Martingale Ideas in Elementary Probability

Lecture course
Higher Mathematics College, Independent University of Moscow
Spring 1966

William Faris
University of Arizona
Fulbright Lecturer, IUM, 1995–1996

Preface

These lecture notes were distributed to students in the second year probability course at the Higher Mathematics College, Independent University of Moscow, during the spring semester, 1996. They are an introduction to standard topics in theoretical probability, including the laws of large numbers and the central limit theorem. The plan was to create a course that would cover this material without being a boring duplication of existing standard courses. Hence came the idea of organizing the course around the concept of martingale. The elementary examples of martingales in the first part of the lectures are obtained by applying gambling schemes to sums of independent random variables. The entire exposition makes no use of the concept of conditional probability and exposition, although these would be central to a more advanced development.

Alexander Shen and Alexei Melnikov were both helpful in the development of the course. In particular, Shen read an early version of the notes and contributed many insights. The students came up with several original solutions of the problems; among the most surprising were those by Kostya Petrov. Several of the students in the course helped with the final version of the notes. These included Alexander Cherepanov, Sam Grushevsky, Kostya Rutin, Dmitry Schwarz, Victor Shuvalov, and Eugenia Soboleva. Victor Shuvalov deserves special mention; he was the organizer and a most able and enthusiastic participant.

The author was a Fulbright Lecturer at the Independent University of Moscow during the 1995–1996 academic year. He thanks the faculty members of the university for their skillful arrangements and generous hospitality. Alexei Roudakov, Yulii Ilyashenko, and Sergei Lando were extraordinarily helpful, and Alexei Sossinsky was particularly ingenious in solving all sorts of problems. Askol'd Khovanski and Nikolai Konstantinov were kind enough to include the author as a participant in their calculus seminar. It was a constant pleasure to deal with such intelligent and well-organized people.

Contents

Lecture 1. Winning a fair game

Summary: Symmetric random walk is perhaps the most basic example of a fair game. For this game the strategy "play until you win a fixed amount" always succeeds in the long run. (Unfortunately one needs infinite reserves and unlimited time to make this work in practice.)

I have no interest in games or gambling. However I am interested in probability models, especially in physics, and hence in the notion of random fluctuation. One of the most important methods of studying random fluctuations is to think of them as successive values of a fair game. Such a game is called a *martingale*, for reasons having to do with certain gambling schemes known by that name.

One of the most important ideas of this theory is that one cannot make gains without risk. We shall see this illustrated over and over in the following lectures.

One unusual feature of these lectures is that I will develop martingale theory without the concept of conditional expectation. Since the emphasis will be on simple concrete examples, there will not be much emphasis on developing the theory of measure and integration. However the basic limit theorems will be presented and illustrated.

Perhaps the most important probability model is that of Bernoulli trials. In this model there is an experiment with various possible outcomes. Each such *outcome* is a sequence of values which are either a success S or a failure F. In other words, an outcome is a function from the indexing set of trials $\{1, 2, 3, \ldots\}$ to the two element set $\{S, F\}$.

We first consider the most symmetric version of the model. An *event* is a set of outcomes to which a probability is assigned. Consider a given function from the first n trials $\{1, 2, \ldots n\}$ to $\{S, F\}$. Consider the event A consisting of all outcomes that agree with this given function on the first n trials. This event is given probability $1/2^n$.

In general, if an event is a finite or countably infinite union of disjoint events, then the probability of this event is the corresponding sum of the probabilities of the constituent events.

Let N_n be the function from the set of outcomes to the natural numbers $0, 1, 2, 3, \ldots$ that counts the number of successes in the first n elements of the outcome. Consider k with $0 \leq k \leq n$. Let $N_n = k$ denote the event consisting of all outcomes for which the value of the function N_n is k. We have the fundamental formula for the probability of k successes in n trials:

$$\mathbf{P}[N_n = k] = \binom{n}{k} \frac{1}{2^n}. \tag{1.1}$$

The derivation of this formula is simple. There is a bijective correspondence between the functions from $\{1, \ldots, n\}$ to $\{S, F\}$ with exactly k successes and the subsets of the set $\{1, \ldots, n\}$ with exactly k elements. The function corresponds to the subset on which the S values are obtained. However the number of k element subsets of an n element set is the binomial coefficient $\binom{n}{k}$. So the event $N_n = k$ is the disjoint union of $\binom{n}{k}$ events each of which have probability $1/2^n$.

If one plots $\mathbf{P}[N_n = k]$ as a function of k for a fixed reasonably large value of n one can already start to get the impression of a bell-shaped curve. However we want to begin with another line of thought.

Let $S_n = 2N_n - n$. This is called *symmetric random walk*. The formula is obtained by counting 1 for each success and -1 for each failure in the first n trials. Thus $S_n = N_n - (n - N_n)$ which is the desired formula. It is easy to see that

$$\mathbf{P}[S_n = j] = \mathbf{P}[N_n = \frac{n+j}{2}] \tag{1.2}$$

when n and j have the same parity.

We can think of S_n as the fortune at time n in a fair game. Of course S_n is a function from the set of outcomes to the integers.

Exercise 1.1. Show that $\mathbf{P}[S_{2m} = 2r] \to 0$ as $m \to \infty$. Hint: Write $\mathbf{P}[S_{2m} = 2r]$ as a multiple of $\mathbf{P}[S_{2m} = 0]$. For each r this multiple approaches the limit one as $m \to \infty$. On the other hand, for each k the sum for $|r| \leq k$ of $\mathbf{P}[S_{2m} = 2r]$ is $\mathbf{P}[|S_{2m}| \leq 2k] \leq 1$.

Exercise 1.2. Fix $k > 0$. Find the limit as $n \to \infty$ of $\mathbf{P}[S_n \geq k]$.

Exercise 1.3. Write $k \mid m$ to mean that k divides m. It is obvious that $\mathbf{P}[2 \mid S_n]$ does not converge as $n \to \infty$. Does $\mathbf{P}[3 \mid S_n]$ converge as $n \to \infty$?

We want to see when we win such a game. Say that we want to win an amount $r \geq 1$. Let T_r be the function from the set of outcomes to the numbers $\{1, 2, 3, \ldots, \infty\}$ defined so that T_r is the least n such that $S_n = r$. Thus T_r is the time when we have won r units.

Figure 1.

Example 1.1. Take $r = 1$. Then clearly $\mathbf{P}[T_1 = 1] = 1/2$, $\mathbf{P}[T_1 = 3] = 1/8$, $\mathbf{P}[T_1 = 5] = 2(1/32)$, and $\mathbf{P}[T_1 = 7] = 5(1/128)$. What is the general formula? What is the sum of these numbers?

Let us first look at the sum. This is the chance of ever getting from a fortune of zero to a fortune of $r > 0$:

$$\rho_r = \mathbf{P}[T_r < \infty] = \sum_{n=0}^{\infty} \mathbf{P}[T_r = n]. \tag{1.3}$$

We claim that $\rho_r = 1$; if you wait long enough you will win the game.

Here is a proof. The proof will use sophisticated reasoning, but the intuition is nevertheless compelling. We will later give other proofs, so one should not have too many worries yet about this point. The proof is correct, given enough theoretical background.

Clearly $\rho_r = \rho_1^r$, since to win r units one has to have r independent wins of one unit. However using the same reasoning

$$\rho_1 = \frac{1}{2} + \frac{1}{2}\rho_2 = \frac{1}{2} + \frac{1}{2}\rho_1^2. \tag{1.4}$$

This is because either at the first stage one immediately wins, or one immediately loses and then has to gain back two units. This is a quadratic equation that may be solved for ρ_1. The only solution is $\rho_1 = 1$.

This is a dramatic conclusion. In order to win in a fair game, one has only to wait until luck comes your way. It almost surely will.

Exercise 1.4. Show if you start a symmetric random walk at an arbitrary integer point, then with probability one it will eventually return to that point. This is the property of *recurrence*.

Exercise 1.5. Let $m_r \geq 0$ be the expected amount of time to get from an initial fortune of 0 to r. Clearly $m_r = rm_1$. Derive the formula

$$m_1 = \frac{1}{2}1 + \frac{1}{2}(1 + m_2). \tag{1.5}$$

Solve the resulting linear equation.

Exercise 1.6. There is a non-symmetric version of the probability model in which the probability of success on each trial is p and the probability of failure on each trial is q and $p + q = 1$. The probability of the set of all infinite sequences of successes and failures that coincide on the first n trials with a fixed finite sequence of length n is $p^k q^{n-k}$, where k is the number of successes in the finite sequence. Show that for this model $\rho_r = \rho_1^r$ and

$$\rho_1 = p + q\rho_1^2. \tag{1.6}$$

Find both solutions of this quadratic equation. Guess which solution is the correct probability to gain one unit.

Exercise 1.7. In the preceding problem it is obvious what the correct solution must be when $p > 1/2$. A student in the course pointed out the following remarkable way of finding the correct solution when $p < 1/2$. Let ρ_1 be the probability that one eventually gains one unit, and let ρ_{-1} be the probability that one eventually loses one unit. Show by examining the paths that $\rho_1 = p/q\rho_{-1}$. On the other hand, one can figure out the value of ρ_{-1}.

Exercise 1.8. A non-symmetric random walk with $p \neq 1/2$ is not recurrent. Find the probability of eventual return to the starting point.

Exercise 1.9. In the non-symmetric version derive the formulas $m_r = rm_1$ and

$$m_1 = p1 + q(1 + 2m_1). \tag{1.7}$$

Find both solutions of this linear equation! Guess which solution is the correct expected number of steps to gain one unit.

Exercise 1.10. Show that when $p > q$ the expected time $m_1 = \sum_n n\mathbf{P}[T_1 = n] < \infty$. Hint: Compare with the series $\rho_1 = \sum_n \mathbf{P}[T_1 = n] = 1$ in the case $p = 1/2$.

Lecture 2. The reflection principle

Summary: The reflection principle is a special technique that applies to symmetric random walk. It makes it possible to calculate the probability of winning a specified amount in a fixed amount of time.

Recall that S_n for $n = 1, 2, 3, \ldots$ is a symmetric random walk. If $r \geq 1$, then T_r is the first time n that $S_n = r$.

There is a remarkable explicit formula for probabilities associated with T_r. It is

$$\mathbf{P}[T_r = n] = \frac{r}{n}\mathbf{P}[S_n = r]. \tag{2.1}$$

As time goes on, the probability that the walk visits r for the first time at n is a increasingly small proportion of the probability that the walk visits r at time n.

The formula is remarkable, but it does not cast much light on why the sum over n is one. Let us derive an equivalent version of the formula; that will give us a better idea of its meaning.

The way to get this formula is to derive instead a formula for $\mathbf{P}[T_r \leq n]$. The derivation is based on breaking the event into two parts. We write

$$\mathbf{P}[T_r \leq n] = \mathbf{P}[T_r \leq n, S_n > r] + \mathbf{P}[T_r \leq n, S_n \leq r] = \mathbf{P}[S_n > r] + \mathbf{P}[T_r \leq n, S_n \leq r]. \tag{2.2}$$

Now we use the *reflection principle*:

$$\mathbf{P}[T_r \leq n, S_n \leq r] = \mathbf{P}[T_r \leq n, S_n \geq r]. \tag{2.3}$$

Figure 2.

However this immediately gives

$$\mathbf{P}[T_r \leq n] = \mathbf{P}[S_n > r] + \mathbf{P}[S_n \geq r]. \tag{2.4}$$

This is the desired formula.

We can now see directly why $\mathbf{P}[T_r < \infty] = 1$. We simply take the limit in the above formula and note that the right hand side goes to $1/2 + 1/2 = 1$.

Exercise 2.1. Give a completely elementary derivation of this limiting behavior of the right hand side.

Next we calculate $\mathbf{P}[T_r = n]$. We could do this by a subtraction, but it is more instructive to do this directly by the reflection principle. We have that

$$\mathbf{P}[S_{n-1} = r-1, Y_n = 1] = \mathbf{P}[S_{n-1} = r-1, Y_n = 1, T_r = n] + \mathbf{P}[S_{n-1} = r-1, Y_n = 1, T_r < n]. \tag{2.5}$$

However the event $S_{n-1} = r - 1, Y_n = 1, T_r = n$ is the same as the event $T_r = n$. Furthermore, by the reflection principle

$$\mathbf{P}[S_{n-1} = r - 1, Y_n = 1, T_r < n] = \mathbf{P}[S_{n-1} = r + 1, Y_n = -1, T_r < n], \tag{2.6}$$

since for each path that crosses r before n there is a corresponding path that is reflected across r after the crossing time. Furthermore the event that $S_{n-1} = r+1, Y_n = -1, T_r < n$ is the same event as $S_{n-1} = r + 1, Y_n = -1$, since a path that takes on the value $r + 1$ at $n - 1$ has to cross r before n. So we obtain

$$\mathbf{P}[S_{n-1} = r - 1, Y_n = 1] = \mathbf{P}[T_r = n] + \mathbf{P}[S_{n-1} = r + 1, Y_n = -1]. \tag{2.7}$$

We can write this in the final form

$$\mathbf{P}[T_r = n] = \frac{1}{2}\mathbf{P}[S_{n-1} = r - 1] - \frac{1}{2}\mathbf{P}[S_{n-1} = r + 1]. \tag{2.8}$$

This is a particularly elegant formulation of the result. Notice that it expresses a probability as a difference of probabilities.

Exercise 2.2. Show that this is equivalent to the formula for $\mathbf{P}[T_r = n] = r/n\mathbf{P}[S_n = r]$ given before.

Exercise 2.3. Consider the case of non-symmetric random walk. Show that the formula $\mathbf{P}[T_r = n] = r/n\mathbf{P}[S_n = r]$ of the preceding exercise is also true for this case. Hint: This can be done with little or no computation.

Exercise 2.4. Consider the case of non-symmetric random walk. Find the formula for $\mathbf{P}[T_r = n]$ as a difference of probabilities.

Lecture 3. Martingales

Summary: A martingale is a fair game. One can construct interesting examples of martingales by combining symmetric random walk with a gambling scheme.

The symmetric random walk is an example of a kind of fair game called a *martingale*. We now give examples of other related martingales. Rather than define the general concept of martingale at this point, we will define an elementary class of martingales. These are the martingales that are derived from symmetric random walk by a gambling scheme.

In each case we will have a sequence $X_n = x_0 + g_n(Y_1, \ldots, Y_n)$ of functions of the first n steps in a random walk. We think of X_n as the value of a game at stage n where the starting capital is x_0. The game is said to be *fair* if for each n the sum

$$\frac{1}{2^n} \sum_{y_1=\pm 1, \ldots, y_n=\pm 1} g_n(y_1, \ldots, y_n) = 0. \tag{3.1}$$

This just says that the expected value of the game at each stage is the starting capital.

Consider the steps Y_i in the symmetric random walk, so $Y_i = \pm 1$ depending on whether the ith trial is a success or a failure. Let x_0 be the initial capital and set

$$X_n = x_0 + W_1 Y_1 + W_2 Y_2 + \cdots W_n Y_n, \tag{3.2}$$

where

$$W_i = f_i(Y_1, \ldots, Y_{i-1}). \tag{3.3}$$

This is the *elementary martingale* starting at x_0 and obtained by the *gambling scheme* defined by the f_i. The idea is that in placing the next bet at stage i one can make use of the information obtained by examining the results of play at stages before i. In many examples we take the starting capital $x_0 = 0$.

Exercise 3.1. Show that an elementary martingale is a fair game.

Example 3.1. The symmetric random walk S_n is an elementary martingale.

Example 3.2. Let $r \geq 1$ and let T_r be the time that the symmetric random walk hits r. Let $X_n = S_n$ for $n \leq T_r$ and let $X_n = r$ for $n > T_r$. This is the walk *stopped* at r. It is also an elementary martingale. This is a game that is played until a desired level of winnings is achieved.

Figure 3.

There is a general theorem that says that every martingale that is bounded above by some fixed constant must converge with probability one. (It is also true that every martingale that is bounded below by some fixed constant must converge with probability one.) This example illustrates the theorem, since as we have seen the probability is one that the martingale X_n converges to the constant value r as $n \to \infty$.

Notice that even though this game is fair at each n, it is not fair in the limit as $n \to \infty$. In this limit there is a sure gain of r. The explanation for this remarkable effect is that the martingale is not bounded below. The effect of a lower bound will be illustrated by the next example.

Example 3.3. Let $q \geq 1$ and let T be the first time that the symmetric random walk hits either $-q$ or r. Let $Z_n = S_n$ for $n \leq T$ and let $Z_n = S_T$ for $n > T$. This is the random walk stopped at either $-q$ or r. It is also a martingale.

Figure 4.

There is another general theorem that says that a martingale that is bounded both below and above not only converges with probability one, but there is a sense in which the game stays fair in the limit. This example illustrates this result, since (as we shall see) the limit of Z_n as $n \to \infty$ is $-q$ with probability $r/(r+q)$ and r with probability $q/(r+q)$.

Examples 3.2 and 3.3 above are obtained by the same construction. Let T be the appropriate stopping time. Take $W_i = 1$ if $i \leq T$ and $W_i = 0$ if $i > T$. Notice that one can see whether $i > T$ by looking at the values of $Y_1, \ldots, Y_{i-1}$.

Exercise 3.2. Can one see whether $i \leq T$ by looking at the values of $Y_1, \ldots, Y_{i-1}$?

Exercise 3.3. Can one see whether $i \geq T$ by looking at the values of $Y_1, \ldots, Y_{i-1}$?

Exercise 3.4. Say that one were fortunate enough to have miraculous schemes in which W_i is allowed to be a function of $Y_1, \ldots, Y_n$. Show that the resulting X_n game could be quite unfair.

Now we give some more examples of martingales.

Example 3.4. Let $W_i = 2S_{i-1}$ be the gambling scheme. Then

$$X_n = 2S_1Y_2 + 2S_2Y_3 + \cdots 2S_{n-1}Y_n = S_n^2 - n. \tag{3.4}$$

Exercise 3.5. Prove the last equality.

This example shows that $S_n^2 - n$ is also a fair game. This is perhaps one of the most fundamental and useful principles of probability: random fluctuations of a sum of independent variables grow on the average so that the square of the displacement is proportional to the number of trials.

Exercise 3.6. Prove that

$$\frac{1}{2^n} \sum_{y_1=\pm 1,\ldots,y_n=\pm 1} (y_1 + \cdots + y_n)^2 = n. \tag{3.5}$$

Example 3.5. Let $W_i = 1/i$. Then

$$X_n = Y_1 + \frac{1}{2}Y_2 + \frac{1}{3}Y_3 + \cdots + \frac{1}{n}Y_n. \tag{3.6}$$

This example will turn out to be fundamental for understanding the "law of averages." It is not a bounded martingale, since if the Y_i were all 1 the series would be a divergent series. However we shall see later that the variance is bounded, and under this circumstances the martingale must converge. Therefore we see that the sum

$$X = Y_1 + \frac{1}{2}Y_2 + \frac{1}{3}Y_3 + \cdots + \frac{1}{n}Y_n + \cdots \tag{3.7}$$

converges with probability one.

Exercise 3.7. In the preceding example, calculate $\bar{Y}_n = X_n - (1/n)(X_1 + \cdots + X_{n-1})$ in terms of $Y_1, \ldots, Y_n$.

Exercise 3.8. Does $\bar{Y}_n$ arise as a martingale from the gambling scheme construction?

Exercise 3.9. Let $x_n \to x$ as $n \to \infty$. Find the limit of $z_n = (1/n)(x_1 + \cdots + x_{n-1})$ as $n \to \infty$.

Example 3.6. Here is an example of the kind of gambling scheme that was called a "martingale". Let $Y_i = \pm 1$ be the steps in a symmetric random walk. Let T be the first i such that $Y_i = 1$. Let $T \wedge n$ the minimum of T and n. Then

$$Z_n = Y_1 + 2Y_2 + 4Y_3 + \ldots 2^{T\wedge n-1} Y_{T\wedge n}. \tag{3.8}$$

Thus one doubles the bet until a final win.

If $T \leq n$ then $Z_n = 1$, since the last gain more than compensates for all the previous losses. This has probability $1 - 1/2^n$. However, if $T > n$, then $Z_n = 1 - 2^n$, which is a catastrophe. This has probability $1/2^n$. The game is fair at each n. However the limit as $n \to \infty$ is 1, so in the limit it is no longer fair, just as in the case of random walk. (It converges faster than the random walk, but the risks of a long wait are also greater.)

Example 3.7. Consider repeated trials of the gambling martingale of the last example with $n = 1, 2, 3, \ldots$ trials. This is a new martingale $Z_1, Z_2, Z_3, \ldots$. Observe that $Z_n = 1$

with probability $1 - 1/2^n$ and $Z_n = 1 - 2^n$ with probability $1/2^n$. Let $S_n = Z_1 + \cdots + Z_n$ be the accumulated earnings. Let A be the event that $S_n \to \infty$ as $n \to \infty$. Even though this is a fair game, the probability of A is $\mathbf{P}[A] = 1$.

Figure 5.

In order to see this, let B be the event that Z_n are eventually one for all large enough n. Clearly $B \subset A$. Furthermore, the complement B^c consists of all outcomes for which $Z_n \neq 1$ for arbitrarily large n. Let C_k be the event that there exists an $n \geq k$ such that $Z_n \neq 1$. Then for each k we have $B^c \subset C_k$. Let C_{kn} be the event that n is the first time $n \geq k$ such that $Z_n \neq 1$. Since C_k is the union for $n \geq k$ of the disjoint events C_{kn}, we have

$$\mathbf{P}[C_k] = \sum_{n=k}^{\infty} \mathbf{P}[C_{kn}] \leq \sum_{n=k}^{\infty} \mathbf{P}[Z_n \neq 1] = \frac{1}{2^{k-1}}. \tag{3.9}$$

Hence $\mathbf{P}[B^c] \leq \mathbf{P}[C_k] \leq 1/2^{k-1}$. Since k is arbitrary, we must have $\mathbf{P}[B^c] = 0$. It follows that $\mathbf{P}[B] = 1$, and so also $\mathbf{P}[A] = 1$.

Notice that what makes this possible is the fact that there is no lower bound on the game. The gambler has to be willing initially to go into heavy debt.

Example 3.8. Let N_n be the number of successes in n independent trials when the probability of success on each trial is $1/2$. Let $S_n = 2N_n - n$ be the symmetric random walk. Let p be a number with $0 < p < 1$ and let $q = 1 - p$. Then

$$X_n = (2p)^{N_n}(2q)^{n-N_n} = (\sqrt{4pq})^n \left(\sqrt{\frac{p}{q}}\right)^{S_n} \tag{3.10}$$

is a martingale. (We multiply by $2p$ for each success, by $2q$ for each failure.)

Exercise 3.10. Show that $X_n - X_{n-1} = (p - q)Y_n X_{n-1}$.

Exercise 3.11. Show that X_n arises from the random walk martingale by a gambling scheme.

Exercise 3.12. A martingale that is bounded below converges with probability one. Clearly $X_n \geq 0$. Thus the event A that the limit of the martingale X_n as $n \to \infty$ exists has probability one. That is, for each outcome ω in A there is a number $X(\omega)$ such that the limit as $n \to \infty$ of the numbers $X_n(\omega)$ is $X(\omega)$. Suppose $p \neq 1/2$. Find the values of X on the outcomes in A.

Note: This sort of martingale is important in physics; it is one of the first basic examples in statistical mechanics. In that context we might call S_n the *energy* and write $\sqrt{p/q} = e^{-\beta}$ and call $1/\beta$ the *temperature*. Then $p-q = \tanh(-\beta)$ and $\sqrt{4pq} = 1/\cosh(\beta)$, so this may be written

$$X_n = \frac{e^{-\beta S_n}}{\cosh^n \beta}. \tag{3.11}$$

It is called the *canonical Gibbs density*.

Lecture 4. The probability framework

Summary: The standard framework for probability experiments is a probability space. This consists of a set of outcomes, a σ-field of events, and a probability measure. Given a probability space describing one trial of an experiment, there is a standard construction of a probability space describing an infinite sequence of independent trials.

We now describe the three elements of a probability space in some detail. First one is given a set Ω. Each point ω in Ω is a possible *outcome* for the experiment.

Second, one is given a σ-field of $\mathcal{F}$ subsets of Ω. Each subset A in $\mathcal{F}$ is called an *event*. Recall that a σ-field of subsets is a collection of subsets that includes $\emptyset$ and Ω and is closed under countable unions, countable intersections, and complements.

The event $\emptyset$ is the impossible event, and the event Ω is the sure event. (This terminology derives from the fact that an experiment is sure to have an outcome.)

Third and finally, one is given a probability measure $\mathbf{P}$. This is a function that assigns to each event A a probability $\mathbf{P}[A]$ with $0 \leq \mathbf{P}[A] \leq 1$. This is the *probability* of the event A. The probability measure must satisfy the properties that $\mathbf{P}[\emptyset] = 0$, $\mathbf{P}[\Omega] = 1$. It also must satisfy *countable additivity*: for every disjoint sequence A_n of events $\mathbf{P}[\bigcup_n A_n] = \sum_n \mathbf{P}[A_n]$. It follows from this that the probability of the complement A^c of an event A is given by $\mathbf{P}[A^c] = 1 - \mathbf{P}[A]$.

Exercise 4.1. Show that $A \subset B$ implies $\mathbf{P}[A] \leq \mathbf{P}[B]$.

Exercise 4.2. Show that $\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B]$.

Exercise 4.3. Events A and B are said to be *independent* if $\mathbf{P}[A \cap B] = \mathbf{P}[A]\mathbf{P}[B]$. Show that if A and B are independent, then so are A and B^c.

Exercise 4.4. Show that if A_n are a sequence of events, the one has *countable subadditivity*

$$\mathbf{P}[\bigcup_n A_n] \leq \sum_n \mathbf{P}[A_n]. \tag{4.1}$$

We define the convergence of a sequence of events. We say that $A_n \to A$ if for every ω there exists an N such that for all $n \geq N$, $\omega \in A_n$ if and only if $\omega \in A$.

Exercise 4.5. Let A_n be an increasing sequence of events. Show that $A_n \to A$ as $n \to \infty$ implies $\mathbf{P}[A_n] \to \mathbf{P}[A]$.

Exercise 4.6. Let A_n be a decreasing sequence of events. Show that $A_n \to A$ as $n \to \infty$ implies $\mathbf{P}[A_n] \to \mathbf{P}[A]$.

Exercise 4.7. Let A_n be a sequence of events. Show that $A_n \to A$ as $n \to \infty$ implies $\mathbf{P}[A_n] \to \mathbf{P}[A]$.

We often indicate an event A (a subset of Ω) by a corresponding condition α (true or false depending on the particular outcome $\omega \in \Omega$). The condition that α is true for the outcome ω is that $\omega \in A$. Conversely, the subset A corresponding to the condition α is $\{\omega \mid \alpha(\omega)\}$. When conditions are combined by the logical operations α or β, α and β, not α, the corresponding sets are combined by the set theoretical operations of union, intersection, and complement: $A \cup B$, $A \cap B$, A^c. Similarly, for an existential condition $\exists n\, \alpha_n$ or a universal condition $\forall n\, \alpha_n$ the corresponding set operations are the union $\bigcup_n A_n$ and intersection $\bigcap_n A_n$.

In this context we often denote a conjunction or the intersection of sets by a comma, so that for instance $\mathbf{P}[A, B]$ is the probability of the conjunction or intersection of the events A and B. Thus we might for instance write a special case of the additivity law as

$$\mathbf{P}[A] = \mathbf{P}[A, B] + \mathbf{P}[A, B^c]. \tag{4.2}$$

An event A is said to be an *almost sure* event if $\mathbf{P}[A] = 1$. A large part of the charm of probability is that one can show that various interesting events are almost sure. Thus an event that is not a logical necessity can nevertheless be a sure bet. Similarly, we could call an event A such that $\mathbf{P}[A] = 0$ an *almost impossible* event; this terminology is not as common, but it is quite natural.

Exercise 4.8. The event that A_n occurs infinitely often as $n \to \infty$ is $\bigcap_k \bigcup_{n \geq k} A_n$. The condition that the outcome $\omega \in \bigcap_k \bigcup_{n \geq k} A_n$ is the same as the condition that $\forall k \exists n\, \omega \in A_n$, so this event is the set of outcomes for which the events in the sequence happen for arbitrarily large index. Prove the first Borel-Cantelli lemma: If $\sum_n \mathbf{P}[A_n] < \infty$, then the event that A_n occurs infinitely often is almost impossible. Hint: For each k, $\mathbf{P}[\bigcap_k \bigcup_{n \geq k} A_n] \leq \mathbf{P}[\bigcup_{n \geq k} A_n]$.

Exercise 4.9. Do Exercise 3.7 using explicitly the first Borel-Cantelli lemma.

When an experiment is performed it has an outcome ω. If A is an event and ω is in A, then the event is said to happen. The probability of an event A is a mathematical prediction about the proportion of times that an event would happen if the experiment were repeated independently many times. Whether or not the event actually happens on any particular experiment is a matter of fact, not of mathematics.

Example 4.1. Discrete space; one trial. Let Ω^1 be a countable set. Let $\mathcal{F}^1$ consist of all subsets of Ω^1. Let $p : \Omega^1 \to [0, 1]$ be a function such that $\sum_{x \in \Omega_1} p(x) = 1$. This is called a *discrete density*. Define

$$\mathbf{P}^1[A] = \sum_{x \in A} p(x). \tag{4.3}$$

This is a basic example of a probability space.

Example 4.2. Continuous space; one trial. Let Ω^1 be the real line. Let $\mathcal{F}^1$ be the smallest σ-field containing all intervals. Let $\rho \geq 0$ be an integrable function such that $\int_{-\infty}^{\infty} \rho(x)\, dx = 1$. This is called a *density*. Let

$$\mathbf{P}^1[A] = \int_A \rho(x)\, dx. \tag{4.4}$$

This is a second basic example.

Independent trial construction

Let Ω^1 be a probability space for one trial of an experiment, say as in example 4.1 or example 4.2. Let Ω^∞ be the set of all functions from the set of trials $\{1, 2, 3, \ldots\}$ to Ω^1. Each outcome ω in Ω^∞ is a sequence of outcomes ω_i for $i = 1, 2, 3, \ldots$ in the space Ω^1. The set Ω^∞ is the set of outcomes for the repeated trials experiment. For each i, let X_i be the function from Ω^∞ to Ω^1 given by $X_i(\omega) = \omega_i$.

If A is an event in $\mathcal{F}^1$ describing what happens on one trial, then for each i the event $X_i \in A$ is an event in the repeated trials experiment defined by what happens on the ith trial. Explicitly this event is the set of all sequences ω such that $X_i(\omega) = \omega_i$ is in A. Let $\mathcal{F}^\infty$ be the smallest σ-field of subsets of Ω^∞ containing all such events $X_i \in A$.

Let $A_1, A_2, A_3, \ldots$ be a sequence of single trial events in $\mathcal{F}^1$. Then the events $X_i \in A_i$ are each in $\mathcal{F}^\infty$ and specify what happens on the ith trial. The event $\bigcap_i [X_i \in A_i]$ is an event in $\mathcal{F}^\infty$ that specifies what happens on each trial. The probability measure $\mathbf{P}^\infty$ for *independent repeated trials* is specified by

$$\mathbf{P}^\infty[\bigcap_i [X_i \in A_i]] = \prod_i \mathbf{P}^1[A_i]. \tag{4.5}$$

This says that the probability of the conjunction or intersection of events defined by distinct trials is the product of the probabilities of the events for the single trial experiments.

Note that if $A_i = \Omega^1$ is the sure event for one trial, then its probability is one, and so it does not contribute to the product. Similarly, the event $X_i \in \Omega^1$ is the sure event for repeated trials, and so it does not change the intersection. This definition is typically used when all but finitely many of the events are the sure event; in this case all but finitely many of the factors in the product are one.

Example 4.3. Discrete space; independent trials. In this case the set of outcomes Ω^∞ consists of all sequences of points each belong to the countable set Ω^1. The probability of the set of all sequences whose initial n values are $x_1, \ldots, x_n$ is the product $p(x_1) \cdots p(x_n)$.

Example 4.4. Continuous space: independent trials. In this case the set of outcomes Ω^∞ consists of all sequences of real numbers. The probability of the set B^n of all sequences ω whose initial n values lie in a set B in n dimensional space is $\mathbf{P}^\infty[B^n] = \int_B \rho(x_1) \cdots \rho(x_n)\, dx_1 \cdots dx_n$.

Example 4.5. Bernoulli trials. This is the special case where the discrete space has two points. We begin with the space $\{S, F\}$ consisting of the possible outcomes from one trial, a success or a failure. The σ-field consists of all four subsets $\emptyset$, $\{S\}$, $\{F\}$, and $\{S, F\}$. The discrete density assigns probability p to S and probability $q = 1 - p$ to F.

Next we consider independent trials of this experiment. An outcome is a sequence of S or F results. Among the events are success on the ith trial and failure on the ith trial. We can also specify what happens on a sequence of trials. If we take any finite subset of the trials and specify success or failure on each of the elements of the subset, then the probability of this is $p^k q^{m-k}$, where k is the number of successes and $m - k$ is the number of failures.

Let N_n be the number of successes on the first n trials. If we ask what is the probability of the event $N_n = k$ of having k successes on the first n trials, then one should realize that this event is the union of $\binom{n}{k}$ events where the subset of the first n trials on which the k successes occur is specified. Thus the probability is obtained by adding $p^k q^{n-k}$ that many times. We obtain the famous *binomial distribution*

$$\mathbf{P}[N_n = k] = \binom{n}{k} p^k q^{n-k}. \tag{4.6}$$

Example 4.6. Normal trials. This is the classical example for continuous variables. Let

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{4.7}$$

This is the *Gaussian* or *normal* density with mean μ and variance σ^2. Consider independent trials where this density is used on each individual trial. An outcome of this experiment is a sequence of real numbers, and the probabilities of events are given by multiple integrals involving this Gaussian density.

This example is much used in statistics, in large part for convenience. However it also has a fundamental theoretical justification. We shall see later that martingales that are the sum of many contributions of comparable magnitude (and consequently do not converge) tend to have their behavior characterized by the Gaussian law.

Exercise 4.10. Discrete waiting times. The probability space for one trial consists of the numbers $\{1, 2, 3, \ldots\}$. Fix p with $0 < p \leq 1$ and let $q = 1 - p$. The probability of $\{r\}$ is pq^{r-1}. This represents the waiting time for the next success. The probability space for repeated trials is all sequences of such numbers. For each sequence ω let $W_k(\omega)$ be the kth number in the sequence. This represents the additional waiting time for the next success after $k - 1$ previous successes. For each trial we have $\mathbf{P}[W_k = r] = pq^{r-1}$. It follows by summation that $\mathbf{P}[W_k > r] = q^r$. Furthermore, by the independent trial construction $\mathbf{P}[W_1 = r_1, \ldots, W_n = r_n] = \mathbf{P}[W_1 = r_1] \cdots \mathbf{P}[W_n = r_n]$.

Let $T_k = W_1 + \cdots + W_k$. This is the total waiting time for the kth success. Find $\mathbf{P}[T_k = r]$. Hint: Sum over all possible values of $W_1, \ldots, W_n$.

Exercise 4.11. Let $N_n = \max\{m \mid T_m \leq n\}$ be the number of successes up to time n. Find $\mathbf{P}[N_n = k]$. Hint: $N_n = k$ is the same as $T_k \leq n, W_{k+1} > n - T_k$. Sum over all possible values of $W_1, \ldots, W_n$.

Exercise 4.12. Continuous waiting times. The probability space for one trial consists of the real numbers $s \geq 0$. Fix $\lambda > 0$. The probability density is $\lambda \exp(-\lambda s)$. The probability of waiting more than t is the integral ds of this from t to infinity, which is $\exp(-\lambda t)$. The probability space for repeated trials is all sequences of positive real numbers. For each sequence ω, let $W_k(\omega)$ be the kth number in the sequence. This is the waiting time for the next arrival after $k-1$ previous arrivals. For each trial we have $\mathbf{P}[s < W_k < s+ds] = \exp(-\lambda s)\, ds$ and consequently $\mathbf{P}[t < W_k] = \exp(-\lambda t)$. Furthermore, the probabilities for distinct trials multiply.

Let $T_k = W_1 + \ldots + W_k$ be the total waiting time for the kth arrival. Show that

$$\mathbf{P}[t < T_k] = \sum_{m=0}^{k-1} \frac{(\lambda t)^m}{m!} e^{-\lambda t}. \tag{4.8}$$

Hint: Show that

$$\begin{aligned} \mathbf{P}[t < T_k] &= \mathbf{P}[t < T_{k-1}] + \mathbf{P}[t - T_{k-1} < W_k, T_{k-1} \leq t] \\ &= \mathbf{P}[t < T_{k-1}] + \int_0^t \mathbf{P}[t - s < W_k, s < T_{k-1} \leq s + ds]. \end{aligned} \tag{4.9}$$

and use $\mathbf{P}[t - s < W_k, s < T_{k-1} \leq s + ds] = \mathbf{P}[t - s < W_k]\mathbf{P}[s < T_{k-1} \leq s + ds]$.

Exercise 4.13. Let $N(t) = \max\{r \mid T_r \leq t\}$ be the number of arrivals up to t. Find $\mathbf{P}[N(t) = k]$.

Lecture 5. Random variables

Summary: A random variable assigns a number to each outcome of the experiment. Every positive random variable has a well-defined expectation. Random variables that are not positive may have well-defined expectations if there is no ambiguity involving infinity minus infinity.

Terminology note: I use the term *positive* to mean greater than or equal to zero. I use *strictly positive* to mean greater than zero. Similarly, *increasing* means that increments are positive; *strictly increasing* means that increments are strictly positive.

Consider a probability space with set of outcomes Ω, σ-field of events $\mathcal{F}$, and probability measure $\mathbf{P}$. A function X from Ω to the real numbers assigns an experimental number $X(\omega)$ to each outcome ω. We want to make probability predictions about the values of such functions X.

Consider an interval I of real numbers. We would like to specify the probability $\mathbf{P}[X \in I]$ of the event that X has values in X. In other words, we would like the set of all outcomes ω such that $X(\omega)$ is in I to be an event in the σ-field $\mathcal{F}$. If this occurs for every interval I, then X is said to be a *random variable*, and $\mathbf{P}[X \in I]$ is well-defined.

When an experiment is performed, it has an outcome ω, and the value of the random variable for that outcome is an experimental number $X(\omega)$. Unfortunately probability theory does not tell us what this number will be.

Example 5.1. Consider the Bernoulli trials example. Let N_n be the function from Ω to the natural numbers that counts the number of successes in the first n trial. Then $N_n = k$ is a condition that specifies an event in $\mathcal{F}$. To see that this is an event in $\mathcal{F}$, one notices that it is a union of $\binom{n}{k}$ disjoint events in $\mathcal{F}$, each of which is obtained as an intersection of events associated with individual trials. Thus we can legitimately compute $\mathbf{P}[N_n = k] = \binom{n}{k}p^k q^{n-k}$. For instance, when $p = 1/2$, $\mathbf{P}[S_7 = 6] = 7/2^7$ and $\mathbf{P}[S_7 \geq 6] = 7/2^7 + 1/2^7 = 1/16$.

On January 1, 1996 I conducted 7 Bernoulli trials and for the outcome ω of that particular experiment $N_7(\omega) = 6$. The event that $N_7 \geq 6$ happened for that ω, but nothing in probability theory could have predicted that.

If X is a random variable defined on a space Ω with probability measure $\mathbf{P}$, then X defines a probability measure $\mathbf{P}^1$ defined for subsets of the real line by the formula by

$$\mathbf{P}^1[I] = \mathbf{P}[X \in I]. \tag{5.1}$$

This new probability measure is called the *distribution* of X. Much of classical probability theory consists of calculating the distributions of various random variables. The distribution of a random variable is nothing more than a summary of the probabilities defined by that random variable in isolation from other random variables.

In some cases one can use a *density* to describe the distribution. This is when the associated probability measure can be represented by integrals of the form $\mathbf{P}^1[I] = \int_I \rho(x)\,dx$.

Exercise 5.1. Let X_1 and X_2 represent the results of the first two trials of the continuous independent trials experiment. Show that the distribution of $X_1 + X_2$ is given by

$$\mathbf{P}^1[I] = \int_I \rho_2(y)\,dy, \tag{5.2}$$

where the density ρ_2 is given by the convolution integral

$$\rho_2(y) = \int_{-\infty}^{\infty} \rho(x)\rho(y-x)\,dx. \tag{5.3}$$

If X is a random variable and f belongs to a very general class of functions (Borel measurable), then $f(X)$ is also a random variable. If X has density ρ, then it may be that $f(X)$ also has a density, but this density must be computed with an awkward change of variable. It is thus often convenient to continue to use the density of X for computations, as in the formula $\mathbf{P}[f(X) \in J] = \int_{\{x \mid f(x) \in J\}} \rho(x)\,dx$.

Each positive random variable $X \geq 0$ has an *expectation* $\mathbf{E}[X]$ satisfying $0 \leq \mathbf{E}[X] \leq \infty$. If X is discrete, that is, if X has only a countable set S of values, then

$$\mathbf{E}[X] = \sum_{x \in S} x \cdot \mathbf{P}[X = x]. \tag{5.4}$$

In the general case, for each $\epsilon > 0$ let X_n be the random variable that has the value $k/2^n$ on the set where X is in the interval $[k/2^n, (k+1)/2^n)$, for $k = 0, 1, 2, 3, \ldots$. Then for each n the random variable X_n has only a countable sequence of values. Furthermore for each outcome the X_n values are increasing to the corresponding X value, so the expectations $\mathbf{E}[X_n]$ are also increasing. We define

$$\mathbf{E}[X] = \lim_n \mathbf{E}[X_n]. \tag{5.5}$$

Exercise 5.2. Let N_n be the number of successes in n Bernoulli trials where the probability of success on each trial is p. Find $\mathbf{E}[N_n]$ from the definition.

Exercise 5.3. Let W be a discrete waiting time random variable with $\mathbf{P}[W > k] = q^k$. Find the expectation of W directly from the definition.

Exercise 5.4. Let $W \geq 0$ be a continuous waiting time random variable such that $\mathbf{P}[W > t] = \exp(-\lambda t)$ with $\lambda > 0$. Find $\mathbf{E}[W]$ from the definition.

Exercise 5.5. Let X be a positive random variable with density ρ. Prove from the definition that $\mathbf{E}[X] = \int_0^\infty x\rho(x)\,dx$.

Exercise 5.6. Let $W \geq 0$ be the continuous waiting time random variable with $\mathbf{P}[W > t] = \exp(-\lambda t)$. Find $\mathbf{E}[W]$ by computing the integral involving the density.

Exercise 5.7. Let $f(X)$ be a positive random variable such that X has density ρ. Prove that $\mathbf{E}[X] = \int_{-\infty}^\infty f(x)\rho(x)\,dx$.

Exercise 5.8. Compute $\mathbf{E}[W^2]$ for the continuous waiting time random variable.

Note: There are positive random variables that are neither discrete nor have a density. However they always have an expectation (possibly infinite) given by the general definition.

It is easy to extend the definition of expectation to positive random variables that are allowed to assume values in $[0, \infty]$. The expectation then has an additional term $\infty \cdot \mathbf{P}[X = \infty]$. This is interpreted with the convention that $\infty \cdot 0 = 0$ while $\infty \cdot c = \infty$ for $c > 0$.

With this convention we have the identity $\mathbf{E}[aX] = a\mathbf{E}[X]$ for all real numbers $a \geq 0$ and random variables $X \geq 0$. Another useful and fundamental property of the expectation is that it preserves order: If $0 \leq X \leq Y$, then $0 \leq \mathbf{E}[X] \leq \mathbf{E}[Y]$.

It is shown in measure theory that for a sequence of positive random variables $X_i \geq 0$ we always have *countable additivity*

$$\mathbf{E}[\sum_i X_i] = \sum_i \mathbf{E}[X_i]. \tag{5.6}$$

The sum on the left is defined pointwise: For each outcome ω, the value of the random variable $\sum_i X_i$ on ω is the number $\sum_i X_i(\omega)$.

The notions of event and probability may be thought of as special cases of the notions of random variable and expectation. For each event A there is a corresponding random variable 1_A that has the value 1 on the outcomes in A and the value 0 on the outcomes not in A. The expectation of this *indicator* random variable is the probability:

$$\mathbf{E}[1_A] = \mathbf{P}[A]. \tag{5.7}$$

Exercise 5.9. Prove the most basic form of *Chebyshev's inequality*: If $Y \geq 0$, then for each $\epsilon > 0$ we have $\epsilon \mathbf{P}[Y \geq \epsilon] \leq \mathbf{E}[Y]$.

Exercise 5.10. Let $X \geq 0$ and let ϕ be an increasing function on the positive reals, for instance $\phi(x) = x^2$. Prove Chebyshev's inequality in the general form that says that for $\epsilon > 0$ we have $\phi(\epsilon)\mathbf{P}[X \geq \epsilon] \leq \mathbf{E}[\phi(X)]$.

Exercise 5.11. Let $0 \leq X \leq M$ for some constant M and let ϕ be an increasing function. Prove that $\mathbf{E}[\phi(X)] \leq \phi(\epsilon) + \phi(M)\mathbf{P}[X > \epsilon]$.

Exercise 5.12. Show that countable additivity for probabilities is a special case of countable additivity for positive random variables.

If we have a sequence of random variables, then we say $X_n \to X$ as $n \to \infty$ if for every ω we have $X_n(\omega) \to X(\omega)$ as $n \to \infty$.

Exercise 5.13. Prove that if X_n is an increasing sequence of positive random variables, and $X_n \to X$, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

Exercise 5.14. Consider this assertion: If X_n is a decreasing sequence of positive random variables, and $X_n \to X$, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$. Is it true in general? Are there special circumstances when it is true?

If we have a random variable X that is not positive, then we can write it as the difference of two positive random variables X^+ and X^-. We can define

$$\mathbf{E}[X] = \mathbf{E}[X^+] - \mathbf{E}[X^-] \tag{5.8}$$

provided that at least one of the expectations on the right is finite. Otherwise we have an ambiguous $\infty - \infty$ and the expectation is not defined. This is a not just a technical point; often much depends on whether an expectation is unambiguously defined.

The absolute value of a random variable is $|X| = X^+ + X^-$. The absolute value $|X|$ is also a random variable, and since it is positive, its expectation is defined. The expectation of X is defined and finite if and only if the expectation of $|X|$ is finite. In that case

$$|\mathbf{E}[X]| \leq \mathbf{E}[|X|]. \tag{5.9}$$

The expectation defined on the space of random variables with finite expectation is linear and order-preserving.

Example 5.2. Let $X_1, \ldots, X_n$ represent the results of the first n trials of the discrete independent trials experiment. Let $Y = f(X_1, \ldots, X_n)$. Then

$$\mathbf{E}[Y] = \sum_{x_1,\ldots,x_n} f(x_1, \ldots, x_n)p(x_1)\cdots p(x_n). \tag{5.10}$$

Example 5.3. Let $X_1, \ldots, X_n$ represent the results of the first n trials of the continuous independent trials experiment. Let $Y = f(X_1, \ldots, X_n)$. Then

$$\mathbf{E}[Y] = \int_{\mathbf{R}^n} f(x_1, \ldots, x_n)\rho(x_1)\cdots\rho(x_n)\,dx_1\cdots dx_n. \tag{5.11}$$

If X is a random variable and $\mathbf{P}^1$ is its distribution, then there is an expectation $\mathbf{E}^1$ associated with the distribution. It may be shown that $\mathbf{E}[f(X)] = \mathbf{E}^1[f]$.

Exercise 5.15. If X_1 and X_2 are the results of the first 2 trials of the continuous independent trials experiment, show that

$$\mathbf{E}[f(X_1 + X_2)] = \int_{-\infty}^{\infty} f(y)\rho_2(y)\,dy, \tag{5.12}$$

where ρ_2 is the convolution defined above.

Lecture 6. Variance

Summary: A random variable may be centered by subtracting its expectation. The variance of a random variable is the expectation of the square of its centered version. This is a simple but powerful concept; it gives rise to a form of the law of averages called the weak law of large numbers.

A random variable is said to have finite variance if $\mathbf{E}[X^2] < \infty$. In this case its *length* is defined to be $\|X\| = \sqrt{\mathbf{E}[X^2]}$.

Theorem 6.1 Schwarz inequality. If X and Y have finite variance, then their product XY has finite expectation, and

$$|\mathbf{E}[XY]| \leq \|X\|\|Y\|. \tag{6.1}$$

Proof: Let $a > 0$ and $b > 0$. Then for each outcome we have the inequality

$$\frac{\pm XY}{ab} \leq \frac{1}{2}\left(\frac{X^2}{a^2} + \frac{Y^2}{b^2}\right). \tag{6.2}$$

Since taking expectations preserves inequalities, we have

$$\frac{\pm\mathbf{E}[XY]}{ab} \leq \frac{1}{2}\left(\frac{\|X\|^2}{a^2} + \frac{\|Y\|^2}{b^2}\right). \tag{6.3}$$

Thus the left hand side is finite. Take $a = \|X\|$ and $b = \|Y\|$.

In the situation described by the theory we also write $\langle XY \rangle = \mathbf{E}[XY]$ and call it the *inner product* of X and Y.

If a random variable has finite variance then it also has finite expectation $\mathbf{E}[X] = \langle X \rangle$ obtained by taking the inner product with 1.

The *mean* of a random variable with finite expectation is defined to be

$$\mu_X = \mathbf{E}[X]. \tag{6.4}$$

The *variance* of a random variable with finite variance is defined to be

$$\mathrm{Var}(X) = \sigma_X^2 = \mathbf{E}[(X - \mu_X)^2]. \tag{6.5}$$

Thus σ_X^2 is the mean square deviation from the mean. The *standard deviation* σ_X is the square root of the variance.

The following identity is not intuitive but very useful.

$$\sigma_X^2 = \mathbf{E}[X^2] - \mathbf{E}[X]^2. \tag{6.6}$$

The reason it is not intuitive is that it writes the manifestly positive variance as the difference of two positive numbers.

Exercise 6.1. Prove it.

If X is a random variable with finite mean, then the *centered version* of X is defined to be $X - \mu_X$. It has mean zero. Obviously the variance of X is just the square of the length of the centered version of X.

If X is a random variable with finite non-zero variance, then its *standardized version* is defined to be $(X - \mu_X)/\sigma_X$. It has mean zero and variance one.

We define the *covariance* of X and Y to be

$$\mathrm{Cov}(X, Y) = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)]. \tag{6.7}$$

This is just the inner product of the centered versions of X and Y.

The *correlation* is the standardized form of the covariance:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{6.8}$$

This has the geometrical interpretation of the cosine of the angle between the centered random variables.

Again there is a non intuitive formula:

$$\mathrm{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]. \tag{6.9}$$

Exercise 6.2. Prove it.

Two random variables are said to be *uncorrelated* if their covariance (or correlation) is zero. This says that the centered random variables are orthogonal.

Theorem 6.2 Let $S_n = Y_1 + \cdots Y_n$ be the sum of uncorrelated random variables with finite variances $\sigma_1^2, \ldots, \sigma_n^2$. Then the variance of the sum is the sum $\sigma_1^2 + \cdots + \sigma_n^2$ of the variances.

Exercise 6.3. Prove this (theorem of Pythagoras)!

Example 6.1. Say the means are all zero and the variances are all the same number σ^2. Then S_n is a generalization of the random walk we considered before. The theorem says that $\mathbf{E}[S_n^2] = n\sigma^2$, or $\sqrt{\mathbf{E}[S_n^2]} = \sigma\sqrt{n}$. The random fluctuations of the walk are so irregular that it travels a typical distance of only $\sigma\sqrt{n}$ in time n.

Let $Y_1, \ldots, Y_n$ be uncorrelated random variables as above, all with the same mean μ. Define

$$\bar{Y}_n = \frac{Y_1 + \cdots + Y_n}{n} \tag{6.10}$$

to be the *sample mean*. Since this is a random variable, its value depends on the outcome of the experiment. For each n its expectation is the mean:

$$\mathbf{E}[\bar{Y}_n] = \mu. \tag{6.11}$$

Theorem 6.3 The variance of the sample mean is

$$\sigma_{\bar{Y}_n}^2 = \frac{\sigma_1^2 + \cdots + \sigma_n^2}{n^2}. \tag{6.12}$$

In many circumstances the right hand side goes to zero. This says that the sample mean has small deviation from the mean. This is a form of the "law of averages", known technically as the *weak law of large numbers*.

Assume that the variances of the Y_i are all the same σ^2. Then the variance of the sample mean is $n\sigma^2/n^2 = \sigma^2/n$. The standard deviation of the sample mean is $\sigma/\sqrt{n}$. It is exceedingly important to note that while $1/\sqrt{n}$ goes to zero as n tends to infinity, it does so relatively slowly. Thus one needs a quite large sample size n to get a small standard deviation of the sample mean. This $1/\sqrt{n}$ factor is thus both the blessing and the curse of statistics.

Example 6.2. Consider discrete waiting times $W_1, \ldots, W_n, \ldots$ with $\mathbf{P}[W_i = k] = pq^{k-1}$ for $k = 1, 2, 3, \ldots$ and with $\mathbf{P}[W_1 = k_1, \ldots, W_n = k_n] = \mathbf{P}[W_1 = k_1] \cdots \mathbf{P}[W_n = k_n]$. The mean of W_i is $\mu = 1/p$.

Exercise 6.4. Compute the mean from

$$\mu = \sum_{k=1}^{\infty} k\mathbf{P}[W_i = k]. \tag{6.13}$$

For the discrete waiting time the variance is $\sigma^2 = q/p^2$.

Exercise 6.5. Compute the variance from

$$\sigma^2 = \sum_{k=1}^{\infty} (k-\mu)^2 \mathbf{P}[W_i = k]. \tag{6.14}$$

Notice that if p is small, then the standard deviation of the waiting time W_i is almost as large as the mean. So in this sense waiting times are quite variable. However let $\bar{W}_n$ be the sample mean with sample size n. Then the mean is still μ, but the standard deviation is $\sigma/\sqrt{n}$. Thus for instance if $p = 1/2$, then the mean is 2 and the standard deviation of the sample mean is $\sqrt{2/n}$. A sample size of $n = 200$ should give a result that deviates from 2 by only something like 1/10. It is a good idea to perform such an experiment and get a value of $\bar{W}_n(\omega)$ for the particular experimental outcome ω. You will either convince yourself that probability works or astonish yourself that it does not work.

Example 6.3. Consider the Bernoulli process. Let $Y_i = 1$ if the ith trial is a success, $Y_i = 0$ if the ith trial is a failure. The mean of Y_i is p and its variance is $p - p^2 = pq$.

Exercise 6.6. Compute this variance.

If $i \neq j$, then the covariance of Y_i and Y_j is $p^2 - p^2 = 0$.

Exercise 6.7. Compute this covariance.

Let $N_n = Y_1 + \cdots + Y_n$ be the number of successes in the first n trials. Then the mean of N_n is np and the variance of N_n is npq.

Exercise 6.8. Compute the mean of N_n directly from the formula

$$\mathbf{E}[N_n] = \sum_k k\mathbf{P}[N_n = k] = \sum_k k \binom{n}{k} p^k q^{n-k}. \tag{6.15}$$

Let $F_n = N_n/n$ be the fraction of successes in the first n trials. This is the *sample frequency*. Then the mean of F_n is p and the variance of F_n is pq/n. We see from this that if n is large, then the sample frequency F_n is likely to be rather close to the probability p.

It is more realistic to express the result in terms of the standard deviation. The standard deviation of the sample frequency is $\sqrt{pq}/\sqrt{n}$.

This is the fundamental fact that makes statistics work. In statistics the probability p is unknown. It is estimated experimentally by looking at the value of the sample frequency F_n (for some large n) on the actual outcome ω. This often gives good results.

Exercise 6.9. Say that you are a statistician and do not know the value of p and $q = 1-p$. Show that the standard deviation of the sample frequency satisfies the bound

$$\frac{\sqrt{pq}}{\sqrt{n}} \leq \frac{1}{2\sqrt{n}}. \tag{6.16}$$

Results in statistics are often quoted in units of two standard deviations. We have seen that an upper bound for two standard deviations of the sample frequency is $1/\sqrt{n}$.

Exercise 6.10. How many people should one sample in a poll of public opinion? Justify your answer. (How large does n have to be so that $1/\sqrt{n}$ is 3 per cent?)

The Bernoulli example is unusual in that the variance $p(1-p)$ is a function of the mean p. For more general distributions this is not the case. Consider again a sequence of random variables Y_i all with the same mean μ and variance σ^2. Statisticians estimate the variance using the *sample variance*

$$V_n = \frac{(Y_1 - \bar{Y}_n)^2 + (Y_2 - \bar{Y}_n)^2 + \cdots + (Y_n - \bar{Y}_n)^2}{n-1}. \tag{6.17}$$

It has the property that it requires no knowledge of the mean μ and is unbiased:

$$\mathbf{E}[V_n] = \sigma^2 \tag{6.18}$$

Exercise 6.11. Prove this property. What is the intuitive reason for the $n-1$ in the denominator? Would a statistician who used n encounter disaster?

Exercise 6.12. The weak law of large numbers is also true if the covariances are not zero but merely small. Let $Y_1, \ldots, Y_n, \ldots$ be a sequence of random variables with mean μ such that for $j \leq i$ we have $\mathrm{Cov}(Y_i, Y_j) \leq r(i-j)$. Require that the bound $r(k) \to 0$ as $k \to \infty$. Show that the variance of the sample mean $\bar{Y}_n$ goes to zero as $n \to \infty$.

Exercise 6.13. In the case of bounded variances and zero covariances the variance of the sample mean goes to zero like $1/n$. Consider the more general weak law with a bound on the covariances. What kind of bound will guarantee that the variance of the sample mean continues to go to zero at this rate?

Exercise 6.14. In this more general weak law there is no requirement that the negative of the covariance satisfy such a bound. Are there examples where it does not?

Lecture 7. Independence

Summary: Some probability calculations only require that random variables be uncorrelated. Others require a stronger martingale property. The strongest property of this sort is independence.

Recall that the condition that two random variables X and Y be *uncorrelated* is the condition that

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]. \tag{7.1}$$

Let $Y_c = Y - \mathbf{E}[Y]$ be the centered random variable. The condition that the random variables are uncorrelated may also be written in the form

$$\mathbf{E}[XY_c] = 0. \tag{7.2}$$

There are stronger conditions that are very useful. One is the condition that for all functions f for which the relevant expectations are finite we have

$$\mathbf{E}[f(X)Y] = \mathbf{E}[f(X)]\mathbf{E}[Y]. \tag{7.3}$$

It says that Y is uncorrelated with every function of X. The condition is linear in Y but non-linear in X. In terms of the centered random variable $Y_c = Y - \mathbf{E}[Y]$ it says that

$$\mathbf{E}[f(X)Y_c] = 0. \tag{7.4}$$

If we think of a game with two stages, and we think of weighting the bet Y_c at the second stage with the result of a gambling strategy $f(X)$ based on the first stage result X, then this is the condition that the modified game remains fair, that is, that Y_c be a martingale difference. For the purposes of this discussion we shall call this the *martingale property.*

Exercise 7.1. Let Z have values 1 and 0 with equal probabilities, and let W have values 1 and -1 with equal probabilities, and let Z and W be independent. Show that $X = ZW$ and $Y = (1 - Z)$ are uncorrelated but do not have the martingale property.

There is an even stronger condition. Two random variables are *independent* if for all functions f and g for which the relevant expectations are finite,

$$\mathbf{E}[f(X)g(Y)] = \mathbf{E}[f(X)]\mathbf{E}[g(Y)]. \tag{7.5}$$

It says simply that arbitrary non-linear functions of X and Y are uncorrelated.

Exercise 7.2. Show that if X and Y are independent, and if I and J are intervals, then $\mathbf{P}[X \in I, Y \in J] = \mathbf{P}[X \in I]\mathbf{P}[Y \in J]$.

Exercise 7.3. Events A and B are said to be independent if their indicator functions 1_A and 1_B are independent. Find a single equation that characterizes independence of two events.

Exercise 7.4. Let Z have values 1 and 0 with equal probabilities, and let W have values 1 and -1 with equal probabilities, and let Z and W be independent. Show that $X = ZW$ and $Y = (1 - Z)W$ have the martingale property but are not independent.

We can generalize all these definitions to a sequence $Y_1, \ldots, Y_n$ of random variables. The condition that Y_n be uncorrelated with $Y_1, \ldots, Y_{n-1}$ is that

$$\mathbf{E}[(a_1Y_1 + \cdots + a_{n-1}Y_{n-1} + b)Y_n] = \mathbf{E}[a_1Y_1 + \cdots + a_{n-1}Y_{n-1} + b]\mathbf{E}[Y_n] \tag{7.6}$$

for all choices of coefficients.

The condition that $Y_n - \mathbf{E}[Y_n]$ is a martingale difference is that

$$\mathbf{E}[f(Y_1, \ldots, Y_{n-1})Y_n] = \mathbf{E}[f(Y_1, \ldots, Y_{n-1})]\mathbf{E}[Y_n] \tag{7.7}$$

for all functions f. This says that even if the bet $Y_n - \mathbf{E}[Y_n]$ is weighted by a gambling scheme based on the previous trials the expected gain remains zero.

The condition that Y_n be independent of $Y_1, \ldots, Y_{n-1}$ is that

$$\mathbf{E}[f(Y_1, \ldots, Y_{n-1})g(Y_n)] = \mathbf{E}[f(Y_1, \ldots, Y_{n-1})]\mathbf{E}[g(Y_n)] \tag{7.8}$$

for all functions f and g.

Exercise 7.5. Show that if we have a sequence $Y_1, \ldots, Y_n$ of random variables such that each Y_j is independent of the Y_i for $i < j$, then

$$\mathbf{E}[f_1(Y_1) \cdots f_n(Y_n)] = \mathbf{E}[f_1(Y_1)] \cdots \mathbf{E}[f_n(Y_n)] \tag{7.9}$$

for all functions $f_1, \ldots, f_n$.

Exercise 7.6. Say that we have three random variables Y_1, Y_2, and Y_3 such that Y_2 is independent of Y_1 and such that Y_3 is independent of Y_1 and Y_3 is independent of Y_2. Must Y_3 be independent of Y_1, Y_2?

How do we get random variables satisfying such conditions? The independence condition is the strongest of these conditions, so let us see if we can find independent random variables. Consider a probability space constructed from an sequence of independent trials. If Y_i depends on the ith trial, then Y_n is independent from $Y_1, \ldots, Y_{n-1}$.

It is easiest to see this in the discrete case. Then

$$\mathbf{E}[f(Y_1, \ldots, Y_{n-1})g(Y_n)] = \sum_{y_1,\ldots,y_n} f(y_1, \ldots, y_{n-1})g(y_n)\mathbf{P}[Y_1 = y_1, \ldots Y_n = y_n]. \tag{7.10}$$

However this is the same as

$$\begin{aligned} &\mathbf{E}[f(Y_1, \ldots, Y_{n-1})]\mathbf{E}[g(Y_n)] \\ &= \sum_{y_1,\ldots,y_{n-1}} \sum_{y_n} f(y_1, \ldots, y_{n-1})g(y_n)\mathbf{P}[Y_1 = Y_1, \ldots, Y_{n-1} = y_{n-1}]\mathbf{P}[Y_n = y_n]. \end{aligned} \tag{7.11}$$

The fact that the probability of the intersection event is the product of the probabilities is the defining property of the independent trials probability space.

In the continuous case one needs to use a limiting argument. We will leave this for another occasion.

It is certainly possible to have independent random variables that do not arise directly from the independent trials construction.

Example 7.1. Geometric waiting times. It is possible to construct a probability model in which discrete waiting times are directly constructed so that they will be independent. However it is also possible to construct instead the model for Bernoulli trials and see that the discrete waiting times arise as independent random variables, but not directly from the construction.

Let $Y_1, Y_2, \ldots, Y_n, \ldots$ be 1 or 0 depending whether the corresponding Bernoulli trial is a success or a failure. These random variables are independent, by their construction.

Let $T_0 = 0$ and let T_r for $r \geq 1$ be the trial on which the rth success takes place. Let $W_r = T_r - T_{r-1}$ be the *waiting time* until the rth success.

Exercise 7.7. Show that $\mathbf{P}[W_1 = k] = q^{k-1}p$. This is the probability distribution of a geometric *waiting time*. (Actually this is a geometric distribution shifted to the right by one, since the values start with 1 rather than with 0.)

Exercise 7.8. Show that $W_1, \ldots, W_n$ are independent. Hint: Show that the probability of the intersection event $\mathbf{P}[W_1 = k_1, \ldots, W_r = k_n] = \mathbf{P}[W_1 = k_1] \cdots [W_r = k_n]$.

Exercise 7.9. Find $\mathbf{P}[T_r = n]$. Hint: $T_r = n$ if and only if $N_{n-1} = r - 1$ and $Y_n = 1$.

Exercise 7.10. Find the mean and variance of T_r.

Exercise 7.11. Consider uncorrelated random variables $Y_1, \ldots, Y_n$ each with mean μ and variances σ^2. We have seen that the weak law of large numbers says that

$$\mathbf{E}[(\bar{Y}_n - \mu)^2] = \frac{\sigma^2}{n}. \tag{7.12}$$

Show that

$$\mathbf{P}[|\bar{Y}_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}. \tag{7.13}$$

Exercise 7.12. Consider independent random variables $Y_1, \ldots, Y_n, \ldots$ each with the same distribution, in particular each with mean μ and variance σ^2. Assume that the fourth moment $\mathbf{E}[(Y_1 - \mu)^4]$ is finite. Show that

$$\mathbf{E}[(\bar{Y}_n - \mu)^4] = \frac{1}{n^4}\left\{n\mathbf{E}[(Y_1 - \mu)^4] + \binom{4}{2}\binom{n}{2}\sigma^4\right\}. \tag{7.14}$$

Be explicit about where you use independence.

Exercise 7.13. Show that in the preceding exercise we have

$$\mathbf{E}[(\bar{Y}_n - \mu)^4] \leq \frac{C}{n^2}. \tag{7.15}$$

Exercise 7.14. Continuing, show that

$$\mathbf{E}[\sum_{n=k}^{\infty}(\bar{Y}_n - \mu)^4] \leq \frac{C}{k-1}. \tag{7.16}$$

Exercise 7.15. Show that

$$\mathbf{P}[\exists n\, n \geq k, |\bar{Y}_n - \mu| \geq \epsilon] \leq \frac{C}{(k-1)\epsilon^4}. \tag{7.17}$$

This remarkable result says that the probability that the sample mean ever deviates from the mean at any point in the entire future history is small. This is a form of the strong law of large numbers.

Exercise 7.16. Consider Bernoulli trials. Show that

$$\mathbf{P}[\exists n\, n \geq k, |F_n - p| \geq \epsilon] \leq \frac{1}{4(k-1)\epsilon^4} \tag{7.18}$$

for $k \geq 4$. This says that no matter what the value of p is, for n large enough the sample frequencies F_n are likely to get close to the probability p and stay there forever.

Lecture 8. Supermartingales

Summary: A martingale is a fair game. A supermartingale is a game that can be either fair or unfavorable. One can try to get a positive return with a supermartingale by a strategy of "quit when you are ahead" or of "buy low and sell high." This strategy can work, but only in a situation where there is also a possibility of a large loss at the end of play.

A martingale is a fair game. It is useful to have a more general concept; a *supermartingale* is a game that is unfavorable, in that on the average you are always either staying even or losing.

Let S_0 and $Y_1, Y_2, Y_3, \ldots$ be a sequence of random variables with finite expectations. Let $S_n = S_0 + Y_1 + \cdots + Y_n$. We think of the Y_i as the outcomes of the stages of a game, and S_n is the cumulated gain.

We want to specify when S_n is a supermartingale, that is, when the Y_i for $i \geq 1$ are supermartingale differences.

Let $W_1, W_2, W_3, \ldots$ be an arbitrary gambling scheme, that is, a sequence of bounded random variables such that $W_i \geq 0$ is a positive function of S_0 and of $Y_1, \ldots, Y_{i-1}$:

$$W_i = f_i(S_0, Y_1, \ldots, Y_{i-1}). \tag{8.1}$$

The requirement that the Y_i for $i \geq 1$ are supermartingale differences is that

$$\mathbf{E}[W_i Y_i] \leq 0 \tag{8.2}$$

for all $i \geq 1$ and for all such gambling schemes. We take this as a definition of supermartingale; it turns out to be equivalent to other more commonly encountered definitions.

Notice that we can take as a special case all $W_i = 1$. We conclude that for a supermartingale difference we always have

$$\mathbf{E}[Y_i] \leq 0 \tag{8.3}$$

for $i \geq 1$. By additivity we have

$$\mathbf{E}[S_n] \leq \mathbf{E}[S_0]. \tag{8.4}$$

On the average, the game is a losing game.

Example 8.1. Let S_0, $Y_0, Y_1, \ldots, Y_n, \ldots$ be a sequence of independent random variables with finite expectations and with $\mathbf{E}[Y_i] \leq 0$ for $i \geq 1$. Then the partial sums $S_n = S_0 + Y_1 + \ldots + Y_n$ form a supermartingale. In fact, if we have $W_i = f_i(S_0, Y_1, \ldots, Y_{i-1}) \geq 0$, then $\mathbf{E}[W_i Y_i] = \mathbf{E}[W_i]\mathbf{E}[Y_i] \leq 0$.

Theorem 8.1 Consider a supermartingale $S_n = S_0 + Y_1 + \cdots + Y_n$. Let $X_0 = f(S_0)$ be a function of S_0. Consider a gambling scheme $W_i = f_i(S_0, Y_1, \ldots, Y_{i-1}) \geq 0$. Form the sequence

$$X_n = X_0 + W_1 Y_1 + W_2 Y_2 + \cdots W_n Y_n. \tag{8.5}$$

Then X_n is also a supermartingale.

The proof of the theorem is immediate. One consequence is that

$$\mathbf{E}[X_n] \leq \mathbf{E}[X_0]. \tag{8.6}$$

All that a gambling scheme can do to a supermartingale is to convert it into another supermartingale, that is, into another losing game.

There is a related concept of *submartingale*, which is intended to model a winning game, in which all $\leq$ relations are replaced by $\geq$ relations. However this reduces to the concept of supermartingale by applying the theory to the negatives of the random variables.

The terminology "super" and "sub" may seem to be backward, since the supermartingale is losing and the submartingale is winning. However the terminology is standard. It may help to remember that "super" and "sub" refer to the initial value.

The concept of martingale is also subsumed, since a martingale is just a process which is both a supermartingale and a submartingale. In that case the inequalities are replaced by equalities.

Example 8.2. Let $X_n = S_0 + W_1Y_1 + W_2Y_2 + \cdots W_nY_n$ using some gambling scheme on the supermartingale generated by independent random variables of the previous example. Then this remains a supermartingale, even though it is no longer a sum of independent random variables.

The Bernoulli process with $p \leq 1/2$ is a supermartingale. In particular, the examples generated from the Bernoulli process with $p = 1/2$ are all martingales.

Now we are going to prove two fundamental theorems about unfavorable games using the principle: no large risk—little chance of gain. In the first theorem the gain comes from the strategy: quit when you win a fixed amount.

Theorem 8.2 Let S_n be a supermartingale with $\mathbf{E}[S_0] = a$. Let $b > a$ and let T be the first time $n \geq 1$ with $S_n \geq b$. Then

$$(b-a)\mathbf{P}[T \leq n] - \mathbf{E}[(S_n - a)^-] \leq 0. \tag{8.7}$$

Here $(S_n - a)^-$ is the negative part of $S_n - a$.

This theorem says that the possible winnings from gaining a fixed amount sometime during the play are balanced by the possible losses at the end of the game. Thus if there is little risk of loss, then the probability of a large gain is correspondingly small.

Proof: Let $X_n = S_n$ for $n \leq T$ and let $X_n = S_T$ for $n > T$. Thus S_n is produced by the gambling scheme W_i that is 1 for $i \leq T$ and 0 for $i > T$. The reason that this is a gambling scheme is that the event $i > T$ is determined by the sequence of values up to time $i - 1$.

Figure 6.

We have that

$$X_n - a \geq (b-a)1_{T \leq n} + (S_n - a)1_{T>n}. \tag{8.8}$$

When we take expectations we get

$$0 \geq (b-a)\mathbf{P}[T \leq n] + \mathbf{E}[(S_n - a)1_{T>n}] \geq (b-a)\mathbf{P}[T \leq n] - \mathbf{E}[(S_n - a)^-]. \tag{8.9}$$

Exercise 8.1. For the symmetric random walk martingale starting at zero, show that $\mathbf{P}[T_r \leq n] \leq \sqrt{n}/r$.

Exercise 8.2. Consider the martingale Z_n in which one doubles the bet each time until a win. Thus $Z_n = 1 - 2^n$ for $n < T_1$ and $Z_n = 1$ for $n \geq T_1$. Show that the inequality is an equality in this case.

Exercise 8.3. Here is another inequality that looks nearly the same but is based on quite another idea. Let $S_n = S_0 + Y_1 + \ldots + Y_n$ be a submartingale. Let a be arbitrary and $b > a$ and T be the first time n with $S_n \geq b$. Show that

$$\mathbf{E}[(S_n - a)^+] - (b-a)\mathbf{P}[T \leq n] \geq 0. \tag{8.10}$$

Hint: The strategy here is to donate a fixed amount of winnings to charity. The result is that for a favorable game the possible donations are more than compensated by the total returns at the end.

Exercise 8.4. Let S_n be the symmetric random walk starting at zero. It is a martingale. Show that S_n^2 is a submartingale.

Exercise 8.5. Let S_n be the symmetric random walk starting at zero. Show that $\mathbf{P}[\exists k\, 1 \leq k \leq n, |S_k| \geq r] \leq n/r^2$.

The second fundamental theorem involves a more complicated gambling strategy, where we try to gain each time the value of the game S_n rises through a specified range of values.

Let $a < b$ and let T_1 be the first time when S_n has a value below a, and let T_2 be the first time after that when S_n has a value above b. The i with $T_1 < i \leq T_2$ are the times of the first upcrossing. Consider the first time T_3 after all that when S_n has a value below a, and the first time T_4 after that when S_n has a value above b. The i with $T_3 < i \leq T_4$ are the times of the second upcrossing. The later upcrossings are defined similarly.

Figure 7.

We now let $S_n = S_0 + Y_1 + \ldots + Y_n$. Let $W_i = 1$ if i belongs to one of the upcrossing intervals, otherwise let $W_i = 0$. Then the W_i form a gambling scheme. Let $X_n = W_1 Y_1 + \cdots + W_n Y_n$ be the resulting accumulated winnings.

Theorem 8.3 (Upcrossing inequality) Let S_n be a supermartingale. Let $a < b$ and let U_n be the number of upcrossings of (a, b) in the first n trials. Then

$$(b-a)\mathbf{E}[U_n] - \mathbf{E}[(S_n - a)^-] \leq 0. \tag{8.11}$$

This theorem says that the expected winnings from a buy-low and sell-high strategy are balanced by the expected losses at the conclusion.

Proof: Consider the winnings X_n with the above gambling scheme. This continues to be a supermartingale, so $\mathbf{E}[X_n] \leq 0$. However

$$(b-a)U_n - (S_n - a)^- \leq X_n. \tag{8.12}$$

This is because the scheme gives winnings of $b - a$ for each completed upcrossing. However it may give a loss if n is part of an incompleted upcrossing, and this is the origin of the second term.

If we take expectations we obtain

$$(b-a)\mathbf{E}[U_n] - \mathbf{E}[(S_n - a)^-] \leq \mathbf{E}[X_n] \leq 0. \tag{8.13}$$

The expected winnings by this scheme must be balanced by the possibility of ending with a catastrophic loss.

Lecture 9. The supermartingale convergence theorem

Summary: A supermartingale that is bounded below is almost sure to converge to a limit. If it did not converge it would fluctuate forever. Then one could use a "buy low and sell high" strategy to make a gain with no compensating risk.

In the following discussion we shall need two basic convergence theorems. The first is the *monotone convergence theorem* for positive random variables. This says that if $0 \leq X_n \uparrow X$, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

This is proved by writing $X = X_0 + \sum_{k=1}^{\infty}(X_k - X_{k-1})$. Since $X_0 \geq 0$ and $X_k - X_{k-1} \geq 0$, we can apply countable additivity for positive random variables. This says that

$$\mathbf{E}[X] = \mathbf{E}[X_0] + \sum_{k=1}^{\infty} \mathbf{E}[X_{k+1} - X_k]. \tag{9.1}$$

This in turn is

$$\mathbf{E}[X] = \mathbf{E}[X_0] + \lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{E}[X_k - X_{k-1}] = \lim_{n\to\infty} \mathbf{E}[X_n]. \tag{9.2}$$

The second convergence theorem is *Fatou's lemma* for positive random variables. This says that if $X_n \geq 0$ for each n and $X_n \to X$ and $\mathbf{E}[X_n] \leq M$, then $\mathbf{E}[X] \leq M$. That is, when the convergence is not monotone, one can lose expectation in the limit, but never gain it.

This is proved by noting that for each n we have $0 \leq \inf_{k\geq n} X_k \leq X_n$. Thus $\mathbf{E}[\inf_{k\geq n} X_k] \leq \mathbf{E}[X_n] \leq M$. Furthermore, as $n \to \infty$ we have $\inf_{k \geq n} X_k \uparrow X$. Hence by the monotone convergence theorem $\mathbf{E}[X] \leq M$.

Theorem 9.1 Supermartingale convergence theorem. Consider a supermartingale S_n for $n = 0, 1, 2, 3, \ldots$ whose negative part has bounded expectation: there is a constant C such that for all n the expectation of the negative part $\mathbf{E}[S_n^-] \leq C$. Then there is a random variable S such that $S_n \to S$ as $n \to \infty$ almost surely.

In words: a losing game that keeps fluctuating must be a game with no floor on the average losses.

Note that the theorem has an equivalent statement in terms of submartingales: A submartingale whose positive part has bounded expectation must converge.

Proof: Consider $\limsup_n S_n = \lim_n \sup_{k\geq n} S_k$ and $\liminf_n S_n = \lim_n \inf_{k\geq n} S_k$. Clearly $\liminf_n S_n \leq \limsup_n S_n$. We show that they are equal with probability one.

Fix rational numbers $a < b$. Let U_n be the number of upcrossings of (a, b) up to time n. By the upcrossing inequality, the expectation $\mathbf{E}[U_n] \leq \mathbf{E}[(S_n - a)^-]/(b-a) \leq (C + a^+)/(b-a)$ which is bounded independently of n, by assumption. As $n \to \infty$ the random variables U_n increase to the random variable U that counts the number of upcrossings of the interval. By the monotone convergence theorem, $\mathbf{E}[U_n] \to \mathbf{E}[U]$ as $n \to \infty$. Hence

$$\mathbf{E}[U] \le (C + a^+)/(b-a). \tag{9.3}$$

There are certainly outcomes of the experiment for which there are infinitely many upcrossings of (a, b). However the set of all such outcomes must have probability zero; otherwise $\mathbf{P}[U = \infty] > 0$ and this would imply $\mathbf{E}[U] = \infty$, contrary to the estimate.

Since there are countably many pairs of rational numbers a, b, it follows from countable subadditivity that the probability that there exists an interval (a, b) with $a < b$ both rational and with infinitely many upcrossings is zero. This is enough to show that the probability that $\liminf_n S_n < \limsup_n S_n$ is zero. Let S be their common value. We have shown that $S_n \to S$ almost surely as $n \to \infty$.

We conclude by showing that S is finite almost everywhere. Since $\mathbf{E}[S_n^-] \le C$, it follows from Fatou's lemma that $\mathbf{E}[S^-] \le C$. Hence $S > -\infty$ almost surely. Furthermore, $\mathbf{E}[S_n^+] = \mathbf{E}[S_n] + \mathbf{E}[S_n^-] \le \mathbf{E}[S_0] + C$, since S_n is a supermartingale. Again by Fatou's lemma, $\mathbf{E}[S^+] < \infty$. Hence $S < \infty$ almost everywhere.

Exercise 9.1. Let S_n be the symmetric random walk starting at zero. Use the martingale convergence theorem to show that S_n is recurrent, that is, from every integer one reaches every other integer with probability one.

Exercise 9.2. Let S_n be the random walk with $p \le 1/2$. Use the supermartingale convergence theorem to show that from every integer one reaches every strictly smaller integer with probability one.

Exercise 9.3. Let S_n be the random walk with $p < 1/2$. Show that $(q/p)^{S_n}$ is a martingale. What is its limit as $n \to \infty$? What does this say about the limit of S_n as $n \to \infty$?

Exercise 9.4. Let $0 < a < 1$. Let $X_n = Y_1 \cdots Y_n$, where the Y_i are independent with values $1 + a$ and $1 - a$ with equal probability. Show that X_n converges with probability one. What is the limit?

Exercise 9.5. Let $X_n = \sum_{k=1}^{n} \frac{1}{k} Y_k$, where the $Y_k = \pm 1$ are the steps in a symmetric random walk. Show that $X_n \to X$ as $n \to \infty$ with probability one.

Lecture 10. Dominated convergence theorems

Summary: If a sequence of random variables converges, then it does not follow in general that the expectations converge. There are conditions that ensure that this happens. The most important ones involve the concept of domination by a fixed random variable with finite expectation.

Perhaps the fundamental convergence theorem is the monotone convergence theorem, which is just another formulation of countable additivity. Let X_n be a sequence of random variables. We say that $X_n \uparrow X$ if for each outcome ω the $X_n(\omega)$ are increasing to the limit $X(\omega)$ as $n \to \infty$.

The monotone convergence theorem says that if the X_n are positive random variables, then $X_n \uparrow X$ implies that $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

There are two very useful generalizations of the monotone convergence theorem. They are obvious corollaries of the usual monotone convergence theorem, but note carefully the domination hypothesis!

The *dominated below monotone convergence theorem* says that if the X_n are random variables with $\mathbf{E}[X_0] > -\infty$, then $X_n \uparrow X$ implies $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

The *dominated above monotone convergence theorem* says that if the X_n are random variables with $\mathbf{E}[X_0] < \infty$, then $X_n \downarrow X$ implies $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

Both the theorems would be false in general without the domination hypothesis. In general we will say that a sequence of random variables X_n is *dominated below* if there exists a random variable Y with $\mathbf{E}[Y] > -\infty$ and with $X_n \geq Y$ for all n. Similarly, we say that X_n is *dominated above* if there exists a Z with $\mathbf{E}[Z] < \infty$ and $X_n \leq Z$ for all n. It is important to notice that in probability theory it is often possible to take the dominating function to be a constant!

Example 10.1. Let X be a positive random variable with $\mathbf{E}[X] = \infty$. Let $X_n = X/n$. Then $X_n \downarrow 0$, but $\mathbf{E}[X_n] = \infty$ for each n.

We often encounter convergence that is not monotone. We say that $X_n \to X$ if for each outcome ω we have $X_n(\omega) \to X(\omega)$ as $n \to \infty$. There is a fundamental trick for reducing this more general kind of convergence to monotone convergence.

Suppose that $X_n \to X$. Note that

$$X_n \geq \inf_{k \geq n} X_k \tag{10.1}$$

and that the random variables on the right are increasing with n to X. Similarly note that

$$X_n \leq \sup_{k \geq n} X_k \tag{10.2}$$

and the random variables on the right are decreasing with n to X. It follows from the first inequality that if the X_k are dominated below, then

$$\mathbf{E}[X_n] \geq \mathbf{E}[\inf_{k \geq n} X_k] \to \mathbf{E}[X]. \tag{10.3}$$

This says that domination below implies that one can only lose expectation in the limit. Similarly, if the X_k are dominated above, then

$$\mathbf{E}[X_n] \leq \mathbf{E}[\sup_{k \geq n} X_k] \to \mathbf{E}[X]. \tag{10.4}$$

This says that domination above implies that one can only gain expectation in the limit. These results are both forms of *Fatou's lemma.*

Example 10.2. Let W be a discrete waiting time random variable with $\mathbf{P}[W = k] = (1/2)^k$ and consequently $\mathbf{P}[n < W] = (1/2)^n$. Let $X_n = 2^n$ if $n < W < \infty$ and $X_n = 0$ otherwise. Thus we get a large reward 2^n if we have to wait more than time n. On the other hand, we get no reward if we wait forever. This gives an example where $X_n(\omega) \to 0$ as $n \to \infty$ for each outcome ω, but $\mathbf{E}[X_n] = 1$ for all n.

Exercise 10.1. Show in the previous example that $\sup_{k \geq n} X_k = 2^{W-1}$ for $n < W < \infty$ and otherwise is zero. Thus we get a huge reward 2^{W-1} if we have to wait more than time n. However this becomes less and less advantageous as n increases. Again we get no reward if we wait forever. Find the expectation of this random variable. What does this say about domination from above?

In the martingale examples, we saw that a fair game could be favorable in the limit. This is only because the game was not dominated below, so it had a sort of infinite reserve to draw on, and so could gain expectation in the limit.

If we have domination both above and below, then the limit of $\mathbf{E}[X_n]$ is $\mathbf{E}[X]$. This is the famous and fundamental *dominated convergence theorem.*

We say that $X_n \to X$ *almost surely* if there is an event A with $\mathbf{P}[A] = 1$ such that for all ω in A we have $X_n(\omega) \to X(\omega)$ as $n \to \infty$. All the above theorems on convergence are true when convergence at every outcome is replaced by almost sure convergence. This is because the values of expectations are not influenced by what happens on the set A^c with probability zero.

We should record our results from martingale theory in the form of theorems. First recall the almost sure convergence theorems.

Theorem 10.1 A supermartingale whose negative part has bounded expectation must converge almost surely.

Theorem 10.2 A submartingale whose positive part has bounded expectation must converge almost surely.

Notice that the above results do not require domination. Domination is a stronger condition than a bound on expectation. The following are the basic variants of Fatou's lemma.

Theorem 10.3 If $\mathbf{E}[X_n] \leq M$ and $X_n \to X$ almost surely and the $X_n \geq -Y$ with $\mathbf{E}[Y] < \infty$, then $\mathbf{E}[X] \leq M$. In particular, a supermartingale that is dominated below can only lose in the limit.

Theorem 10.4 If $\mathbf{E}[X_n] \geq N$ and $X_n \to X$ almost surely and the $X_n \leq Y$ with $\mathbf{E}[Y] < \infty$, then $\mathbf{E}[X] \geq N$. In particular, a submartingale that is dominated above can only gain in the limit.

Theorem 10.5 If $X_n \to X$ almost surely and the $|X_n| \leq Y$ with $\mathbf{E}[Y] < \infty$, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$ as $n \to \infty$. In particular, a martingale that is dominated below and above remains fair in the limit.

Exercise 10.2. Let Y_k be a sequence of random variables such that $\sum_k \mathbf{E}[|Y_k|] < \infty$. Show that $\mathbf{E}[\sum_k Y_k] = \sum_k \mathbf{E}[Y_k]$. Hint: Find a dominating function for $\sum_{k=1}^n Y_k$.

Exercise 10.3. Let A_n be a sequence of events such that $A_n \to A$ as n tends to infinity. This means that for each outcome ω, we have $\omega \in A$ if and only if $\omega \in A_n$ for all sufficiently large n. Show that $\mathbf{P}[A_n] \to \mathbf{P}[A]$ as n tends to infinity.

Sometimes it is impossible to find a dominating function. Another device to control the expectation of the limit is to get control on the variance or second moment. In this situation one can show that one has approximate dominance by a large constant.

Theorem 10.6 If $\mathbf{E}[X_n^2] \leq C^2$ and $X_n \to X$ almost surely, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$ as $n \to \infty$. In particular, a martingale with a bound on its second moment remains fair in the limit.

Proof: It is easy to see that $\mathbf{E}[(X_n - X_m)^2] \leq 4C^2$. Let $m \to \infty$. By Fatou's lemma $\mathbf{E}[(X_n - X)^2] \leq 4C^2$. Now consider a large constant K. Then

$$|\mathbf{E}[X_n] - \mathbf{E}[X]| \leq \mathbf{E}[|X_n - X|] = \mathbf{E}[|X_n - X|1_{|X_n - X| \leq K}] + \mathbf{E}[|X_n - X]1_{|X_n - X| > K}] \quad (10.5)$$

The second term on the right is bounded by

$$\mathbf{E}[|X_n - X]1_{|X_n - X| > K}] \leq \frac{1}{K}\mathbf{E}[(X_n - X)^2] \leq \frac{4C^2}{K}. \quad (10.6)$$

Take K so large that this is very small. The first term then goes to zero by the dominated convergence theorem (with dominating function K).

In the following examples we want to use martingales to do calculations. The technique is summarized in the following two theorems.

Theorem 10.7 Let S_n be a martingale. Let T be the first time that the martingale reaches some set. Write $T \wedge n$ for the minimum of T and n. Then $S_{T\wedge n}$ is a martingale. In particular $\mathbf{E}[S_{T\wedge n}] = \mathbf{E}[S_0]$.

Proof: The martingale $S_{T\wedge n}$ is equal to S_n for $n < T$ and to S_T for $n \geq T$. If $S_n = S_0 + Y_1 + \cdots + Y_n$, then $S_{T\wedge n} = S_0 + W_1 Y_1 + \cdots + W_n Y_n$, where W_i is 1 or 0 depending on whether $i \leq T$ or $i > T$. Since one can ascertain whether $i > T$ by the history up to time $i - 1$, this is a gambling scheme. The strategy is to play as long as you have not arrived at the set.

Theorem 10.8 Let S_n be a martingale. Let T be the first time that the martingale reaches some set. Assume that $T < \infty$ with probability one. Assume that $S_{T\wedge n}$ is dominated above and below by random variables that do not depend on n and that have finite expectation. Then $\mathbf{E}[S_T] = \mathbf{E}[S_0]$.

Proof: Since $S_{T\wedge n} \to S_T$ as $n \to \infty$, this follows from the previous result and the dominated convergence theorem.

Example 10.3. Symmetric random walk martingale stopped at $r > 0$. Recall the symmetric random walk $S_n = Y_1 + \cdots + Y_n$, where the Y_i are independent random variables with $Y_i = \pm 1$ and have mean zero. Then S_n is a martingale. Let T be the first time that the process hits $r > 0$. Write $T \wedge n$ for the minimum of T and n. If we look at the process $S_{T\wedge n}$ where we stop S_n at the point $r > 0$, then $S_{T\wedge n}$ is also a martingale. Since $S_{T\wedge n}$ is a submartingale that is bounded above, it must converge almost surely, and in fact $S_{T\wedge n} \to r$ as $n \to \infty$. This is compatible with Fatou's lemma: $S_{T\wedge n}$ is dominated above by r, and the limiting expectation is $r \geq 0$. Since it is not dominated below, the fair game can be favorable in the limit. With infinite reserves, opportunism is a winning strategy.

We can also consider the same process $S_{T\wedge n}$ as a supermartingale. Note that $S_{T\wedge n} - r \leq 0$, so $\mathbf{E}[(S_{T\wedge n} - r)^-] = \mathbf{E}[r - S_{T\wedge r}] = r$. This shows that the negative part has bounded expectation, so it must converge almost surely (as we already know). However it is not dominated below, and in fact it gains in the limit! This important counterexample show that there is really a need for an extra hypothesis to prevent a supermartingale from gaining in the limit.

Example 10.4. Symmetric random walk martingale stopped at $-s < 0$ or $r > 0$. Instead let T be the first time that S_n hits either $-s < 0$ or $r > 0$. The stopped process $S_{T\wedge n}$ is a bounded martingale and hence must converge almost surely. It converges to S_T as $n \to \infty$. By the dominated convergence theorem $\mathbf{E}[S_T] = 0$, and from this one can work out the probability of the event of hitting either point.

Exercise 10.4. Work out these probabilities.

Also $S_n^2 - n$ is a martingale. Let T be the hitting time of S_n for $-s$ or r. The process $S_{T\wedge n}^2 - T \wedge n$ is a martingale. There is a slight problem in that it is not bounded below. However $\mathbf{E}[T \wedge n] = \mathbf{E}[S_{T\wedge n}^2] \leq \max(r^2, s^2)$. It follows from the monotone convergence theorem that $\mathbf{E}[T] \leq \max(r^2, s^2)$. Hence T is a dominating function for $T \wedge n$. It follows from the dominated convergence theorem that $\mathbf{E}[T] = \mathbf{E}[S_T^2]$. From this one can calculate $\mathbf{E}[T]$.

Exercise 10.5. Do this calculation.

Example 10.5. Random walk supermartingale stopped at $r > 0$. Now consider random walk $S_n = Y_1 + \cdots + Y_n$, but now the probability that $Y_i = 1$ is p and the probability that $Y_i = -1$ is $q = 1 - p$. This is not a martingale when $p \neq 1/2$. We must look for other martingales. (This is done systematically in the theory of Markov chains, but here we only present the results for this example.)

The first martingale that we may associate with this problem is $(q/p)^{S_n}$.

Exercise 10.6. Check that this martingale is obtained from a sum of independent random variables by a gambling scheme.

Let T be the hitting time of S_n for a point $r > 0$. Consider the process $S_{T\wedge n}$ that is equal to S_n for $n \leq T$ and to $S_T = r$ for $n > T$. This is the non-symmetric random walk stopped at r. Another useful martingale is $(q/p)^{S_{T\wedge n}}$.

Exercise 10.7. Show that this is obtained from the previous martingale by a gambling scheme.

Assume that $p < 1/2$. Then $(q/p)^{S_{T\wedge n}}$ is a bounded martingale. Hence it converges almost surely, and the limit is $(q/p)^r$ with probability $\mathbf{P}[T < \infty]$. Furthermore, the game is fair in the limit. Hence $1 = (q/p)^r\mathbf{P}[T < \infty]$. Thus we have computed the hitting probability of $r > 0$ to be $(p/q)^r$.

Figure 8.

Exercise 10.8. Check the limit of this martingale.

Example 10.6. Random walk supermartingale stopped at $-s < 0$. Now continue to take $p < 1/2$, but for variety let T be the time that S_n hits $-s < 0$. Let $S_{T\wedge n}$ be the process that is equal to S_n for $n \leq T$ and to $S_T = -s$ for $n > T$. This is the process stopped at $-s$. This is a supermartingale that is bounded below, so it must converge almost surely, and the limit is clearly $-s$.

We want to manufacture a martingale out of this. It is clear that $X_n = S_n - (p-q)n$ is a martingale. This martingale stopped at T is $X_{T\wedge n} = S_{T\wedge n} - (p-q)(T\wedge n)$. It is clear that $X_{T\wedge n}$ converges almost surely to $-s-(p-q)T$ as $n \to \infty$.

We would like to argue that the martingale remains fair in the limit, but $S_{T\wedge n}$ and $T \wedge n$ are not bounded above uniformly in n. (They are both bounded below.) However we can try to dominate these random variables above by random variables independent of n.

Since $X_{T\wedge n}$ is a martingale, we have $(q-p)\mathbf{E}[T\wedge n] = -\mathbf{E}[S_{T\wedge n}] \leq s$. Since $T \wedge n \uparrow T$ as $n \to \infty$, the monotone convergence theorem implies that $(q-p)\mathbf{E}[T] \leq s$. This shows that T is a dominating function for $T \wedge n$.

Clearly $S_{T\wedge n} \leq \sup_n S_n$. However on the set of outcomes where $\sup_n S_n = r$ the hitting time of r is finite, and we have seen that the probability of this event is $(p/q)^r$. Hence

$$\mathbf{E}[\sup_n S_n] \leq \sum_r r\left(\frac{p}{q}\right)^r < \infty. \tag{10.7}$$

We have found a dominating function $\sup_n S_n$ for $S_{T\wedge n}$.

By the dominated convergence theorem the martingale is fair in the limit. Hence

$$0 = \mathbf{E}[S_T] - (p-q)\mathbf{E}[T] = -s + (q-p)\mathbf{E}[T]. \tag{10.8}$$

In conclusion,

$$\mathbf{E}[T] = \frac{s}{q-p} \tag{10.9}$$

In a strictly unfavorable game you lose fast.

Figure 9.

Lecture 11. The strong law of large numbers

Summary: The strong law of large numbers is a form of the law of averages that says that sample means approach a limiting value and stay close to it throughout all future history. This can be proved by applying the martingale convergence theorem to the sum of the martingale parts of the differences of successive sample means.

Let $Y_1, \ldots, Y_n, \ldots$ be a sequence of martingale differences with mean zero and variances $\sigma_1^2, \ldots, \sigma_n^2, \ldots$. Let

$$\bar{Y}_n = \frac{Y_1 + \cdots + Y_n}{n} \tag{11.1}$$

be the sample mean. The change in the sample mean as the sample size is increased by one has a simple and useful decomposition. This is

$$\bar{Y}_n - \bar{Y}_{n-1} = \frac{1}{n} Y_n - \frac{1}{n} \bar{Y}_{n-1}. \tag{11.2}$$

It says that the change in the sample mean consists of a new contribution and a contribution from the previous sample mean. The new contribution is a martingale difference. The other term is predictable from the previous result.

This formula is useful as a stable way to update the mean. We can write it as

$$\bar{Y}_n = \frac{1}{n} Y_n + \frac{n-1}{n} \bar{Y}_{n-1}. \tag{11.3}$$

Let

$$Z_n = Y_1 + \frac{1}{2} Y_2 + \cdots + \frac{1}{n} Y_n \tag{11.4}$$

with $Z_0 = 0$. The update equation can be written as

$$n\bar{Y}_n = n(Z_n - Z_{n-1}) + (n-1)\bar{Y}_{n-1}. \tag{11.5}$$

This can be solved by iteration to give

$$n\bar{Y}_n = nZ_n - (Z_{n-1} + \cdots + Z_0), \tag{11.6}$$

or

$$\bar{Y}_n = Z_n - \frac{1}{n}(Z_0 + \cdots Z_{n-1}). \tag{11.7}$$

This expresses the sample mean as the difference between the present value of the martingale and the average of the past values of the martingale.

Theorem 11.1 (Kolmogorov's strong law) Let $Y_1, \ldots, Y_n, \ldots$ be a sequence of mean zero martingale difference random variables with finite variances $\sigma_1^2, \ldots, \sigma_n^2, \ldots$. Assume that the variances are uniformly bounded, or more generally that the variances satisfy

$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} = M^2 < \infty. \tag{11.8}$$

Then with probability one the sample means $\bar{Y}_n$ converge to zero as $n \to \infty$.

Proof: Let Z_n be the sum as above. It is obtained from a martingale by a gambling scheme, therefore it is a martingale. Its variance $\mathbf{E}[Z_n^2] \leq M^2$. As a consequence $\mathbf{E}[|Z_n|] \leq M$. Therefore it converges to a limit Z. For each outcome for which the martingale $Z_n \to Z$, it follows that the sample mean $\bar{Y}_n \to Z - Z = 0$.

Of course if the random variables of interest do not have mean zero, then this theorem is applied to their centered versions.

One can ask in what way this strong law improves on the weak law of large numbers. The weak law implies that if n is large, then the sample mean $\bar{Y}_n$ is likely to be close to the mean. The strong law says that if n is large, then all the sample means $\bar{Y}_k$ for $k \geq n$ are likely to be close to the mean. This is a much stronger assertion.

Here is a precise way of making the comparison. Take $\epsilon > 0$. The weak law implies that

$$\epsilon^2 \mathbf{P}[|\bar{Y}_n| \geq \epsilon] \leq \mathbf{E}[\bar{Y}_n^2] \to 0 \tag{11.9}$$

as $n \to \infty$. In other words, for each $\epsilon > 0$ $\mathbf{P}[|\bar{Y}_n| < \epsilon] \to 1$ as $n \to \infty$.

The strong law implies that

$$\mathbf{P}[\exists n \forall k \geq n \, |\bar{Y}_k| < \epsilon] = 1. \tag{11.10}$$

Since the events $\forall k \geq n \, |\bar{Y}_k| < \epsilon$ are increasing with n to an event with probability one, by the monotone convergence theorem their probability approaches one. Thus $\mathbf{P}[\forall k \geq n \, |\bar{Y}_k| < \epsilon] \to 1$ as $n \to \infty$.

Exercise 11.1. Show that the condition of almost sure convergence

$$\mathbf{P}[\forall \epsilon > 0 \exists n \forall k \geq n \, |X_k - X| < \epsilon] = 1 \tag{11.11}$$

is equivalent to the condition $\forall \epsilon > 0\, \mathbf{P}[\exists n \forall k \geq n\, |X_k - X| < \epsilon] = 1$ and that this is in turn equivalent to the convergence in probability of the entire future history:

$$\forall \epsilon > 0 \lim_{n\to\infty} \mathbf{P}[\forall k \geq n\, |X_k - X| < \epsilon] = 1. \tag{11.12}$$

Show that any of these conditions implies the usual condition of convergence in probability for the individual random variables: $\forall \epsilon > 0 \lim_{n\to\infty} \mathbf{P}[|X_n - X| < \epsilon] = 1$.

Exercise 11.2. There is a slight technical difference between the hypotheses in the two theorems. Show that the hypothesis on the variances in the strong law of large numbers implies the hypothesis on the variances in the weak law of large numbers.

Example 11.1. Consider a sum of independent and identically distributed random variables with mean zero and finite variance. The path $S_n = Y_1 + \ldots + Y_n$ is a generalization of the random walk that we have been considering. Let $\epsilon > 0$ and consider the cone $|s| \leq \epsilon n$. The weak law says that for n sufficiently large the point (n, S_n) is very likely to satisfy $|S_n| \leq \epsilon n$ and so be inside this cone. The strong law says that with probability one there is an n such that for all $k \geq n$ the points (k, S_k) all lie in this cone. It is a prediction about the entire future history of the process.

Let $p + q = 1$ with $p \neq q$. Consider the number of successes N_n in n trials. Write

$$X_n = 2^n p^{N_n} q^{n-N_n}. \tag{11.13}$$

Then X_n is the ratio of the probability of a path computed with parameter p to the probability of the same path computed with parameter $1/2$.

Exercise 11.3. We have seen that when the probability model is independent trials with probability of success on one trial equal to $1/2$, then the ratios X_n form a martingale, and X_n converges to zero almost surely as $n \to \infty$. Give an independent proof of this by applying the strong law of large numbers to the sample frequencies N_n/n.

Exercise 11.4. Show that if the probability model is independent trials with probability of success on one trial equal to p, then X_n is a submartingale and X_n converges to infinity almost surely as $n \to \infty$.

Exercise 11.5. Kolmogorov's strong law of large numbers implies that if $Y_1, \ldots, Y_n, \ldots$ is a sequence of martingale differences with $\mathbf{E}[Y_n^2] \leq C^2$, then $\bar{Y}_n \to 0$ as $n \to \infty$ with probability one. This version has a particularly simple proof. Write the difference

$$\bar{Y}_n - \bar{Y}_{n-1} = \frac{Y_n}{n} - \frac{1}{n}\bar{Y}_{n-1} \tag{11.14}$$

as the sum of a martingale difference and a predictable part. Show that

$$\mathbf{E}[|\bar{Y}_{n-1}|] \leq \frac{C}{\sqrt{n-1}}. \tag{11.15}$$

Show that the sum of the predictable parts converges absolutely with probability one.

Exercise 11.6. The conclusion of Kolmogorov's strong law of large numbers makes no mention of variances, so it is tempting to imagine a theorem of the following form: If $Y_1, \ldots, Y_n, \ldots$ is a sequence of martingale differences with $\mathbf{E}[|Y_n|] \leq C$, then $\bar{Y}_n \to 0$ with probability one. This is false. Find a counterexample.

Exercise 11.7. The counterexample must not satisfy the hypotheses of Kolmogorov's theorem. Check this.

Exercise 11.8. Let $S_n = Y_1 + \ldots + Y_n$. Prove the update formula

$$\frac{S_n}{a_n} = \frac{Y_n}{a_n} + \frac{a_{n-1}}{a_n}\frac{S_{n-1}}{a_{n-1}}. \tag{11.16}$$

Exercise 11.9. In the previous exercise let $Z_0 = 0$ and $Z_n = \sum_{j=1}^{n}(Y_j/a_j)$. Show that

$$\frac{S_n}{a_n} = Z_n - \sum_{k=1}^{n} \frac{a_k - a_{k-1}}{a_n} Z_{k-1} \tag{11.17}$$

where $a_0 = 0$.

Exercise 11.10. In the previous exercise, show that if the a_n are increasing to infinity and $Z_n \to Z$, then $S_n/a_n \to 0$.

Exercise 11.11. Consider martingale differences Y_i with mean zero and variance σ^2. Show that $\mathbf{E}[S_n^2] = n\sigma^2$ and $\mathbf{E}[|S_n|] \leq n^{1/2}\sigma$. Thus the sum S_n of n random terms grows on the average like $n^{1/2}$.

Exercise 11.12. Take $a_n = n^{1/2}(\log n)^{1/2+\epsilon}$ with $\epsilon > 0$. Consider martingale differences Y_i with mean zero and variance σ^2. Show that $S_n/a_n \to 0$ with probability one. Thus even the largest fluctuations of a sum S_n of n random terms grow at most only slightly faster than $n^{1/2}$.

Lecture 12. Convergence of distributions

Summary: The distribution of a random variable describes the various probabilities that can be calculated using the values of the random variable. Given a sequence of random variables, it is possible that the corresponding sequence of distributions converges in some appropriate sense. In order to prove that the distributions converge, it is sufficient to prove the convergence of expectations involving certain smooth functions of the random variables.

We say that the distribution of X_n *converges weakly* to the distribution of X as $n \to \infty$ if for each bounded continuous function f we have $\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)]$ as $n \to \infty$.

Exercise 12.1. Say that the distribution of X_n converges weakly to the distribution of X. Let g be a continuous function. Show that the distribution of $g(X_n)$ converges weakly to the distribution of $g(X)$.

Exercise 12.2. For each $p = 1/n$ let W_n be a discrete waiting time random variable, so that $\mathbf{P}[W_n = k] = (1-p)^{k-1}p$. Fix $\lambda > 0$. Show that the distribution of $W_n/(n\lambda)$ converges weakly to the distribution of a continuous exponential waiting time with parameter λ.

Example 12.1. Say that $X_n = a_n$ and $X = a$ are constant random variables. The distribution of each of these random variables is a probability measures that assigns probability one to a single point. Suppose $a_n \to a$ as $n \to \infty$. Then the distribution of X_n converges weakly to the distribution of a. This just says that $f(a_n) \to f(a)$. Note that the continuity of f is essential.

Example 12.2. It would be a mistake to think that if the distribution of X_n converges weakly to the distribution of X, then X_n must converge to X. Take for instance a situation where Y has the values 1 and -1 each with probability 1/2. Let $X_n = Y$ and $X = -Y$. The $\mathbf{E}[f(X_n)] = \mathbf{E}[f(X)] = f(1)(1/2) + f(-1)(1/2)$. However the random variable X_n is very far from the random variable X.

Exercise 12.3. Show that it is possible that the distribution of X_n approaches the distribution of X, the distribution of Y_n approaches the distribution of Y, but the distribution of $X_n + Y_n$ does not approach the distribution of $X + Y$.

Exercise 12.4. Show that the distribution of Y_n approaches the distribution of a if and only if for every $\epsilon > 0$ we have $\mathbf{P}[|Y_n - a| < \epsilon] \to 1$ as $n \to \infty$. Hint: Approximate the indicator function of the interval $(a - \epsilon, a + \epsilon)$ above and below by continuous functions that have the value 1 at a.

Exercise 12.5. Show that in the weak law of large numbers the distribution of the sample means $\bar{Y}_n$ approaches the distribution of the constant μ.

Exercise 12.6. Show that if the distribution of X_n approaches the distribution of X, and the distribution of Y_n approaches the distribution of a, then the distribution of $X_n + Y_n$ approaches the distribution of $X + a$.

Exercise 12.7. Show that if the distribution of X_n approaches the distribution of X, and the distribution of Y_n approaches the distribution of a, then the distribution of $Y_n X_n$ approaches the distribution of aX.

Exercise 12.8. Let g be a smooth function, say bounded with bounded derivatives. Show that if the distribution of X_n approaches the distribution of a and the distribution of $(X_n - a)/b_n$ approaches the distribution of Z, then the distribution of $(g(X_n) - g(a))/b_n$ approaches the distribution of $g'(a)Z$.

Ideally we would like to have $\mathbf{P}[X_n \leq a]$ converge to $\mathbf{P}[X \leq a]$ as $n \to \infty$ for each a. However this is too much to expect in general.

Example 12.3. Let X_n be the constant random variable whose value is always $1/n$. Let X be the constant random variable whose value is always 0. The distribution of X_n, which is a probability measure concentrated on the point $1/n$, converges weakly to the distribution of X, a probability measure concentrated on the point 0. However $\mathbf{P}[X_n \leq 0] = 0$ for all n, while $\mathbf{P}[X \leq 0] = 1$.

Theorem 12.1 If the distribution of X_n converges weakly to the distribution of X, and if $\mathbf{P}[X = a] = 0$, then $\mathbf{P}[X_n \le a] \to \mathbf{P}[X \le a]$ as $n \to \infty$.

Proof: Let f be the indicator function of $(-\infty, a]$. Even though this is not a continuous function, we want to show that $\mathbf{E}[f(X_n)]$ converges to $\mathbf{E}[f(X)]$.

Let $\epsilon > 0$. Since $\mathbf{P}[X = a] = 0$, it follows that $\mathbf{P}[a - 1/k \le X \le a + 1/k] < \epsilon/2$ for k sufficiently large.

Let g be the continuous function that is 1 below $a - 1/k$, 0 above a and linear in between. Let h be the continuous function that is 1 below a, 0 above $a + 1/k$, and linear in between. Then $g \le f \le h$ and $f - g \le 1$ and $h - f \le 1$. Thus $\mathbf{E}[g(X)] \le \mathbf{E}[f(X)] \le \mathbf{E}[h(X)]$, and the two extreme numbers are both within $\epsilon/2$ of the number in the middle.

Now $\mathbf{E}[g(X_n)] \le \mathbf{E}[f(X_n)] \le \mathbf{E}[h(X_n)]$, and for n sufficiently large $\mathbf{E}[g(X_n)]$ and $\mathbf{E}[h(X_n)]$ are each within $\epsilon/2$ of $\mathbf{E}[g(X)]$ and $\mathbf{E}[h(X)]$. It follows that $\mathbf{E}[f(X_n)]$ is within ϵ of $\mathbf{E}[f(X)]$. This completes the proof.

The rest of the material in this section is technical. It is useful to be able to prove weak convergence by checking convergence only for smooth functions. Let C_c^∞ be the space of all smooth functions that each vanish outside of some interval (depending on the function). Let X_n be a sequence of random variables and let X be another random variable. The distribution of X_n *converges in the sense of Schwartz* to the distribution of X if for all f in C_c^∞

$$\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)]. \tag{12.1}$$

The space of continuous functions that vanish at infinity is often a natural space in measure theory, so we also introduce it here. Let C_0 be the space of all continuous functions that vanish at infinity. The distribution of X_n *converges in the sense of Radon* to the distribution of X if for all f in C_0

$$\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)]. \tag{12.2}$$

Recall also the definition of weak convergence. Let BC be the space of all bounded continuous functions. The distribution of X_n *converges weakly* to the distribution of X if for all f in BC

$$\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)]. \tag{12.3}$$

Theorem 12.2 If the distribution of X_n converges to the distribution of X in the sense of Schwartz, then it converges in the sense of Radon.

Proof: This follows form the fact that every f in C_0 may be approximated uniformly with arbitrary precision by a function g in C_c^∞. (This fact is proved in courses of analysis.)

Theorem 12.3 If the distribution of X_n converges to the distribution of X in the sense of Radon, then it converges weakly.

Proof: It is not true that every function f in the space BC of bounded continuous functions can be approximated uniformly by a function in C_0 that also vanishes at infinity. So we must use an approximation argument. Let f be bounded above and below by some constant C.

Let $\epsilon > 0$. Let h be a function that is 1 on some large interval $[-M, M]$, 0 on $(-\infty, -M-1]$ and on $[M+1, \infty)$ and linear in between. By the monotone convergence theorem we can choose M so large that $\mathbf{E}[h(X)] \geq 1 - \epsilon/(6C)$. Then since h is continuous and vanishes at infinity, we can take n so large that $\mathbf{E}[h(X_n)] \geq 1 - \epsilon/(3C)$.

Now consider the function hf obtained by multiplying the values of the two functions; it is continuous and vanishes at infinity. The function $(1-h)f$ need not vanish at infinity, but we have chosen h so that $\mathbf{E}[(1-h)(X_n)]$ and $\mathbf{E}[(1-h)(X)]$ are both bounded above by $\epsilon/(3C)$. Furthermore, we have

$$\mathbf{E}[f(X_n)]-\mathbf{E}[f(X)] = \mathbf{E}[(1-h)(X_n)f(X_n)]+\mathbf{E}[hf(X_n)]-\mathbf{E}[hf(X)]+\mathbf{E}[(1-h)(X_n)f(X)]. \tag{12.4}$$

For n large enough we can make the absolute value of the difference of the two middle terms on the right hand side smaller than $\epsilon/3$. It follows that the absolute value of the difference on the left hand side is smaller than ϵ.

Lecture 13. The central limit theorem

Summary: The central limit theorem says that under rather general circumstances the distribution of a standardized sum of independent random variables approaches the standard normal distribution. This gives a detailed description of the deviations from average behavior.

The central limit theorem says that the normal distribution (or Gaussian distribution) is in some sense universal. Before beginning, we should ask the question: why the normal distribution? The answer is that this is the distribution with finite variance that reproduces itself under addition of independent random variables and appropriate normalization. Recall that the standard normal distribution has density $(1/\sqrt{2\pi})\exp(-z^2/2)$.

Theorem 13.1 Let $Z_1, \ldots, Z_n$ be independent standard normal random variables with mean zero. Let $\sigma_1^2 + \cdots + \sigma_n^2 = 1$. Then $W = \sigma_1 Z_1 \cdots + \sigma_n Z_n$ is also standard normal.

Proof: We give the proof for the case $n = 2$. The proof of the general case can be done using the same idea or derived from the case $n = 2$. This is left as an exercise.

The idea is to look at the joint density of the standard normal random variables Z_1 and Z_2. Since the random variables are independent, this is just the product

$$\rho(z_1)\rho(z_2) = \frac{1}{2\pi}\exp(-\frac{1}{2}z_1^2)\exp(-\frac{1}{2}z_2^2) = \frac{1}{2\pi}\exp(-\frac{1}{2}(z_1^2+z_2^2)). \tag{13.1}$$

This uses the special fact that the product of exponentials is the exponential of a sum. The probability that the point (Z_1, Z_2) is in some subset of the plane is given by the integral of the joint density over this subset. This is obvious for rectangles, and it extends to more general sets by an approximation argument.

The other special fact that we use is that the quadratic form $z_1^2 + z_2^2$ is invariant under rotation. Let $w_1 = \sigma_1 z_1 + \sigma_2 z_2$ and $w_2 = -\sigma_2 z_1 + \sigma_1 z_2$. The w_i arise from the z_j by a rotation in the plane. Define corresponding random variables W_i in terms of the Z_j by the same rotation formula.

We compute the joint density of W_1 and W_2. To find the probability that (W_1, W_2) are in some set in the plane, we integrate the joint density of (W_1, W_2) over this set. This joint density is given by taking the joint density of Z_1 and Z_2, dividing by the Jacobian determinant of the transformation, and expressing the z_j in terms of the w_i. The Jacobian determinant of a rotation is one, so there is no problem with that. Furthermore $z_1^2 + z_2^2 = w_1^2 + w_2^2$ is invariant under rotation. The joint density is thus

$$\frac{1}{2\pi}\exp(-\frac{1}{2}(w_1^2 + w_2^2)) = \frac{1}{2\pi}\exp(-\frac{1}{2}w_1^2)\exp(-\frac{1}{2}w_2^2) = \rho(w_1)\rho(w_2). \tag{13.2}$$

So W_1 and W_2 are also independent standard normal variables. This completes the proof.

Exercise 13.1. Show that the case $n = 2$ of the theorem implies the general case.

Exercise 13.2. Show that $\int_{-\infty}^{\infty} \rho(z)\,dz = 1$. The trick is to first show that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \rho(z_1)\rho(z_2)\,dz_1\,dz_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}e^{-\frac{1}{2}(z_1^2+z_2^2)}\,dz_1\,dz_1 = 1. \tag{13.3}$$

Since the standard normal distribution is universal, it is useful to remember some numerical facts about it. The probability that $|Z| < 1$ is about 68 percent, somewhat over two thirds. The probability that $|Z| < 2$ is about 95 percent. These are rough figures, but they are simple enough to memorize.

Exercise 13.3. Find $\mathbf{P}[0 < Z < 1]$. Find $\mathbf{P}[1 < Z < 2]$. Find $\mathbf{P}[2 < Z]$.

Exercise 13.4. Show that for $x > 0$ we have $\mathbf{P}[Z > x] < (1/x)(1/\sqrt{2\pi})\exp(-x^2/2)$.

Exercise 13.5. Estimate $\mathbf{P}[Z > 3]$.

Now we turn to the central limit theorem. Recall that in the weak law of large number we took a sum of independent random variables and divided by the number of terms. The result was convergence to a constant. However this does not give a very precise idea of how the sum diverges. The central limit theorem considers the situation when one divides by the square root of the number of terms. Then there is a remarkable universal behavior given by the normal (Gaussian) distribution.

For convenience in the following we shall deal with centered random variables, so that we do not have to mention the means explicitly. One should keep in mind that we are thinking about fluctuations about the mean.

Consider a sequence of independent random variables $Y_1, \ldots, Y_n$ with mean zero and with finite variances $\sigma_1^2, \ldots, \sigma_n^2$. Let

$$s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2 \tag{13.4}$$

be the variance of the sum of the first n random variables. Consider the ratio Y_i/s_n of an individual observation to the standard deviation of the sum. Fix $\epsilon > 0$ independent of n. The observation Y_i is *large* if the ratio $|Y_i/s_n| > \epsilon$. Let

$$\left(\frac{Y_i}{s_n}\right)_{>\epsilon} = \left(\frac{Y_i}{s_n}\right) 1_{|Y_i/s_n|>\epsilon}$$

be the large observation contribution to the ratio.

The *Lindeberg condition* is that for each $\epsilon > 0$ the contribution of the large observations is small in the sense that

$$\mathbf{E}[\sum_{i=1}^{n} \left(\frac{Y_i}{s_n}\right)^2_{>\epsilon}] \to 0 \tag{13.5}$$

as $n \to \infty$.

Example 13.1. Let the Y_i all have the same distribution, so each $\sigma_i^2 = \sigma^2$ and $s_n^2 = n\sigma^2$. This is the most classical case of the central limit theorem. Then

$$\mathbf{E}[\sum_{i=1}^{n} \left(\frac{Y_i}{s_n}\right)^2_{>\epsilon}] = \frac{1}{\sigma^2}\mathbf{E}[Y_1^2 1_{|Y_1|>\sqrt{n}\sigma\epsilon}] \to 0 \tag{13.6}$$

by the monotone convergence theorem. The Lindeberg condition is automatic in this case, since the contribution from each of the observations is comparable.

Example 13.2. . Let there be a fixed constant C such that each $|Y_i| \leq C$. Assume also that the total variance $s_n^2 \to \infty$ as $n \to \infty$. Then for n large enough there are no large observations. Clearly the Lindeberg condition is satisfied in this case.

Example 13.3. Of course there are examples where the Lindeberg condition is violated. The simplest such example is where Y_1 is a random variable and all $Y_i = 0$ for $i \geq 2$. Then $Y_1/s_n = Y_1/\sigma_1$, and so the first observation is large.

Remark: The Lindeberg condition implies that the ratio of an individual variance to the total variance is arbitrarily small. This is because $\sigma_i^2/s_n^2 = \mathbf{E}[(Y_i/s_n)^2]$ is bounded by $\epsilon^2 + \mathbf{E}[(Y_i/s_n)^2_{>\epsilon}]$. For each $\epsilon > 0$ we can make this less than $2\epsilon^2$ for n sufficiently large.

Theorem 13.2 (Central Limit Theorem) Let $Y_1, \ldots, Y_n, \ldots$ be a sequence of independent random variables with mean zero and with finite variances $\sigma_1^2, \ldots, \sigma_n^2, \ldots$. Assume the Lindeberg condition. Then the distribution of

$$\frac{Y_1 + \cdots + Y_n}{\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}} \tag{13.7}$$

converges weakly to the standard normal distribution as $n \to \infty$.

Proof: It is sufficient to show that the distributions converge in the Schwartz sense. The key will be the observation that Y_n and $Y_n^2 - \sigma_n^2$ are martingale differences, and hence weighting them with a gambling system does not change the mean from zero.

The proof uses the device of comparing the independent random variables $Y_1, \ldots, Y_n$ with independent normal random variables $\sigma_1 Z_1, \ldots, \sigma_n Z_n$. We take the Z_i to be standard normal random variables, and we take $\sigma_i^2 = \mathbf{E}[Y_i^2]$. Write $s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$ as before. Our explicit calculation shows that $Z = (\sigma_1 Z_1 + \cdots \sigma_n Z_n)/s_n$ is standard normal.

Let f be a function in C_c^∞. We want to estimate the expectation of the difference

$$f(\frac{Y_1 + \cdots + Y_n}{s_n}) - f(Z). \tag{13.8}$$

We replace the Y_k by the $\sigma_k Z_k$ one by one. Let

$$U_k = \frac{Y_1 + \cdots + Y_{k-1} + \sigma_{k+1} Z_{k+1} + \cdots \sigma_n Z_n}{s_n}. \tag{13.9}$$

Notice that this depends only on the first $k-1$ observations. Then the difference is the sum of the differences

$$f(U_k + \frac{Y_k}{s_n}) - f(U_k + \frac{\sigma_k Z_k}{s_n}) \tag{13.10}$$

from $k = 1$ to n. So we have n terms to estimate.

Figure 10.

Expand each difference in a Taylor series about U_k. The zero order term is zero. The first order term is

$$f'(U_k)\frac{Y_k - \sigma_k Z_k}{s_n} \tag{13.11}$$

and by independence has expectation zero. The second order term is

$$\frac{1}{2} f''(U_k)\frac{Y_k^2 - \sigma_k^2 Z_k^2}{s_n^2} \tag{13.12}$$

and by independence it also has expectation zero.

We need to worry about the remainder. The most convenient way to do this is to use Taylor's theorem with a second order remainder. We have already shown that the second order terms cancel exactly, so it is sufficient to estimate

$$\frac{1}{2}\left(f''(U_k + \alpha\frac{Y_k}{s_n}) - f''(U_k)\right)\frac{Y_k^2}{s_n^2} \tag{13.13}$$

where $|\alpha| \leq 1$ and

$$\frac{1}{2}\left(f''(U_k + \beta\frac{\sigma_k Z_k}{s_n}) - f''(U_k)\right)\frac{\sigma_k^2 Z_k^2}{s_n^2} \tag{13.14}$$

where $|\beta| \leq 1$.

The first of these expressions involves Y_k^2/s_n^2. Break it into parts where $|Y_k/s_n|$ is $\leq \epsilon$ (small observation) and $> \epsilon$ (large observation). For the small observation part we bound the difference of the second derivatives by a constant time the difference of the arguments, which is bounded by ϵ. It follows that the small observation part is bounded by a constant times $\epsilon Y_k^2/s_n^2$. For the large observation part we can simply bound the second derivative. This part is bounded by a constant times $(Y_k/s_n)^2_{>\epsilon}$.

Fix $\epsilon > 0$ small and consider the sum of the expectations from $k = 1$ to n. The small observation contribution part is bounded by a multiple of ϵ. Then for large n the large observation component is arbitrarily small, by the Lindeberg condition.

It remains to deal with the other expression, the one involving Z_k. We bound the difference of the second derivatives by a constant times the difference of the arguments, which is bounded by $(\sigma_k/s_n)|Z|$. The resulting bound on the entire expression is a constant times σ_k^3/s_n^3 times $|Z|^3$. The expectation is bounded by a constant times σ_k^3/s_n^3. Write the last factor as σ_k/s_n times σ_k^2/s_n^2. The factor σ_k/s_n is arbitrarily small for n large enough, and the sum from $k = 1$ to n of σ_k^2/s_n^2 is one. This completes the proof.

Notice that the proof uses martingale ideas, in particular weighting by a "gambling scheme" that depends only on previous observations. This suggests that the proof could be extended to certain martingales, and in fact it can, in a rather routine way.

Exercise 13.6. If the Y_k are not assumed to be independent but only to be martingale differences, then it no longer follows that the $Y_k^2 - \sigma_k^2$ are martingale differences. Show that the proof can be carried out under the hypothesis that the $Y_k^2 - \sigma_k^2$ are approximate martingale differences in an average sense. This means that for every uniformly bounded sequence of gambling schemes W_k we have

$$\sum_{k=1}^{n} \mathbf{E}[W_k\left(\frac{Y_k^2 - \sigma_k^2}{s_n^2}\right)] \to 0 \tag{13.15}$$

What can we conclude from this theorem? It says that if we take n independent random variables and consider their sum, and if there are no significant contributions to the variance from relatively large values, then the standardized version of the sum is approximately normal. Notice that there is no requirement that the distributions have the same distribution or even the same variance. This may help explain why the normal distribution sometimes arises in nature.

A most important special case is that of independent and identically distributed random variables with finite variance. An experimenter may create this situation by repeating an experiment under identical conditions. We state this as a corollary.

Theorem 13.3 (Central Limit Theorem) Let $Y_1, \ldots, Y_n, \ldots$ be a sequence of independent and identically distributed random variables with mean μ and with finite variances σ^2. Then the distribution of

$$\frac{Y_1 + \cdots + Y_n - n\mu}{\sigma\sqrt{n}} \tag{13.16}$$

converges weakly to the standard normal distribution as $n \to \infty$.

The conclusion may also be stated in the form: the distribution of

$$\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \tag{13.17}$$

approaches a standard normal distribution as $n \to \infty$.

One fact about the normal distribution that is easy to remember is that $\mathbf{P}[|Z| \leq 2]$ is approximately .95. So the probability that the sample mean is within $2\sigma/\sqrt{n}$ of the mean is about ninety five percent.

Recall that for the fraction of successes in binomial trials the σ is bounded above by $1/2$. So the probability that the sample proportion is within $1/\sqrt{n}$ of the probability of success exceeds ninety five percent. If one forgets everything else in these lectures, that fact is worth retaining.

Exercise 13.7. How large a sample needs to be taken in a poll so that the probability that the sample proportion is within 2.5 percent of the population proportion is at least 95 percent?

Exercise 13.8. A population of athletic men has an average weight of 93 kilograms with a standard deviation of 14 kilograms. A statistician does not know these figures, but is able to take a sample of 49 men and compute the sample mean. What is the probability that such a sample mean is with 2 pounds of 93?

Exercise 13.9. Consider the previous problem. The statistician can still only afford samples of size 49. How good are these samples? There is a value of a so that the probability that the sample mean is within a of 93 is ninety-five percent. Find this value.

Exercise 13.10. Consider the previous problems. The statistician is now willing to pay for a larger sample. How large should the sample size be so that the probability that the sample mean is within 2 pounds of 93 is ninety-five percent?

Exercise 13.11. Assume that angles are distributed uniformly over the interval from 0 to 2π. A statistician takes an independent sample of n angles Θ_i and computes the sample mean $\bar{\Theta}_n$. What does the central limit theorem say about the distribution of such sample means? Find constants c_n so that $(\bar{\Theta}_n - \pi)/c_n$ has a non-trivial limiting distribution.

Exercise 13.12. Another statistician computes the tangents $Y_i = \tan(\Theta_i)$ of the angles. Find the density of the random variables Y_i. Say that the statistician takes an independent sample of n tangents and computes the sample mean $\bar{Y}_n$. What does the central limit theorem say about the distribution of such sample means?

Exercise 13.13. The central limit theorem says that the sample mean $\bar{Y}_n$ when standardized is approximately normally distributed. How about $g(\bar{Y}_n)$ suitably standardized? Will it be approximately normally distributed? Assume that g is smooth and that $g'(\mu) \neq 0$.

Exercise 13.14. In the previous problem what happens if $g'(\mu) = 0$?

Lecture 14. Statistical estimation

Summary: One kind of statistical problem is when one has independent random variables with a fixed but unknown expectation. One wants to use some quantity such as the sample mean to estimate this unknown quantity. The central limit theorem gives a description of how well such procedures work.

This lecture is a very brief introduction to statistics. The most classical situation is an observation of a sequence of independent and identically distributed random variables $Y_1, \ldots, Y_n$. However the distribution is unknown.

The statistician is allowed to perform the experiment and obtain a finite number of experimental numbers $Y_1(\omega), \ldots, Y_n(\omega)$ depending on the outcome ω. From these numbers the statistician is to guess the distribution, or at least to guess certain information about the distribution. In the usual language of statistics, knowledge of the distribution is knowledge of the *population*. The values of the random variable are the *sample*. One wants to infer knowledge of the population from observations of the sample.

For instance, assume that the Y_i all have mean $\mathbf{E}[Y_i] = \mu$. The statistician may decide to use the value of the sample mean

$$\bar{Y}_n = \frac{\sum_{i=1}^n Y_i}{n} \tag{14.1}$$

to guess the population mean μ. What are the consequences of this choice of method?

In the long run, the statistics community in adopting this method is measuring random variables of the form $\bar{Y}_n$. We know that these random variables have mean μ and standard deviation $\sigma/\sqrt{n}$, where σ is the standard deviation of an individual observation Y_i. Assume for the moment that σ is finite. Then for n large enough, depending on the (unknown) population standard deviation σ, the standard deviation of the sample mean $\sigma/\sqrt{n}$ is small. Furthermore, by the central limit theorem, the distribution of the sample mean is approximately Gaussian. So statisticians who use large enough samples should most of the time get rather accurate estimates.

The next problem is to guess the population standard deviation σ. It is natural to use the sample variance

$$V_n = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{n-1} \tag{14.2}$$

to estimate the population variance σ^2. The factor $n-1$ in the denominator is put there to ensure that $\mathbf{E}[V_n] = \sigma^2$. This is a convenient convention, but it obviously has little importance for large n.

Exercise 14.1. Show that $\mathbf{E}[\sqrt{V_n}] \neq \sigma$. Is it bigger or smaller?

Of course the mean and variance do not give in general all the information about the population. Furthermore, if the variance is large, then the sample mean may not be at all the best way of estimating the population mean! All these difficulties disappear if the population is known to be normal or Gaussian. In this case the distribution of each Y_i is determined by the two parameters μ and σ^2, and it is exceedingly natural to estimate these by the sample mean and sample variance.

In the general case one might want to think of other methods. There are many devices. The simplest one is the use of *order statistics*. We consider the case when the distribution of the Y_i is determined by the (unknown) function $F(y) = \mathbf{P}[Y_i \leq y]$. Let $Y_{[1]} \leq \ldots \leq Y_{[i]} \leq \ldots \leq Y_{[n]}$ be the $Y_1, \ldots, Y_n$ arranged in increasing order. Then $Y_{[i]}$ is called the ith order statistic. The order statistics divide the real axis into $n+1$ random sample intervals.

Exercise 14.2. Find the distribution function of the ith order statistic.

Exercise 14.3. Show that the order statistics are not independent.

Fix a value of t between 0 and 1, and assume that there is a value of y with $F(y) = t$. Thus the proportion of the population with values less than or equal to y is t. Then one might want pick i so that $i/(n+1)$ is close to t and use $Y_{[i]}$ to estimate y. This is reasonable because the proportion of the sample intervals containing values less than $Y_{[i]}$ is $i/(n+1)$.

The case when the population is divided into two equal halves is of special importance. If $t = 1/2$, then the corresponding y is called the *population median.* If n is odd the middle order statistic $Y_{[(n+1)/2]}$ is called the *sample median.* One can take the sample median as an estimate of the population median.

How well do the procedures based on order statistics work? The following theorem provides the answer.

Theorem 14.1 Let $F(y) = t$ and assume that the density $F'(y) = f(y) > 0$. Then as $n \to \infty$ with $\sqrt{n}(i/n - t) \to 0$, the order statistic $Y_{[i]}$ is approximately normally distributed with mean y and standard deviation $\sqrt{t(1-t)}/(f(y)\sqrt{n})$.

Proof: Let $N(x)$ be the number of i such that $Y_i \leq x$. This is a binomial random variable with mean $nF(x)$ and variance $nF(x)(1-F(x))$. Furthermore $Y_{[i]} \leq x$ is equivalent to $N(x) \geq i$. If we let

$$Z = \frac{N(x)/n - F(x)}{\sqrt{F(x)(1-F(x))/n}} \tag{14.3}$$

be the standardized random variable corresponding to $N(x)$, then $Y_{[i]} \leq x$ is equivalent to

$$-Z \leq \frac{F(x) - i/n}{\sqrt{F(x)(1 - F(x))/n}} \tag{14.4}$$

Take $x = y + \epsilon/\sqrt{n}$. Then the right hand side converges to $\epsilon f(y)/\sqrt{t(1-t)}$. So the probability that $Y_{[i]} - y \leq \epsilon/\sqrt{n}$ is the same as the probability that a standard normal random variable $-Z$ does not exceed $\epsilon f(y)/\sqrt{t(1-t)}$.

Notice that it is possible that the statistician knows that the population mean is the same as the population median. In this case one can use either the sample mean or the sample median to estimate this number. Which is better to use? For the sample mean to be appropriate, the variance must be small—extremely large observations must be unlikely. For the sample median to be appropriate, the density at the population mean must not be too close to zero—however large observations are irrelevant. Knowing which to use seems to require some knowledge of the population, which is exactly what the statistician does not have.

Exercise 14.4. Compare the variances of the sample mean and the sample median for the case of a normal population. Of course this is a case when the advantage will go to the sample mean; it will have the smaller variance. By enough to make a difference?

Exercise 14.5. Find an example where the sample median has a much smaller variance than the sample mean.

Appendix: Final Examination

Martingale Ideas in Elementary Probability
Spring 1996
Final Examination

0. Basic definitions

A *supermartingale* is a sum of random variables $S_n = S_0 + Y_1 + \ldots + Y_n$, where each Y_i is a *supermartingale difference*. This means that for every sequence of *gambling scheme* random variables $W_i \geq 0$ that depend only on $S_0, Y_1, \ldots Y_{i-1}$ we have $\mathbf{E}[W_i Y_i] \leq 0$. Intuition: a supermartingale is an unfavorable game that cannot be made favorable by a gambling scheme.

A gambling scheme converts a supermartingale into a supermartingale. Thus if $S_n = S_0 + Y_1 + \cdots Y_n$ is a supermartingale and if X_0 is a function of S_0 and $Z_i = W_i Y_i$, then $X_n = X_0 + Z_1 + \cdots + Z_n$ is also a supermartingale.

The *supermartingale convergence theorem* says that a supermartingale that is bounded below must converge with probability one.

A *submartingale* is defined in the same way, except that $\mathbf{E}[W_i Y_i] \geq 0$. A *martingale* is both a supermartingale and a submartingale. For a martingale $\mathbf{E}[W_i Y_i] = 0$ without the restriction on the sign of W_i.

1. Waiting times

A random variable W is an exponential waiting time with mean $1/\lambda$ if

$$\mathbf{P}[W > t] = e^{-\lambda t}$$

for $t \geq 0$. Thus it has density $\lambda e^{-\lambda t}$ for $t \geq 0$. Let $W_0, W_1, \ldots, W_n, \ldots$ be independent exponential waiting times with mean $1/\lambda$. Think of W_0 as the time you have to wait for the tram. Let N be the least $n \geq 1$ so that $W_n > W_0$. Thus N is the number of trams until a longer wait. Find the probability $\mathbf{P}[N \geq k]$. Find the probability $\mathbf{P}[N = k]$. Find the expectation $\mathbf{E}[N]$.

2. Matched pairs

In a medical experiment people with similar characteristics are paired. There are 100 such pairs. In each pair a drug is given to one individual and a placebo is given to the other. An independent evaluator is told to guess which individual has been given the drug. The drug has no effect. What is the chance that the evaluator guesses the correct individual in more than 60 of the pairs?

3. Random growth

A random variable Y is uniform on $[0, a]$ if $\mathbf{P}[Y \leq y] = y/a$ for $0 \leq y \leq a$. Thus it has density $1/a$ for $0 \leq y \leq a$. Here is a model for random growth of a population. The first generation is some number $X_0 > 0$. Each succeeding generation $X_{n+1} = Y_n X_n$, where Y_n

is distributed uniformly on the interval from 0 to a. One might think that there would be an eventual exponential growth rate λ of the X_n. In that case, the rate λ could be computed as a limit of $\log(X_n)/n$. What is λ? In what sense does the convergence take place? How large must a be so that the growth rate is positive?

4. Statistics

A random variable Z is standard normal if it has density

$$\rho(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

A statistician is measuring independent standard normal random variables $Z_1, \ldots, Z_n$. However he is worried that he might be instead dealing with normal random variables that have mean $\epsilon > 0$ and variance one. He therefore decides to compute the "likelihood ratio"

$$X_n = \frac{\rho(Z_1 - \epsilon) \cdots \rho(Z_n - \epsilon)}{\rho(Z_1) \cdots \rho(Z_n)}.$$

He will be reassured if this ratio is small.

a. Lemma: A product $X_n = R_1 \cdots R_n$ of independent positive random variables $R_i > 0$ with means $\mathbf{E}[R_i] = 1$ is a martingale. Prove this lemma.

b. Show that the likelihood ratio is not small on the average; in fact that X_n is a martingale whose expectation is always one.

c. What is the limit as $n \to \infty$ of the likelihood ratios X_n?

5. Insurance

An government insurance company has an equal chance of collecting a premium or paying a claim each day on which there is activity. However there are days when there is no activity at all. It starts with an initial reserve of $X_0 = x \geq 1$, where x is an integer. If it has a reserve $X_n \geq 1$ at day n, then at the next day it will gain or lose Y_{n+1}, where $Y_{n+1} = 1, 0, -1$ with probabilities $p_n/2$, $(1 - p_n)$ and $p_n/2$. Thus $X_{n+1} = X_n + Y_{n+1}$ for $X_n \geq 1$. If $X_n = 0$, then the company is declared bankrupt and $X_{n+1} = 0$.

a. Show that the reserve X_n is a martingale. Show that for each n the expected value of X_n is x.

b. Show that with probability one X_n has a limit X_∞ as $n \to \infty$.

c. Show that if there is rapidly decreasing activity, that is, $\sum_n p_n < \infty$, then the expected value of X_∞ is x.

d. Show that otherwise, if $\sum_n p_n = \infty$, then the expected value of X_∞ is not x. What is it? Hint: If activity eventually ceases, then there is some n such that $Y_k = 0$ for all $k \geq n$. Calculate the probability that $Y_k = 0$ for all $k \geq n$.

6. A queue

A random variable ξ has a Poisson distribution with mean $\mu > 0$ if

$$\mathbf{P}[\xi = k] = \frac{\mu^k}{k!} e^{-\mu}.$$

Here is a model for emptying a queue. Let $\xi_1, \xi_2, \xi_3, \ldots$ be independent random variables, each with a Poisson distribution with mean μ. There are X_n people in the queue. If $X_n \geq 1$, then one person is served and then ξ_{n+1} new people enter the queue. Thus

$$X_{n+1} = X_n - 1 + \xi_{n+1}$$

when $X_n \geq 1$. If $X_n = 0$, then the queue is empty. Work stops, and $X_{n+1} = 0$. We start with $X_0 = x \geq 1$.

a. When is the queue X_n a supermartingale, a submartingale, a martingale? What does the supermartingale convergence theorem say about the long term behavior of the queue when $\mu \leq 1$?

b. Show that if $\mu > 1$ there is a solution $\rho < 1$ of the equation $\mathbf{E}[\rho^\xi] = \rho$.

c. Show that ρ^{X_n} is a martingale.

d. Use this martingale to find the probability in the case $\mu > 1$ that the queue ever empties, that is, that X_n is eventually zero.

e. Let $\tilde{X}_n = X_n + (1 - \mu)n$ when $X_n \geq 1$ and $\tilde{X}_n = (1 - \mu)T$ when $X_n = 0$, where T is the least n such that $X_n = 0$. Show that $\tilde{X}_n$ is a martingale. Use this martingale to show that when $\mu < 1$ the expected time $\mathbf{E}[T]$ to empty the queue is finite.

Answers

1. The joint distribution of $W_0, \ldots, W_{k-1}$ is symmetric under permutations of the random variables. Thus $\mathbf{P}[N \geq k] = \mathbf{P}[W_0 \geq W_1, \ldots, W_0 \geq W_{k-1}] = 1/k$. Hence $\mathbf{P}[N = k] = 1/k - 1/(k+1) = 1/(k(k+1))$. Finally $\mathbf{E}[N] = \sum_{k=1}^{\infty} k\mathbf{P}[N = k] = \sum_{k=1}^{\infty} 1/(k+1) = \infty$.

2. Here $N_{100} = Y_1 + \ldots + Y_{100}$, where $Y_i = 1$ or 0 with probability $1/2$. Thus $\mu = 1/2$ and $\sigma = 1/2$. Hence the standardized random variable is

$$Z = \frac{N_{100} - 100\frac{1}{2}}{\sqrt{100}\frac{1}{2}} = \frac{N_{100} - 50}{5}.$$

According to the central limit theorem

$$\mathbf{P}[N_{100} \geq 60] = \mathbf{P}[Z \geq 2] = .025.$$

3. We have

$$\frac{\log X_n}{n} = \frac{\log X_0}{n} + \frac{\log Y_1 + \cdots + \log Y_n}{n} \to \lambda$$

as $n \to \infty$, where

$$\lambda = \mathbf{E}[Y_1] = \int_0^a \log y \, dy/a = \log a - 1.$$

According to the strong law of large numbers, the convergence takes place with probability one. If $a > e$ then $\lambda > 0$.

4. a. A product $X_n = R_n \cdots R_1$ of independent positive random variables each with mean one is a martingale. This is because the difference $X_{n+1} - X_n = (R_{n+1} - 1)X_n$ is the product of a martingale difference by a weight from a gambling scheme.

b. This is the situation in this model, since

$$\mathbf{E}[\frac{\rho(Z_n - \epsilon)}{\rho(Z_n)}] = \int \rho(z - \epsilon) \, dz = 1.$$

c. This is a positive martingale; hence it must converge with probability one. The only possibility is for it to converge to zero, since otherwise each time multiplying by a factor of

$$\frac{\rho(Z_n - \epsilon)}{\rho(Z_n)} = e^{\epsilon Z_n - \frac{1}{2}\epsilon^2}$$

would produce a large random change in the value. [One can also check convergence to zero by direct application of the strong law of large numbers.]

5. a. The random walk $S_n = x + Y_1 + \cdots Y_n$ is a martingale. The differences $Z_{n+1} = X_{n+1} - X_n$ are equal to Y_{n+1} times a factor that is 1 when $S_n \geq 1$ and to 0 when $S_n = 0$. This is an application of a gambling scheme, so X_n is also a martingale.

b. This martingale is bounded below, so it converges with probability one.

c. Let us compute

$$\mathbf{E}[X_n^2] = x^2 + \sum_{k=1}^n \mathbf{E}[Z_k^2] \leq x^2 + \sum_{k=1}^n \mathbf{E}[Y_k^2] = x^2 + \sum_{k=1}^n p_k \leq x^2 + \sum_{k=1}^\infty p_k.$$

Since this is bounded independently of n, the limit

$$\mathbf{E}[X_\infty] = \lim_n \mathbf{E}[X_n] = x.$$

d. By independence the probability that $Y_k = 0$ for all $k \geq n$ is the infinite product $\prod_{k \geq n}(1 - p_k)$. The logarithm of the infinite product is

$$\log \left(\prod_{k \geq n} (1 - p_k) \right) = \sum_{k \geq n} \log(1 - p_k) \leq - \sum_{k \geq n} p_k = -\infty$$

since $\log(1 - p) = -p - p^2/2 - p^3/3 - \cdots \leq -p$. The probability must thus be zero.

It follows by countable additivity that the probability that there exists n with $Y_k = 0$ for $k \geq n$ is zero. Hence the random walk keeps fluctuating. The only way for the martingale to converge is for it to reach zero. The company will be bankrupt with probability one.

6. a. Consider the random walk $S_n = x + 1 - \xi_1 + \cdots + 1 - \xi_n$. This is a supermartingale, submartingale, or martingale according to whether $\mu \leq 1$, $\mu \geq 1$, or $\mu = 1$. The process X_n is obtained by a gambling scheme in which one stops playing when S_n first reaches zero. Thus it shares the same properties. Suppose $\mu \leq 1$. Then the supermartingale is bounded below, so it must converge. It can only converge to zero. Conclusion: if the queue is not overloaded on the average, then it eventually empties.

b. The equation $\mathbf{E}[\rho^{\xi}] = \rho$ says that

$$e^{\mu(\rho-1)} = \rho.$$

At $\rho = 0$ the left hand side is larger than the right hand side. There is a fixed point at $\rho = 1$. At the fixed point the derivative of the left hand side is μ and the derivative of the right hand is 1. So if $\mu > 1$, then the left hand side must be smaller than the right hand side for $\rho < 1$ but near enough to 1. By the intermediate value theorem there must be a solution strictly between 0 and 1.

c. The difference

$$\rho^{X_{n+1}} - \rho^{X_n} = (\rho^{\xi_{n+1}} - 1)\rho^{X_n}$$

when $X_n \geq 1$ and is zero when $X_n = 0$. This is an application of a gambling scheme. Furthermore, when ρ is a solution of the equation this is an application of a gambling scheme to a martingale, so it is a martingale.

d. The expectation of this martingale ρ^{X_n} is constant, so it is always ρ^x. Since the martingale is bounded below, it converges with probability one. In fact, since $0 < \rho < 1$, it converges to 0 when $X_n \to \infty$ and it converges to 1 when $X_n \to 0$. Since the martingale is also bounded above, the expectations converge. Thus the probability that $X_n \to 0$ is ρ^x.

e. Consider the random walk $x + (\xi_1 - \mu) + \cdots (\xi_n - \mu)$. This is a martingale equal to $S_n + (1-\mu)n$. If we stop this martingale when S_n reaches 0, this remains a martingale. This new martingale is $\tilde{X}_n$. The expectation of $\tilde{X}_n$ is the initial x, and so when $\mu \leq 1$

$$x = \mathbf{E}[(X_n + (1-\mu)n)1_{T>n}] + \mathbf{E}[(1-\mu)T1_{T\leq n}] \geq (1-\mu)\mathbf{E}[T1_{T\leq n}].$$

Hence when $\mu < 1$

$$\mathbf{E}[T1_{T\leq n}] \leq \frac{1}{1-\mu}.$$

By the monotone convergence theorem $\mathbf{E}[T] \leq 1/(1-\mu)$.

Мартингалы в теории вероятностей
Весна 1996
Семестровый экзамен

0. Основные определения.

Последовательность случайных величин $S_0, S_1, S_2, \dots$ называется *супермартингалом*, если последовательность разностей Y_n, определённых соотношениями $S_n = S_0 + Y_1 + \dots + Y_n$, обладает следующим свойством: для любого i и для любой случайной величины $W_i \ge 0$, зависящей только от $S_0, Y_1, \dots Y_{i-1}$, математическое ожидание $\mathbf{E}[W_i Y_i]$ неположительно. (Величины W_i можно рассматривать как ставки в игре; супермартингал — это невыгодная для нас игра, которая остаётся невыгодной, как бы мы не выбирали ставки.)

Последовательность ставок преобразует супермартингал в другой супермартингал. Именно, если $S_n = S_0 + Y_1 + \cdots Y_n$ — супермартингал, X_0 зависит только от S_0, а W_i обладает описанными выше свойствами, то последовательность $X_n = X_0 + Z_1 + \cdots + Z_n$, где $Z_i = W_i Y_i$, является супермартингалом.

Теорема *о сходимости для супермартингалов* утверждает, что любой ограниченный снизу супермартингал сходится с вероятностью 1.

Мы определяем *субмартингалы* аналогичным образом, только теперь требуется, чтобы $\mathbf{E}[W_i Y_i]$ было неотрицательно. Супермартингал, являющийся одновременно субмартингалом, называется *мартингалом*. Для мартингала равенство $\mathbf{E}[W_i Y_i] = 0$ выполнено и для величин W_i, не являющихся неотрицательными.

1. Время ожидания.

Говорят, что величина W имеет показательное распределение со средним $1/\lambda$, если

$$\mathbf{P}[W > t] = e^{-\lambda t}$$

для любого $t \ge 0$. (В этом случае плотность вероятности будет $\lambda e^{-\lambda t}$ для $t \ge 0$.) Пусть величины $W_0, W_1, \dots, W_n, \dots$ независимы и имеют показательное распределение со средним $1/\lambda$. (Придя на трамвайную остановку, мы ждём W_0 минут до первого трамвая, затем W_1 минут от первого до второго и так далее.) Пусть N — наименьшее положительное целое число, для которого $W_N > W_0$ (число трамваев до первого большего интервала). Найти вероятность $\mathbf{P}[N \ge k]$. Найти вероятность $\mathbf{P}[N = k]$. Найти математическое ожидание $\mathbf{E}[N]$.

2. Пары

Для проверки действия нового лекарства было составлено 100 пар людей с близкими физическими данными. В каждой паре один пациент получал лекарство, другой — плацебо (пустышку). Затем независимый наблюдатель должен быль указать (по результатам наблюдений), какой из пациентов (в каждой паре) получал лекарство. Какова вероятность того, что он даст правильный ответ для более чем 60 пар из 100, если на самом деле лекарство не действует вовсе?

3. Случайный рост

Говорят, что случайная величина Y равномерно распределена на отрезке $[0, a]$, если $\mathbf{P}[Y \leq y] = y/a$ для любого y из отрезка $0 \leq y \leq a$. (Она имеет плотность $1/a$ на всём этом отрезке.) Рассмотрим следующую модель роста популяции. В первом поколении имеется некоторое число $X_0 > 0$ особей. В каждом следующем поколении пусть будет $X_{n+1} = Y_n X_n$ особей, где Y_n равномерно распределено в отрезке от 0 до a. Можно ожидать, что число особей будет экспоненциально расти с показателем λ, который можно искать как предел $\log(X_n)/n$ при $n \to \infty$. В каком смысле эта сходимость будет иметь место? Как велико должно быть a, чтобы скорость роста была положительной?

4. Статистика

Случайная величина Z называется стандартно распределённой нормальной случайной величиной, если её плотность есть

$$\rho(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

Статистик наблюдает независимые стандартно распределённые нормальные случайные величины $Z_1, \dots, Z_n$. Однако он подозревает, что на самом деле эти величины имеют математическое ожидание $\epsilon > 0$ и вариацию 1. Поэтому он вычисляет "отношение правдоподобия"

$$X_n = \frac{\rho(Z_1 - \epsilon) \cdots \rho(Z_n - \epsilon)}{\rho(Z_1) \cdots \rho(Z_n)}.$$

— надеясь, что оно окажется малым.

a. Лемма: произведение $X_n = R_1 \cdots R_n$ независимых положительных случайных величин $R_i > 0$ с математическим ожиданием 1 является мартингалом. Докажите это.

b. Покажите, что отношение правдоподобия в среднем вовсе не является малым; на самом деле X_n есть мартингал с единичным математическим ожиданием.

c. Чему равен предел (при $n \to \infty$) отношений X_n?

5. Страховка

Государственная страховая компания имеет равные шансы получить страховой взнос или заплатить страховую сумму (если вообще что-то произойдёт — может не произойти ничего!) Начальный резерв равен $X_0 = x \geq 1$, где x — некоторое целое число. Если к n-му дню резерв был равен $X_n \geq 1$, то на следующий день резерв изменится на Y_{n+1}, где $Y_{n+1} = 1, 0, -1$ с вероятностями соответственно $p_n/2$, $(1 - p_n)$ и $p_n/2$ и станет равным $X_{n+1} = X_n + Y_{n+1}$ (в случае $X_n \geq 1$). Если же $X_n = 0$, то компания считается банкротом и $X_{n+1} = 0$.

a. Покажите, что X_n является мартингалом и что для любого n математическое ожидание величины X_n равно x.

b. Покажите, что с вероятностью 1 последовательность X_n имеет предел при $n \to \infty$ (который мы обозначим X_∞).

c. Покажите, что если активность быстро падает в том смысле, что $\sum_n p_n < \infty$, то математическое ожидание X_∞ равно x.

d. Покажите, что в случае $\sum_n p_n = \infty$ математическое ожидание X_∞ не равно x. Как так может быть? (Указание. Если деятельность в какой-то момент прекращается, то существует такое n, что $Y_k = 0$ для всех $k \geq n$. Найдите вероятность того, что $Y_k = 0$ для всех $k \geq n$.)

6. Очередь

Говорят, что случайная величина ξ имеет пуассоновское распределение со средним $\mu > 0$, если

$$\mathbf{P}[\xi = k] = \frac{\mu^k}{k!}e^{-\mu}.$$

Рассмотрим модель очереди. Пусть $\xi_1, \xi_2, \xi_3, \ldots$ — независимые пуассоновски распределенные случайные величины со средним μ. Величина X_n — длина очереди на n-ом шаге. Если $X_n \geq 1$, то один клиент обслуживается, а в хвост очереди встают ξ_{n+1} новых клиентов, так что

$$X_{n+1} = X_n - 1 + \xi_{n+1}$$

(при $X_n \geq 1$). Если же $X_n = 0$, очередь кончается, лавка закрывается и $X_{n+1} = 0$. В начале $X_0 = x \geq 1$.

a. В каком случае последовательность X_n образует супермартингал, субмартингал, мартингал? Что утверждает теорема о сходимости для супермартингалов по поводу поведения очереди для случая $\mu \leq 1$?

b. Покажите, что при $\mu > 1$ найдется $\rho < 1$, являющееся решением уравнения $\mathbf{E}[\rho^\xi] = \rho$.

c. Покажите, что ρ^{X_n} является мартингалом.

d. Используйте этот мартингал, чтобы найти (для случая $\mu > 1$) вероятность того, что очередь рано или поздно опустеет,. то есть что $X_n = 0$ для достаточно больших n.

e. Пусть $\tilde{X}_n = X_n + (1-\mu)n$ при $X_n \geq 1$ и $\tilde{X}_n = (1-\mu)T$ при $X_n = 0$, где T — наименьшее число n, для которого $X_n = 0$. Покажите, что $\tilde{X}_n$ является мартингалом. Используя это, покажите, что при $\mu < 1$ математическое ожидание $\mathbf{E}[T]$ (где T — время до закрытия лавки) конечно.