Probability Paper

Stephen Zhou

February 2024

1 Introduction

When there are only finitely many events, probability is pretty simple: If you can find the probability of any single event occurring, then the probability of a set of events occurring is just the sum of the events in the set. However, this breaks down when there are an infinite number of events. For example, it is intuitively possible to choose an element of $[0, 1]$ with equal probability. However, the probability any specific number is chosen is 0. So how can you choose any element at all? Also, even though $[0, 1]$ is uncountable and the integers $\mathbb{N}$ are countable, it seems impossible to choose an element of $\mathbb{N}$ at random. It turn out that measure theory is just the right tool we need to resolve these paradoxes.

2 Basic Definitions and Theorems

We first need to rigorously define probability.

Definition. *A probability space is a measure space $(\Omega, \mathcal{F}, P)$ where $P(\Omega) = 1$. We call P a* probability measure *and the elements of $\mathcal{F}$* events. *Ω is called a sample space.*

This definition alone allows us to explain both the paradoxes in the previous section. Measure spaces only have *countable* additivity, so it is possible that $P([0, 1]) = 1$ even though $P(x) = 0$ for any $x \in [0, 1]$. For example, consider the Lebesgue measure λ. This is actually the measure that lets us choose any element with equal probability. (i.e. is translation invariant) If there existed a probability measure on $\mathbb{N}$, where $P(x) = \varepsilon > 0$ for any $x \in \mathbb{N}$, then $P(Z) = \sum_{i=0}^{\infty} P(i) = \sum_{i=0}^{\infty} \varepsilon = \infty$, so there exist no way to randomly choose an element from $\mathbb{N}$, or any countable set, with equal probability.

We will now prove a basic theorem about integration:

Theorem 2.1 (Chebyshev's Inequality)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $f > 0$ almost everywhere. Then*

$$P(x \in X : f(x) \geq \varepsilon) \leq \frac{1}{\varepsilon} \int_{\Omega} f \, dP. \tag{1}$$

Proof. Let $g(x) = 0$ when $f(x) < \varepsilon$ and $g(x) = \varepsilon$ when $f(x) \geq \varepsilon$. Since $g(x) \leq f(x)$,

$$\int_\Omega f\, dP \geq \int_\Omega g\, dP = \varepsilon P(x \in X : f(x) \geq \varepsilon). \tag{2}$$

□

This is often used in the form

$$P(x \in X : |f(x)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \int_\Omega f\, dP. \tag{3}$$

To see this, apply Chebyshev on $f(x)^2$ and ε^2. A similar theorem holds for higher powers. We will now prove our first serious theorem in measure theory:

Theorem 2.2 (First Borel-Cantelli Lemma)**.** *Let $X_1, X_2, \ldots$ be a series of events in a probability space $(\Omega, \mathcal{F}, P)$ with*

$$\sum_{i=0}^{n} P(X_i) < \infty. \tag{4}$$

Then the probability that infinitely many of the X_i occur is 0.

Proof. The set of all $x \in \Omega$ such $x \in X_i$ for infinitely many i is

$$X = \bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} X_i. \tag{5}$$

We get that

$$\mu(X) \leq \mu(\bigcup_{i=j}^{\infty} X_i) \leq \sum_{i=j}^{\infty} \mu(X_i) \to 0 \tag{6}$$

as $j \to \infty$. □

This has a converse if we assume that the X_i are independent.

Definition. *A set of events S is said to be* independent *if for any $E_1, E_2 \in S$, $P(E_1 \cap E_2) = P(E_1)P(E_2)$.*

Theorem 2.3 (Second Borel-Cantelli Lemma)**.** *Let X_i be independent events such that*

$$\sum_{i=1}^{\infty} P(X_i) = \infty. \tag{7}$$

Then it is almost certain that infinitely many of the X_i will occur.

Proof. It suffices to prove that

$$P\left(\bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} X_i\right)^C = 0 \tag{8}$$

This is equivalent to

$$P\left(\bigcup_{j=1}^{\infty}\bigcap_{i=j}^{\infty}X_i^C\right) \leq \sum_{j=1}^{\infty}\prod_{i=j}^{\infty}1-P(X_i) = \sum_{j=1}^{\infty}0 = 0. \tag{9}$$

□

3 Random Walks On Lattices

Here is an interesting application of these lemmas:

Theorem 3.1. *A random walk along a d dimensional lattice starting at the origin that moves to adjacent lattice points with equal probability almost surely returns infinitely many times if $d \leq 2$, and almost never if $d > 2$.*

Proof. Let $E_{d,2n}$ be the probability that a random walk in d dimensions returns to 0 after $2n$ moves. Then $P(E_{d,2n}) = \left(\frac{1}{2^n}\binom{2n}{n}\right)^d$. We need to find when

$$\sum_{i=1}^{n}\left(\frac{1}{4^i}\binom{2i}{i}\right)^d \tag{10}$$

diverges. We will need to take *Stirling's approximation*

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{11}$$

for granted. (A proof can be found here [1]) This gives that

$$\frac{1}{4^i}\binom{2i}{i} \sim \left(\frac{1}{\sqrt{\pi i}}\right)^d. \tag{12}$$

Since

$$\sum_{i=1}^{n}\frac{1}{\sqrt{\pi i}^d} \tag{13}$$

diverges for $d \leq 2$,

$$\sum_{i=1}^{n}\frac{1}{4^i}\binom{2i}{i} = \infty, \tag{14}$$

and the second Borel-Cantelli lemma proves that the origin is almost certainly returned to infinitely many times.

If $d > 2$, then

$$\sum_{i=1}^{n}\frac{1}{4^i}\binom{2i}{i} < \infty, \tag{15}$$

so the first Borel-Cantelli lemma proves that the origin is almost certainly not returned to infinitely many times.

□

4 The Product Measure and Fubini's Theorem

Everything we do in this section and the next also applies to general measure spaces. Imagine performing two experiments, where you want the result A from the first experiment and B from the second. What is the probability you get the desired result from both experiments? Intuitively, the answer is $P(A)P(B)$. So it makes sense to say that $P(A \times B) = P(A)P(B)$. Does this give us a way to define a σ-algebra on the product of two sample spaces? It turns out that it does, and that this measure is unique.

Let $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ be probability spaces. Obviously, the sample space should be $\Omega_1 \times \Omega_2$. It's tempting to let the collection of measurable sets be $\mathcal{F}_1 \times \mathcal{F}_2$, the set of all *measurable rectangles*, but this is not necessarily a σ-algebra. So we'll do the next best thing and let the measurable sets be those in $\mathcal{F}_1 \bigotimes \mathcal{F}_2$, the smallest σ-algebra containing $\mathcal{F}_1 \times \mathcal{F}_2$. We want $(P_1 \bigotimes P_2)(A \times B) = P_3(A \times B) = P_1(A)P_2(B)$ for $A \in \mathcal{F}_1, B \in \mathcal{F}_2$. The measure P_3 is called a product measure.

Example 4.1. The Lebesgue measure on $\mathbb{R}^2$ is the product measure $\lambda \times \lambda$.

Recall that the Lebesgue measure is determined uniquely by its values on the intervals $[a, b]$. Is $(P_1 \times P_2)$ always uniquely determined by its values on measurable rectangles? This turns out to be true, at least for finite measure spaces.

Theorem 4.1. *With notation as before, the measure P_3 exists and is unique.*

This theorem is a special case of Caratheodory's extension theorem, which generalizes the construction of the Lebesgue measure from its value on intervals. We won't prove this here, since that would take us too far off track, but the proof is basically the same as the construction of the Lebesgue measure.

Recall Fubini's theorem on $\mathbb{R}^2$, which allows us to write a double integral as a iterated integral. Is the same thing true for any product measure?

Theorem 4.2 (Fubini's Theorem)**.** *With notation as before,*

$$\int_{A \times B} f(x, y)\, dP_3 = \int_A \int_B f(x, y)\, dP_2\, dP_1 \tag{16}$$

if

$$\int_{A \times B} |f(x, y)|\, dP_3 < \infty. \tag{17}$$

Proof. Notice that the theorem obviously holds when $f = \mathbb{1}_S$ for a measurable rectangle S. Take it for granted that this holds for all P_3 measurable sets S, not just rectangles. (This is actually one way to define the product measure).

Thus this holds for simple f. If we assume f is positive, we can find a sequence of simple functions $0 \leq f_1 \leq f_2 \leq \dots$ such that $\lim_{n \to \infty} f_n = f$. Since the f_i are simple,

$\lim_{n\to\infty} \int_{A\times B} f_n(x,y)\,dP_3 = \lim_{n\to\infty} \int_A \int_B f_n(x,y)\,dP_2\,dP_1$.(18) By Monotone convergence,

$$\int_{A\times B} \lim_{n\to\infty} f_n(x,y)\,dP_3 = \int_A \int_B \lim_{n\to\infty} f_n(x,y)\,dP_2\,dP_1. \tag{19}$$

$\lim_{n\to\infty} f_n = f$, so we are done.

For general f, we write $f = f_+ - f_-$, where f_+, f_- are the positive and negative parts of f respectively and get the theorem as long as neither $\int_{A\times B} f_+$ or $\int_{A\times B} f_-$ is ∞, that is, $\int_{A\times B} |f| < \infty$. □

This also proves Tonelli's Theorem:

Theorem 4.3.

$$\int_{A\times B} f(x,y)\,dP_3 = \int_A \int_B f(x,y)\,dP_2\,dP_1 \tag{20}$$

if f is positive.

5 Random Variables

We also would like to define a random real variable, that is, a variable that takes on real values according to some distribution. It makes more sense to define these as a function $X : \Omega \to \mathbb{R}$ than actual variables, since there is no randomness in variables. Given a Lebesgue measurable set $A \in 2^{\mathbb{R}}$, we would like it if the probability X is in A is defined, or equivalently that $f^{-1}(A) \in \mathcal{F}$ for all measurable sets A. This motivates the following definition:

Definition. *A random variable is a measurable function $X : \Omega \to \mathbb{R}$.*

This actually defines a probability measure on $\mathbb{R}$:

Definition. *A image (or pushforward) measure on $\mathbb{R}$ is a measure obtained from a measure space $(\Omega, \mathcal{F}, P)$ and a random variable X defined as $P_X(A) = P(X^{-1}(A))$.*

It is easy to check that this is a probability measure from the properties of preimages. The image measure can be thought of as the probability that $f(x) \in B$.

Recall that the intervals $[-\infty, b]$ generate the Lebesgue σ-algebra. Thus the values of $P_X([-\infty, b])$ determine the image measure P_X entirely.

Definition. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. The distribution function (or CDF) of a random variable $X : \Omega \to \mathbb{R}$ is defined as $F(x) = P_X([-\infty, b])$.*

If two random variables have the same distribution, they are basically the same, since we don't care about what the events themselves are, only their probabilities.

The next thing we would like to do is define the expected value $\mathbb{E}[X]$ of a random variable X on a probability space $(\Omega, \mathcal{F}, P)$. An obvious definition would be

$$E[X] = \int_\Omega X \, dP \tag{21}$$

(Remember that X is a measurable function $\Omega \to \mathbb{R}$). Another sensible definition is

$$E[X] = \int_{\mathbb{R}} x f(x) \, dx, \tag{22}$$

where $f = F'$.

These two turn out to be equivalent because of a general theorem.

Theorem 5.1 (Change of Variables)**.** *If either side is defined,*

$$\int_{\mathbb{R}} f \, dP_X = \int_\Omega f \circ X \, dP \tag{23}$$

Proof. For indicator functions, this is just the definition of P_X. Thus this holds for all simple functions. There exists a sequence of simple functions $\{f_i\}$ such that $f_n \to f$ as $n \to \infty$ and $f_i \leq f$ for all i. Then

$$\lim_{n\to\infty} \int_{\mathbb{R}} f_n \, dP_X = \lim_{n\to\infty} \int_\Omega f_n \circ X \, dP \tag{24}$$

Dominated convergence tells us that we can swap the limit and integral, giving us the formula. □

By Theorem 4.1,

$$\int_\Omega X \, dP = \int_\Omega I \circ X \, dP = \int_{\mathbb{R}} x \, dP_X = \int_{\mathbb{R}} x f(x) \, dx, \tag{25}$$

the definitions of $E(X)$ are equivalent.

Given the mean $E(X)$ of a variable, we define it's variance as

$$\sigma^2 = \int_\Omega (X - E(X))^2 \, dP = \int_R (x - E(X))^2 \, dP_X. \tag{26}$$

The standard deviation σ is the square root of the variance.

We can now rewrite Chebyshev's inequality

$$P(x \in X : |f(x)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \int_X f(x)^2 dP \tag{27}$$

as

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}. \tag{28}$$

6 The Laws of Large Numbers

We first need define the sum of random variables.

Definition. *The sum of* 2 *random variables* X_1 *and* X_2 *on probability spaces* $(\Omega_1, \mathcal{F}_1, P_1)$ *and* $(\Omega_2, \mathcal{F}_2, P_2)$ *is the random variable* $X+Y$ *over* $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \bigotimes \mathcal{F}_2, P_1 \times P_2)$.

It should be obvious that means are additive.

Theorem 6.1. *Let* X *and* Y *be random variables as before. Then*

$$E(X+Y) = E(X) + E(Y) \tag{29}$$

Proof.

$$\int_{\Omega_1 \times \Omega_2} X+Y \, dP_1 \times P_2 = \int_{\Omega_1} \int_{\Omega_2} X+Y \, dP_2 \, dP_1 = \int_{\Omega_1} E(X)+X \, dP_1 = E(X)+E(Y). \tag{30}$$

□

It turns out that variances are additive too:

Theorem 6.2. *Let* X *and* Y *be random variables. Then*

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) \tag{31}$$

Proof. We can assume that $E(X) = E(Y) = 0$ by considering $X - E(X), Y - E(Y)$ instead.

$$\sigma^2(X+Y) = \int_{\Omega_1 \times \Omega_2} (X+Y)^2 \, dP_1 \times P_2 = \int_{\Omega_1 \times \Omega_2} X^2+2XY+Y^2 \, dP_1 \times P_2 \tag{32}$$

or

$$\sigma^2(X)+\sigma^2(Y)+2\int_{\Omega_1 \times \Omega_2} XY = \sigma^2(X)+\sigma^2(Y)+2E(X)E(Y) = \sigma^2(X)+\sigma^2(Y) \tag{33}$$

□

If we repeat an experiment many times, we should expect the mean of the results to converge (in some way) to the expected value. The laws of large numbers formalize this notion.

Recall the main types of convergence:

Definition. *A sequence* $\{X_n\}$ *of variables in a probability space* $(\Omega, \mathcal{F}, P)$ *is said to converge in probability to* X *if, for all* $\varepsilon > 0$,

$$\lim_{n \to \infty} P(\omega : |X_n - X| > \varepsilon) = 0. \tag{34}$$

(This is called convergence in measure in analysis.)

Definition. *A sequence $\{X_n\}$ of variables in a probability space $(\Omega, \mathcal{F}, P)$ is said to converge almost certainly to X if*

$$\lim_{n\to\infty} P(X_n \neq X) = 0 \tag{35}$$

(This is called convergence almost everywhere in analysis.)

Convergence almost certainly is stronger than convergence in probability, so the strong and weak laws give conditions for each type, respectively. There are several different hypotheses for both laws, but we will only prove one form of each.

Theorem 6.3 (Weak Law of large numbers)**.** *Let $\{X_i\}$ are independent random variables in L^2 with mean 0 and variance σ^2 satisfying*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2(X_i) = 0. \tag{36}$$

then

$$\frac{1}{n} \sum_{i=1}^{n} X_n \to 0 \tag{37}$$

as $n \to \infty$ in probability.

Proof. i) By Chebyshev's inequality,

$$P\left(|\frac{1}{n} \sum_{i=1}^{n} X_i| > \varepsilon \right) \leq \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^{n} \sigma^2(i) \tag{38}$$

for all $\varepsilon > 0$. As $n \to \infty$ the right side approaches 0 by the condition. □

Note that the condition holds when the $\{X_i\}$ are independent identically distributed variables (or i.i.d).

Theorem 6.4 (Strong law of large numbers)**.** *Let $\{X_n\}$ be i.i.d random variables with mean $\mu < \infty$. Then*

$$\frac{1}{n} \sum_{i=1}^{n} X_n \to 0 \tag{39}$$

as $n \to \infty$ almost surely.

The proof given in class is probably the cleanest one that works without extra conditions, so we will give another proof that assumes $E[X^4]$ is finite.

Assuming $E[X^4 \ is finite.] We can assume that \mu = 0.$ Then

$$E\left[\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^4 \right] = \frac{1}{n^4} \sum_{a=1}^{n} \sum_{b=1}^{n} \sum_{c=1}^{n} \sum_{d=1}^{n} E[X_a X_b X_c X_d]. \tag{40}$$

Any of the terms $E[X_a X_b X_c X_d]$ evaluates to 0 if one of $\{a, b, c, d\}$ is different from the others, since each X_i is in a different variable from all the other X_j. Thus we are looking for

$$\frac{nE[X^4] + 3n(n-1)E(X_1^2 X_2^2)}{n^4} \tag{41}$$

since the terms are i.i.d.

Since $x^2 - 2xy + y^2 = (x-y)^2 > 0$, $xy < \frac{x^2+y^2}{2}$. Thus

$$E[X_1^2 X_2^2] < E[X_1^4] \tag{42}$$

and

$$\frac{nE[X^4] + 3n(n-1)E(X_1^2 X_2^2)}{n^4} < \frac{(3n(n-1)+n)E[X_1^4]}{n^4} \tag{43}$$

and

$$\frac{(3n(n-1)+n)E[X_1^4]}{n^4} < \frac{3E[X_1^4]}{n^2}. \tag{44}$$

$$\sum_{i=1}^{n} \frac{3E[X_1^4]}{n^2} \tag{45}$$

is convergent, so its terms must converge to 0. Thus

$$E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^4\right] \tag{46}$$

almost surely converges to 0, and taking fourth roots gives us the theorem. □

References

[1] Tom Carter. *Stirlings Approximation (To n!)* URL: `https://csustan.csustan.edu/~tom/LectureNotes/Stirling/stirling-better.pdf`.