THE LAW OF LARGE NUMBERS

JEREMY SCHWARTZ

Abstract. This paper presents an introduction to probability theory leading up to a proof of the weak law of large numbers. Probability spaces and random variables are covered fairly rigorously. Properties of the expectation operator are proved for the case of discrete random variables, and convergence of random variables and characteristic functions are presented for the purpose of proving the weak law of large numbers. The strong law of large numbers is also presented without proof.

Contents

1. Probability Spaces

Probability is fundamentally concerned with experiments or games which have non-deterministic outcomes. All the possible outcomes can be described as a set:

Definition 1. *A sample space is the set Ω of all possible outcomes. Individual outcomes are typically denoted by ω*

Example 1. *If we rolled a standard six sided die, one way to represent the possible outcomes is the set containing the numbers 1 through 6. In this case, the sample space would be:*

$$\Omega = \{1, 2, 3, 4, 5, 6\} \tag{1.1}$$

We are frequently concerned with more than just the elementary outcomes of an experiment though. Staying with this example, we might be interested in whether the die shows an odd number, a number greater than four, etc. This gives rise to the following definition

Definition 2. *An event is a subset $A \subseteq \Omega$ which can be assigned a probability. The set of all events is denoted by $\mathcal{F}$*

For fairly complicated reasons, not all subsets of the sample space are necessarily events. However, we can use our intuition to determine some properties of $\mathcal{F}$

(1) If A is an event, then the non-occurence of A is an event
i.e. $A \in \mathcal{F} \Rightarrow A^c := \Omega / A \in \mathcal{F}$
(2) If A and B are events, then (A and B) is an event and (A or B) is an event
i.e. $A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$ and $A \cup B \in \mathcal{F}$
(3) There is a sure event (the event that any ω occurs) and an impossible event (the event that no ω occurs)
i.e. $\Omega, \emptyset \in \mathcal{F}$

Any collection $\mathcal{F}$ of subsets which satisfies these conditions is called an algebra. The following example will demonstrate that a good theory of probability ought to include a slightly stronger condition for a collection of events.

Example 2. *You are offered a game where you and an opponent flip a coin until it comes up heads. You win if the first heads appears on an even toss. The sample space here is $\Omega = \{\omega_1, \omega_2, \omega_3, ...\}$ where ω_i is the outcome that the first i-1 tosses are tails and the ith toss is heads. In this case, the event that you win is defined as $A = \{\omega_2, \omega_4, \omega_6, ...\}$. This is an infinite countable union of outcomes, and to discuss its probability, we need it to be an element of $\mathcal{F}$.*

An algebra which is closed under countable unions is called a σ-algebra

Definition 3. *A collection $\mathcal{F}$ of subsets of Ω is a σ-algebra if it satisfies:*

(1) $\emptyset \in \mathcal{F}$
(2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
(3) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \displaystyle\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Remark 1. *A σ-algebra is closed under countable intersections, i.e.*

$$A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in F$$

Proof. This is just an extension of De Morgan's law, which states for $A, B \subseteq \Omega$

$$(A \cap B)^c = A^c \cup B^c$$

□

It is easy to check that a σ-algebra satisfies the conditions of an algebra, and therefore contains our intuitive requirements for the set of events. Finally, we need some way to actually assign probabilities to events. This is accomplished by a probability measure:

Definition 4. *A probability measure $\mathbb{P}$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying*

(1) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$
(2) *if $A_1, A_2, ...$ is a collection of disjoint members of $\mathcal{F}$, in that for all $i \neq j, A_i \cap A_j = \emptyset$, then*

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \tag{1.2}$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

Sometimes, the occurence of one event causes another event to be more or less likely. For instance, if it rains in my neighborhood, there is a greater chance of a bus being late. Results will often only hold for events which do not influence each other, prompting the following definition

Definition 5. *Events A and B are called independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{1.3}$$

More generally, a family $\{A_i : i \in I\}$ of events is called independent if

$$\mathbb{P}(\bigcap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i)$$

for all finite subsets J of I

Example 3. *We roll a six-sided die again, each face appearing with probability 1/6. The event A that the die shows an odd value is not independent from the event B that it shows a 5 because*

$$\mathbb{P}(A \cap B) = \mathbb{P}(5) = 1/6 \tag{1.5}$$

while

$$\mathbb{P}(A)\mathbb{P}(B) = (1/2)(1/6) = 1/12 \tag{1.6}$$

Example 4. *If we choose a card from a standard deck at random (each card having probability 1/52 of being chosen), then the event that the card is a spade is independent of the event that the card is an ace because*

$$\mathbb{P}(spade) = \mathbb{P}(2\spadesuit \cup 3\spadesuit \cup ... \cup ace\spadesuit) = \sum_{1}^{13}(1/52) = 1/4$$

and

$$\mathbb{P}(ace) = \mathbb{P}(ace\spadesuit \cup ace\clubsuit \cup ace\diamondsuit \cup ace\heartsuit) = \sum_{1}^{4}(1/52) = 1/13$$

so if A={the event that a card is a spade} and B={the event that a card is an ace}, the event A∩B={the event that a card is the ace of spades} and

$$\mathbb{P}(A \cap B) = 1/52 = \mathbb{P}(A)\mathbb{P}(B) \tag{1.9}$$

2. Random Variables

It is often the case that outcomes of a probabilistic experiment correspond to real number values. For instance, a gambler gains or loses some amount of money as a result of a game. This value can be of greater interest than the outcome itself.

Definition 6. *A random variable X is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$*

Most people would be hard-pressed to find a function $X : \Omega \to \mathbb{R}$ which didn't satisfy the second part of this definition. Basically, it requires that the pre-image of $[-\infty, x]$ be an event in $\mathcal{F}$. Since events occur with different probabilities, a random variable is more or less likely to take certain values in $\mathbb{R}$. We describe this with a distribution function.

Definition 7. *The distribution function F of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ given by $F(x) = \mathbb{P}(X \leq x)$*

Random variables can either take values in a countable or uncountable subset of $\mathbb{R}$, prompting the following distinction

Definition 8. (1) *A random variable X is called* ***discrete*** *if it only takes values in a countable subset of $\mathbb{R}$. In this case, there is a corresponding* ***probability mass function*** *defined by $f(x) = \mathbb{P}(X = x)$*

(2) *A random variable X is called* ***continuous*** *if its distribution function F can be expressed as $F(x) = \int_{\infty}^{x} f(u)du$ for some integrable function $f : \mathbb{R} \to \mathbb{R}^+$ in which case f is called the* ***probability density function***

Remark 2. *A random variable can be neither continuous nor discrete*

Proof. Consider a random variable which takes the value 0 with probability 1/2 and a value in the interval $[\frac{1}{2}, 1]$ with probability density 1. □

From here on out, definitions will be given for both discrete and continuous random variables, but for the sake of space, results will only be demonstrated for discrete variables. In the previous section, we wanted to know if two events were independent. This generalizes to two random variables.

Definition 9. (1) *Discrete random variables X and Y are independent if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all $x, y \in \mathbb{R}$*

(2) *Continuous random variables X and Y are independent if the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for all $x, y \in \mathbb{R}$*

Since the pre-image of any real number value is an event, this definition is essentially just requiring that the events $X^{-1}(x)$ and $Y^{-1}(y)$ be independent for all x and y. We will frequently make use of one special type of random variable.

Definition 10. *An* ***indicator variable*** *for some event A is the function*

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \tag{2.1}$$

It is clear from definitions that for a random variable X and a real, measurable function $g : \mathbb{R} \to \mathbb{R}$, g(X) is also a random variable. One thing that we would like to show is that for independent random variables X and Y, and real, measurable functions $g, h : \mathbb{R} \to \mathbb{R}$, $g(X)$ and $h(Y)$ are independent.

Theorem 1. *If random variables X and Y are independent, and $g, h : \mathbb{R} \to \mathbb{R}$ are measurable functions, $g(X)$ and $h(Y)$ are independent.*

Proof. Let $A_x = g^{-1}(x)$ and $B_y = h^{-1}(y)$ be events. Now $g(X(\omega_1)) = x$ when $\omega_1 \in A_x$ and $h(Y(\omega_2)) = y$ when $\omega_2 \in B_y$. It follows from the remark after definition 9 that X and Y are independent for the events A_x and B_y. It follows that g(X) and h(Y) are independent for all $x, y \in \mathbb{R}$. □

We will end this section with a proposition which will be useful later.

Proposition 1. *Any discrete random variable can be written as a linear combination of indicator variables.*

Proof. Suppose X is a discrete random variable. This means that X takes values in the set $\{x_1, x_2, ...\}$. Consider the event $A_i = \{X = x_i\}$. Let I_{A_i} be the indicator function of A_i. It is clear that

$$X = \sum_i x_i I_{A_i}$$

□

3. Expectation

Most of us know that we should expect to lose money playing the lottery. Even though it is possible to win large amounts of money it is much more likely that we would lose a small amount of money. This idea, of weighting possible outcomes by the probability that they occur, allows us to determine the expectation for a random variable with an underlying probability space.

Definition 11. (1) *The expectation $\mathbb{E}$ of a discrete random variable X is defined*

$$\mathbb{E}(X) = \sum_{x:f(x)>0} xf(x)$$

where f(x) is the probability mass function.

(2) *The expectation of a continuous random variable X is defined by*

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

where f(x) is the probability density function

This is also sometimes referred to as the mean value or expected value of a random variable. For the sake of convenience, we will often write $\mathbb{E}(X) := \sum_x xf(x)$ for a discrete random variable. This appears to be an uncountable sum, but all but countably many of the entries are 0.

One of the fundamental properties of expectation is that it is linear i.e. for random variables X,Y and scalars a,b, $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$. To show this, we will first prove a very useful lemma

Lemma 1. *If $g : \mathbb{R} \to \mathbb{R}$ is a real valued function and X is a discrete random variable with mass function f, then $\mathbb{E}(g(X)) = \sum_x (g(x)f(x))$*

Proof. Define the event $A_x := \{X = x\}$ so that $f(x) = \mathbb{P}(A_x)$. We want to find the probability mass function for g(X)

$$f_g(z) = \mathbb{P}(g(X) = z) = \mathbb{P}(\bigcup_{x:g(x)=z} (A_x)) = \sum_{x:g(x)=z} \mathbb{P}(A_x) = \sum_{x:g(x)=z} f(x)$$

Now we can plug this into the definition of expectation to get

$$\mathbb{E}(g(X)) = \sum_z g(z)f_g(z) = \sum_z g(z) \sum_{x:g(x)=z} f(x) = \sum_z \sum_{x:g(x)=z} g(z)f(x) \tag{3.4}$$

which is the same as $\sum_x g(x)f(x)$ □

Theorem 2. *For random variables X, Y and $a, b \in \mathbb{R}$, $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.*

Proof. Let $A_x = \{X = x\}, B_y = \{Y = y\}$. As in proposition 1, we can write the random variable $aX + bY$ as a linear combination of indicator variables.

$$aX + bY = \sum_{x,y} (ax + by)I_{A_x \cap B_y} \tag{3.5}$$

Furthermore, Lemma 1 shows that

$$\mathbb{E}(aX+bY) = \sum_{x,y}(ax+by)\mathbb{P}(A_x\cap B_y) = \sum_x ax \sum_y \mathbb{P}(A_x \cap B_y)+\sum_y by \sum_x \mathbb{P}(A_x \cap B_y)$$

However,

$$\sum_y \mathbb{P}(A_x \cap B_y) = \mathbb{P}(A_x \cap (\bigcup_y B_y)) = \mathbb{P}(A_x \cap \Omega) = \mathbb{P}(A_x) \tag{3.6}$$

and likewise $\sum_x \mathbb{P}(A_x \cap B_y) = \mathbb{P}(B_y)$, which gives

$$\mathbb{E}(aX + bY) = a\sum_x x\mathbb{P}(A_x) + b\sum_y y\mathbb{P}(B_y) = a\mathbb{E}(X) + b\mathbb{E}(Y) \tag{3.7}$$

□

And now we can finally make it clear why we've cared so much about independence

Theorem 3. *If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Proof. Let A_x and B_y be as in the proof of Theorem 2. We have

$$XY = \sum_{x,y} xyI_{A_x \cap B_y} \tag{3.8}$$

And so, by independence,

$$\mathbb{E}(XY) = \sum_{x,y} xy\mathbb{P}(A_x \cap B_y) = \sum_{x,y} xy\mathbb{P}(A_x)\mathbb{P}(B_y) = \sum_x x\mathbb{P}(A_x)\sum_y y\mathbb{P}(B_y) = \mathbb{E}(X)\mathbb{E}(Y)$$

□

4. Convergence of Random Variables

Let's return to the lottery example from the beginning of the previous section. Even though somebody wins the lottery every week, we reassure ourselves that we're making the right decision not to play by claiming that we'd surely lose money if we played enough. To study what happens in the long run for a repeated experiment, we consider infinite sequences of random variables. In particular, we are interested in when sequences of random variables converge. Random variables are functions from the sample space to the real numbers, so the most obvious definition is to say a sequence of random variables converges if it converges for all $\omega \in \Omega$. It turns out that, since this definition has nothing to do with probability, this isn't very useful. Instead, we have four different types of convergence.

Definition 12. *Let $X_1, X_2, X_3, ...$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. There are four ways of interpreting $X_n \to X$ as $n \to \infty$*

(1) *$X_n \to X$ **almost surely**, written $X_n \xrightarrow{a.s.} X$, if $\{\omega \in \Omega : X_n(\omega) \to X(\omega)$ as $n \to \infty\}$ is an event whose probability is 1.*

(2) *$X_n \to X$ **in rth mean**, where $r \geq 1$, written $X_n \xrightarrow{r} X$ if $\mathbb{E}|X_n^r| \leq \infty$ for all n and*

$$\mathbb{E}(|X_n - X|^r) \to 0$$

as $n \to \infty$

(3) *$X_n \to X$ **in probability**, written $X_n \xrightarrow{P} X$ if for all $\epsilon > 0$*

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0$$

as $n \to \infty$

(4) *$X_n \to X$ **in distribution**, written $X_n \xrightarrow{D} X$ if*

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$$

as $n \to \infty$ for all points x at which the distribution function $F_X(x) = \mathbb{P}(X \leq x)$ is continuous.

That's a lot to handle, and we will only be working with parts (1) and (4) of this definition. These types of convergence are closely related, and the following implications hold in general:

Proposition 2.

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$$

$$X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$$

In general, the reverse implications do not hold.

The following example shows that convergence in distribution is the weakest type of convergence.

Example 5. *A fair coin shows heads or tails with probability 1/2 for each. Let $X = I_h$ the indicator function for heads and $Y = I_t$ the indicator function for tails. Now define a sequence $X_n = X$ for all n. $X_n \xrightarrow{d} X$ clearly, but $X_n \xrightarrow{d} Y$ also, since X and Y have the same distribution function. However, $|X_n - Y| = 1$ for all n so X_n cannot converge to Y in any other type of convergence.*

We will be especially interested in one particular type of sequence of random variables.

Definition 13. *A sequence of random variables X_n is **independent and identically-distributed** if the X_n's share a common distribution function and they are independent*

5. Characteristic Functions

A very useful class of functions for studying sequences of random variables are characteristic functions

Definition 14. *The **characteristic function** ϕ of a random variable X is the function $\phi : \mathbb{R} \to \mathbb{C}$ defined by*

$$\phi(t) = \mathbb{E}(e^{itX}) \tag{5.1}$$

We will often write ϕ_X to specify the characteristic function of X. There are several results which provide ways to manipulate characteristic functions.

Theorem 4. *If X and Y are independent then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$*

Proof. Plugging into the definition gives

$$\phi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX}e^{itY})$$

Now we can apply Euler's formula: $e^{ix} = \cos(x) + i\sin(x)$, as well as theorem's 1-3 and independence to yield the result. □

Lemma 2. *If $a, b \in \mathbb{R}$ and $Y = aX + b$ then $\phi_Y(t) = e^{itb}\phi_X(at)$*

Proof.

$$\phi_Y(t) = \mathbb{E}(e^{it(aX+b)}) = \mathbb{E}(e^{itb}e^{i(at)X}) \tag{5.2}$$

by theorem 2

$$\phi_Y(t) = e^{itb}\mathbb{E}(e^{i(at)X}) = e^{itb}\phi_X(at) \tag{5.3}$$

□

We will now present several important results without proof

Lemma 3. *If* $\mathbb{E}|X^k| < \infty$ *then*

$$\phi(t) = \sum_{j=0}^{k} \frac{\mathbb{E}(X^j)}{j!}(it)^j + o(t^k) \tag{5.4}$$

Taylor's theorem for a complex variable is the key ingredient in this proof (explaining some of the aesthetic similarities). The next two results show the close relationship between characteristic functions and distribution functions. These results will form the foundation for our proof of the law of large numbers.

Proposition 3. *Random variables X and Y have the same characteristic function if and only if they have the same distribution function.*

Theorem 5. *Suppose that* $F_1, F_2, ...$ *is a sequence of distribution functions with corresponding characteristic functions* $\phi_1, \phi_2, ...$

(1) *If* $F_n \to F$ *for some distribution function F with characteristic function* ϕ*, then* $\phi_n(t) \to \phi(t)$ *for all t*
(2) *Conversely, if* $\phi(t) = \lim_{n\to\infty} \phi_n(t)$ *exists and is continuous at* $t = 0$*, then* ϕ *is the characteristic function of some distribution function F, and* $F_n \to F$

6. Laws of Large Numbers

We are now ready to prove the weak law of large numbers

Theorem 6. *(Weak law of large numbers) Let* $X_1, X_2, ...$ *be a sequence of independent identically distributed random variables with finite expectations* μ*. The sequence of partial sums defined by* $S_n := \sum_{i=1}^n X_i$ *satisfies*

$$\frac{1}{n}S_n \xrightarrow{D} \mu \tag{6.1}$$

as $n \to \infty$

Proof. By theorem 5, it will suffice to show that the characteristic functions of $n^{-1}S_n$ converge to the characteristic function of the constant random variable μ. Since the X_i are identically distributed, by proposition 4, we know they have the same characteristic function which we will denote ϕ_X. Let ϕ_n be the characteristic function of $n^{-1}S_n$. By theorems 4 and lemma 2, we have

$$\phi_n(t) = (\phi_X(t/n))^n \tag{6.2}$$

We can also use lemma 3 to show that $\phi_X(t) = 1 + it\mu + o(t)$. Combining these two equations, we obtain

$$\phi_n(t) = (1 + \frac{i\mu t}{n} + o(\frac{t}{n}))^n \tag{6.3}$$

Using the fact that $\lim_{n\to\infty}(1+\frac{a}{n})^n = e^a$ and waving our hands a bit we obtain

$$\phi_n(t) \to e^{it\mu}$$

as $n \to \infty$. However, $e^{it\mu} = \mathbb{E}(e^{it\mu}) = \phi_\mu(t)$, the characteristic function of the constant random variable μ □

There is a much stronger version of the law of large numbers called (unsurprisingly), the strong law of large numbers. It is presented here without proof

Theorem 7. *(Strong law of large numbers) Let $X_1, X_2, ...$ be a sequence of independent identically distributed random variables. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu \tag{6.4}$$

as $n \to \infty$ for some constant μ, if and only if $\mathbb{E}|X_i| < \infty$. In this case $\mu = \mathbb{E}(X_i)$

In particular, by Proposition 2, it is clear that the strong law implies the weak law.

References

[1] Geoffrey Grimmet and David Stirzaker. Probability and Random Processes. Oxford University Press. 2001.

[2] David Williams. Probability with Martingales. Cambridge University Press. 1991.