeck process.

Let $\{X_t, -\infty < t < \infty\}$ be a zero-mean Gaussian process with

$$K_X(t, s) = \sigma^2 e^{-a|t-s|}.$$

This process is exponentially correlated, wide-sense stationary.

Let's look at some of the properties of the correlation function of a wide-sense stationary process.

- $R_X(\tau) = R_X(-\tau)$. To see this, we note that

$$E(X_t X_{t+\tau}) = E(\underbrace{X_{t+\tau-\tau}}_{t'} \underbrace{X_{t+\tau}}_{t'}) = E(X_{t'-\tau} X_{t'}).$$

- $R_X(\tau)$ has maximum of $R_X(0)$. To establish this fact, we apply the Schwarz inequality and the definition of wide-sense stationarity to obtain:

$$|R_X(\tau)| = |E(X_t X_{t+\tau})| \leq E^{\frac{1}{2}} X_t^2 E^{\frac{1}{2}} X_t^2 = E X_t^2 = R_X(0).$$

- $R_X(\tau)$ is continuous for all τ if it is continuous at $\tau = 0$. We demonstrate this fact by the following:

$$\begin{aligned} |R_X(\tau + \epsilon) - R_X(\tau)| &= |E(X_t(X_{t+\tau+\epsilon} - X_{t+\tau}))| \\ &\leq E^{\frac{1}{2}} X_t^2 E^{\frac{1}{2}} (X_{t+\tau+\epsilon} - X_{t+\tau})^2 \\ &= E^{\frac{1}{2}} X_t^2 [E X_{t+\tau+\epsilon}^2 - 2E(X_{t+\tau+\epsilon} X_{t+\tau}) + E X_{t+\tau}^2]^{\frac{1}{2}} \\ &= R_X^{\frac{1}{2}}(0) [2R_X(0) - 2R_X(\epsilon)]^{\frac{1}{2}} \\ &= [2R_X(0)(R_X(0) - R_X(\epsilon))]^{\frac{1}{2}} \end{aligned}$$

We see immediately that if $R_X(\cdot)$ is continuous at zero, then the limit of the right side of this equation as $\epsilon \rightarrow 0$ becomes zero, which establishes the desired result.

11 Ergodic Theory

Ergodic theory is the study of the long-term average behavior of processes. We begin by analyzing the behavior of sample averages as the length of the sample becomes infinite. Let $\{X_i, i \geq 0\}$ be a discrete-time random process. We define the sample average of the first n samples as

$$S_n = \frac{1}{n} \sum_{i=0}^{n-1} X_i, \quad n = 1, 2, \dots$$

Let us examine the random process $\{S_n, n \geq 0\}$. Clearly, the mean value is

$$E S_n = \frac{1}{n} \sum_{i=0}^{n-1} E X_i,$$

and the variance is

$$\begin{aligned}
 \sigma_{S_n}^2 &= E(S_n - ES_n)^2 \\
 &= E \left[\frac{1}{n} \sum_{i=0}^{n-1} X_i - \frac{1}{n} \sum_{i=0}^{n-1} EX_i \right]^2 \\
 &= \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \underbrace{E[(X_i - EX_i)(X_j - EX_j)]}_{K_X(i,j)} \\
 &= \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K_X(i, j). \tag{32}
 \end{aligned}$$

Now suppose $EX_i = m_X$, a constant, that $\sigma_{X_i}^2 = \sigma_X^2$, a constant, and that $\{X_i\}$ is an uncorrelated sequence:

$$K_X(i, j) = \begin{cases} \sigma_X^2 & i = j \\ 0 & i \neq j \end{cases}$$

Then

$$\sigma_{S_n}^2 = \frac{1}{n} \sigma_X^2 \rightarrow 0 \text{ as } n \rightarrow \infty,$$

so we get

$$\text{l. i. m.}_{n \rightarrow \infty} S_n = m_X. \tag{33}$$

This result is a version of the well-known law of large numbers, which says, essentially, that the long-term average of a sequence of random samples converges to the (common) expectation of the population. This is our first example of an ergodic theorem—a very weak one, since we assumed considerable structure for the process $\{X_i\}$. We are interested in determining whether we can make stronger statements regarding the long-term behavior of sample averages.

First of all, we note that for the example above, the process S_n converged in mean square to a *number*. But was it just a number? In fact, a little thought reveals that the process converged to a random variable, but it was a degenerate one—one with zero variance. The left-hand side of (33) is a random variable, and therefore so is the right-hand side. So we are justified in writing

$$\begin{aligned}
 E \left(\text{l. i. m.}_{n \rightarrow \infty} S_n \right) &= m_X \\
 \text{Var} \left(\text{l. i. m.}_{n \rightarrow \infty} S_n \right) &= 0.
 \end{aligned}$$

Let us now retain the interpretation that the sample average converges to a random variable, but let us relax the requirement that the variance of this random variable be zero.

As an example of such a situation, consider the following compound experiment. Suppose I have three coins, one fair, one with $P(H) = p$, and one with $P(H) = q$, with $p \neq q$. At the beginning of time I flip the fair coin once; if heads shows, I choose the coin with $P(H) = p$, and if tails shows, I choose the coin with $P(H) = q$. I then start tossing the chosen coin repeatedly. For the i -th toss, let $X_i = 1$ if heads occurs, and $X_i = 0$ if tails shows. Also, define S_n as the sample average:

$$S_n = \frac{1}{n} \sum_{i=0}^{n-1} X_i.$$

Clearly, this experiment satisfies the hypotheses of our previous example, so it is evident that S_n converges in mean-square to *something*, but that something is **not** a constant. If, at the beginning of time, I selected the coin with bias p , then $S_n \xrightarrow{m.s} p$, otherwise $S_n \xrightarrow{m.s} q$. If I let \bar{X} denote the limiting random variable, then I can write

$$\bar{X} = \begin{cases} p & \text{with probability } \frac{1}{2} \\ q & \text{with probability } \frac{1}{2} \end{cases}.$$

Note, with this example, that the thing the process converges to is not the expectation, which would be $\frac{p+q}{2}$, even though this is a stationary process.

There are many ergodic theorems, each with slightly different hypotheses. Here, we present the most simple of such theorems.

Theorem 34 *Let $\{X_n, n \in \mathcal{Z}\}$, where \mathcal{Z} is the set of integers, be a wide-sense stationary discrete-time random process, and suppose*

$$\lim_{n \rightarrow \infty} K_X(n) = 0.$$

Then

$$\text{l. i. m.}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} X_i = m_X,$$

where m_X is the common expectation of $\{X_n\}$.

Proof

We first examine the variance of $S_n = \frac{1}{n} \sum_{i=0}^{n-1} X_i$. Using (32) and the assumption of

weak stationarity,

$$\begin{aligned}
\sigma_{S_n}^2 &= \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K_X(i-j) \\
&= \frac{1}{n} K_X(0) + \frac{2(n-1)}{n^2} K_X(1) + \frac{2(n-2)}{n^2} K_X(2) \\
&\quad + \cdots + \frac{2(n-(n-1))}{n^2} K_X(n-1) \\
&= \frac{1}{n} K_X(0) + \frac{2}{n} \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) K_X(k).
\end{aligned}$$

Now suppose $K_X(n) \rightarrow 0$ as $n \rightarrow \infty$ and fix $\epsilon > 0$. Then there exists an integer N such that $k > N$ implies that $|K_X(k)| < \frac{\epsilon}{2}$. Then for $n > N$, and using the fact that $|K_X(k)| \leq K_X(0)$,

$$\begin{aligned}
\sigma_{S_n}^2 &= \frac{1}{n} K_X(0) + \frac{2}{n} \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) K_X(k) \\
&\leq \frac{1}{n} K_X(0) + \frac{2}{n} \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) |K_X(k)| \\
&\leq \frac{1}{n} K_X(0) + \frac{2}{n} \sum_{k=1}^N \left(1 - \frac{k}{n}\right) |K_X(0)| + \frac{1}{n} \sum_{k=N+1}^{n-1} \left(1 - \frac{k}{n}\right) \epsilon.
\end{aligned}$$

Now observe that

$$\sum_{k=1}^N \left(1 - \frac{k}{n}\right) = N - \frac{1}{n} \sum_{k=1}^N k = N - \frac{1}{n} \frac{(N+1)N}{2} < N,$$

where we have used the fact that

$$\sum_{k=1}^N k = \frac{(N+1)N}{2}.$$

Also, $1 - \frac{k}{n} < 1$. Substituting these bound into the above inequality yields

$$\begin{aligned}
\sigma_{S_n}^2 &< \frac{1}{n} K_X(0) + \frac{2N}{n} K_X(0) + \frac{n}{n} \epsilon \\
&= \left(\frac{1}{n} + \frac{2N}{n}\right) K_X(0) + \epsilon.
\end{aligned}$$

Now let $n \rightarrow \infty$. The right-hand side of this expression converges to ϵ , and this must hold for all $\epsilon > 0$. Thus, $\sigma_{S_n} \rightarrow 0$ as $n \rightarrow \infty$, and

$$\text{l. i. m.}_{n \rightarrow \infty} S_n = EX,$$

as required. □

As discussed in the text, this theorem may be used to characterize processes that are asymptotically uncorrelated. Recall that the simple version of the law of large numbers given earlier was in the context of an uncorrelated process. This theorem permits us to relax that requirement to permit the correlation to decay more gradually. For example, the theorem applies to exponentially correlated processes: $K_X(k) = \sigma_X^2 e^{-\alpha|k|}$.

Many versions of this theorem exist; we have proven only the discrete-time case. It is important to note that this theorem may be applied not only to the mean, but to other functions of the random process. Just as we often find it convenient to approximate the mean of a process by a time average, we may find it convenient to approximate the correlation function of a process by an appropriate time average. We will pursue this problem later on.

As a practical issue, this theorem provides us with a fairly rigorous interpretation of the circumstances under which the common phrase “time-averages may be used to approximate ensemble averages” is valid as well as a precise interpretation of this phrase.

The above ergodic theorem also holds for the continuous-time case, but we will postpone discussing that topic until we define the sense in which the integral of a random process is defined.