

1 Probability

Remark 1.1. To build a mathematical model of uncertainty and randomness we start by thinking what it is we will be measuring. We want to be able to measure the likelihood of some event happening. So we will measure events. What's an event though? It is an occurrence of something and this something is comprised of smaller events that the event we are concerned with is comprised of. These smaller events we will call outcomes, or singleton event. So for example, lets say I want to know the probability that my plane crashes. The event of a plane crash is composed of infinitely many outcomes... the event maybe consists of an outcome where a mechanic overlooked something, then something wobbled in just the right way, then the plane took a certain turn which caused some screw to loosen, ..., then the plane crashed. So the screw loosening is one outcome, the plane turning is another, etc. all these outcomes make up the event where a plane crashes. So we will consider sets of outcomes, we will call these events.

Remark 1.2. The below theorems will show us how to build these events, how to count the number of outcomes in the events.

Theorem 1.1. Let X_1, X_2, \dots, X_n be finite sets with cardinalities $|X_1|, |X_2|, \dots, |X_n|$. If a process consists of making sequential choices such that:

- The first choice is made from X_1 ,
- The second choice is made from X_2 ,
- ...,
- The n th choice is made from X_n ,

where the number of choices at each stage is independent of previous choices, then the total number of ways to complete the process is:

$$|X_1| \cdot |X_2| \cdots |X_n| = \prod_{i=1}^n |X_i|.$$

Theorem 1.2. Let n and k be nonnegative integers with $0 \leq k \leq n$. The number of distinct subsets of size k that a set of size n has is given by the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Theorem 1.3. For any integer $n \geq 0$ and any real or complex numbers a, b ,

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k},$$

Theorem 1.4. The number of ways to arrange n distinct objects in a sequence is

$$P(n) = n! = n(n-1)(n-2) \cdots 2 \cdot 1$$

The number of ways to select and arrange k objects from n distinct objects is

$$P(n, k) = \frac{n!}{(n-k)!}.$$

1.1 Axioms of Probability

Remark 1.3. So how will we define the abstract space we will be working in so that we can effectively measure the likelihood of events? So consider some event A that you want to know the likelihood of. If the event A is possible, then it should also be possible that A^c , meaning, we should be able to measure both of these. So lets call the space \mathcal{A} , we will add sets to \mathcal{A} that we think should be possible to measure if it is to be possible to measure A . So we have $A \in \mathcal{A}$ and $A^c \in \mathcal{A}$. If $A \in \mathcal{A}$ and $B \in \mathcal{A}$ then we should be able to measure $A \cup B$ and $A \cap B$, so we include all intersections and unions of possible events in \mathcal{A} .

Definition 1.1 (Algebra and σ -algebra). Let Ω be an abstract space. Let 2^Ω denote all subsets of Ω . With \mathcal{A} being a subset of 2^Ω . Then \mathcal{A} is an algebra if it satisfies (1), (2), and (3). \mathcal{A} is a σ -algebra if it satisfies (1), (2), and (4).

1. $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$
2. If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.
3. If the finite sequence of events $A_1, A_2, \dots, A_n \in \mathcal{A}$ then $\bigcup_{i=1}^n A_i \in \mathcal{A}$ and $\bigcap_{i=1}^n A_i \in \mathcal{A}$.
4. If the countable sequence of events $A_1, A_2, \dots \in \mathcal{A}$ then $\bigcup_{i=1}^\infty A_i \in \mathcal{A}$ and $\bigcap_{i=1}^\infty A_i \in \mathcal{A}$.

Remark 1.4. If $C \subset 2^\Omega$, then the σ -algebra generated by C , denoted $\sigma(C)$, is the smallest σ -algebra containing C .

We choose b_n to be strictly increasing so that the $(a_n, b_n]$ part of the interval $(a_n, b_n]$ converges to (a, b) . This is what allows us to have that $(a, b) = \bigcup_{n=1}^\infty (a_n, b_n]$.

Theorem 1.5 (Borel σ -algebra). If $\Omega = \mathbb{R}$, the Borel σ -algebra is the σ -algebra generated by open sets (or equivalently closed sets). Then the Borel σ -algebra can be generated by intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$.

Proof. Let C denote all open intervals. Since every open set in \mathbb{R} is the countable union of open intervals, we have $\sigma(C)$ = the Borel σ -algebra of \mathbb{R} . Let D denote all intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$. Let $(a, b) \in C$, and let $(a_n)_{n \geq 1}$ be a sequence of rationals decreasing to a and $(b_n)_{n \geq 1}$ be a sequence of rationals strictly increasing to b . Then

$$(a, b) = \bigcup_{n=1}^\infty (a_n, b_n] = \bigcup_{n=1}^\infty ((-\infty, b_n] \cap (a_n, \infty)) = \bigcup_{n=1}^\infty ((-\infty, b_n] \cap (-\infty, a_n]^c)$$

Since the right most expression is of the form of D and since we have that any element of C is equivalent to an element of D , we have $C \subset \sigma(D)$, hence $\sigma(C) \subset \sigma(D)$. However since $(-\infty, a]$ contains all its limit points, we know each element of D is a closed set, since closed sets are Borel sets, we have that $\sigma(D)$ is contained in the Borel sets \mathcal{B} . Thus we have

$$\mathcal{B} = \sigma(C) \subset \sigma(D) \subset \mathcal{B},$$

and hence $\sigma(D) = \mathcal{B}$. □

Remark 1.5. So the theorem above shows that when our sample space is the real numbers, or any space with the proper topology, we can generate the σ -algebra we define the probability measure on by using intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$. Since $a \in \mathbb{Q}$, we have made the σ -algebra from countable sets. Which is what we needed since before we knew that open sets C could cover any set, we had to show that the countable collection of $(-\infty, a]$ could also cover (and thus measure) any set.

Remark 1.6. For our actual probability measure, we need an event that is guaranteed and an event that is impossible, so we include $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$ since we want $P(\emptyset) = 0$ and $P(\Omega) = 1$. We also would want that if we measure two events A and B where A and B share no outcomes $A \cap B = \emptyset$, then we want that $P(A \cup B) = P(A) + P(B)$.

Definition 1.2 (Probability Measure). A probability measure defined on a σ -algebra \mathcal{A} of Ω is a function $P : \mathcal{A} \rightarrow [0, 1]$ that satisfies

1. $P(\Omega) = 1$
2. For every pairwise disjoint $(A_n \cap A_m = \emptyset \text{ whenever } n \neq m)$ countable sequence $(A_n)_{n \geq 1}$ of elements of \mathcal{A} , we have

$$P\left(\bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty P(A_n).$$

Definition 1.3. If $A \in 2^\Omega$, then the indicator function $1_A(\omega)$ be given by

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

We say $A_n \in \mathcal{A}$ converges to A if $\lim_{n \rightarrow \infty} 1_{A_n}(\omega) = 1_A(\omega) \forall \omega \in \Omega$.

Remark 1.7. So if we can define convergence of a sequence of sets, then we must have some conception of convergene of supremum and infimum. How will we define these? We want the supremum of a set to be

Definition 1.4. If $A_n \in \mathcal{A} \forall n \in \mathbb{N}$ then define

$$\begin{aligned} \limsup_{n \rightarrow \infty} A_n &= \bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m \\ \liminf_{n \rightarrow \infty} A_n &= \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} A_m. \end{aligned}$$

Theorem 1.6. Let P be a probability measure, and let A_n be a sequence of events in \mathcal{A} which converges to A . Then $A \in \mathcal{A}$ and $\lim_{n \rightarrow \infty} P(A_n) = P(A)$.

Proof. □

Theorem 1.7. Let A_1, A_2, \dots, A_n be events, then

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - (P(A \cap B) + P(A \cap C) + P(B \cap C)) + P(A \cap B \cap C)$.
3. $P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i) - \sum_{1 \leq i_1 \leq i_2 \leq n} P(A_{i_1} \cap A_{i_2})$
 $+ \sum_{1 \leq i_1 \leq i_2 \leq i_3 \leq n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3})$
 $- \sum_{1 \leq i_1 < i_2 \leq i_3 < i_4 \leq n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap A_{i_4})$
 $+ \dots + (-1)^{n+1} P(A_1 \cap \dots \cap A_n) = \sum_{k=1}^n (-1)^{k+1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k})$

Exercise 1.1. Let $A_i \in \mathcal{A}$ be a sequence of events. Show that

$$P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

Definition 1.5. Let B be an event in the sample space Ω such that $P(B) > 0$. Then for all events A the conditional probability of A given B is defined as

$$P(A | B) = \frac{P(A \cap B)}{P(B)}.$$

Definition 1.6. A collection of events $(A_i)_{i \in I}$ is an independent collection if for every finite subset J of I , one has

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

If the above condition is satisfied for for the whole collection, we say the collection $(A_i)_{i \in I}$ is mutually independent. Also, if A_i and A_j are independent $\forall i, j$ with $i \neq j$, that is if any two events you pick from the collection $(A_i)_{i \in I}$ are independent, then the collection is pariwise independent.

Exercise 1.2. If A and B are independent, so also are A and B^c , A^c and B , and A^c and B^c .

Proposition 1.1. If $A_1, A_2, \dots, A_n \in \mathcal{A}$ and if $P(A_1 \cap \dots \cap A_{n-1}) > 0$, then

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2) \dots P(A_n | A_1 \cap \dots \cap A_{n-1}).$$

Definition 1.7. A countable collection of events B_1, \dots, B_n are a partition of Ω if the sets B_i are pairwise disjoint and together they make up Ω . That is, for all i and j , $B_i \cap B_j = \emptyset$ whenever $i \neq j$ and $\bigcup_{i=1}^n B_i = \Omega$

Proposition 1.2. Suppose that B_1, \dots, B_n is a partition of Ω with $P(B_i) > 0$ for $i = 1, \dots, n$. Then for any event A we have

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A | B_i)P(B_i).$$

Theorem 1.8. Let B_1, B_2, \dots, B_n be a partition of the sample space Ω such that each $P(B_i) > 0$. Then for any event A with $P(A) > 0$, and for any $k = 1, \dots, n$, we have:

$$P(B_k | A) = \frac{P(AB_k)}{P(A)} = \frac{P(A | B_k)P(B_k)}{\sum_{i=1}^n P(A | B_i)P(B_i)}.$$

Definition 1.8. Let A_1, A_2, \dots, A_n and B be events with $P(B) > 0$. Then A_1, A_2, \dots, A_n are conditionally independent, given B , if the following condition holds:

For any $k \in \{2, \dots, n\}$ and indices $1 \leq i_1 < i_2 < \dots < i_k \leq n$,

$$P(A_{i_1}A_{i_2}\dots A_{i_k} | B) = P(A_{i_1} | B)P(A_{i_2} | B)\dots P(A_{i_k} | B).$$

Theorem 1.9. (a) A probability on the countable set Ω is characterized by its values on the atoms:

$$p_\omega = P(\omega), \quad \omega \in \Omega.$$

(b) Let $(p_\omega)_{\omega \in \Omega}$ be a family of real numbers indexed by Ω . Then there exists a unique probability P such that $P(\omega) = p_\omega$ if and only if $p_\omega \geq 0$ and $\sum_{\omega \in \Omega} p_\omega = 1$

Definition 1.9. A random variable is a measurable function $X : \Omega \rightarrow \mathbb{R}$ such that for all Borel measurable sets $B \subseteq \mathbb{R}$, the preimage of B is an event in \mathcal{A} , that is

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}.$$

This means that X is \mathcal{A} -measurable, ensuring that we can compute probabilities of the form $P(X \in B)$

Remark 1.8. So a random variable inputs events or outcomes and outputs a real number, then the probability measure will assign probabilities in $[0, 1]$ to the values of X . We can then define the distribution of X by

$$P^X(A) = P(\omega \mid X(\omega) \in A) = P(X^{-1}(A)) = P(X \in A)$$

This is completely determined by the following

$$p_j^X = P(X = j) = \sum_{\omega \mid X(\omega)=j} p_\omega \quad \text{and} \quad P_X(A) = \sum_{j \in A} p_j^X$$

Definition 1.10. Let X be a real-valued random variable on a countable space Ω . The expectation of X , denoted $E(X)$, is defined to be

$$E(X) = \sum_{\omega} X(\omega)p_\omega \quad \text{or} \quad \int_{-\infty}^{\infty} X(\omega)p_\omega d\omega??$$

provided this sum converges. Notice that if the random variable is discrete we use the finite sum, if it is continuous, we use the continuous sum.

Definition 1.11. The n th moment of the random variable X is the expectation $E(X^n)$.

$$E(X^n) = \sum_{\omega} X^n(\omega)p_\omega \quad \text{or} \quad \int_{-\infty}^{\infty} X^n(\omega)P(X(\omega))d\omega$$

Theorem 1.10. Let $h : \mathbb{R} \rightarrow [0, \infty)$ be a nonnegative function and let X be a real valued random variable. Then

$$P(\{\omega \mid h(X(\omega)) \geq a\}) \leq \frac{E(h(X))}{a}, \quad \forall a > 0.$$

Corollary 1.1 (Markov's Inequality).

$$P(|X| \geq a) \leq \frac{E(|X|)}{a}$$

Definition 1.12. Let X be a real valued random variable with $X^2 \in \mathcal{L}^1$ where \mathcal{L}^1 is the space of real valued random variables on (Ω, \mathcal{A}, P) . The variance of X is defined to be

$$\sigma^2 = \sigma_X^2 = E((X - E(X))^2) = E(X^2) - (E(X))^2$$

The standard deviation of X , σ_X , is the nonnegative square root of the variance.

Corollary 1.2 (Chebyshev's Inequality). If X^2 is in \mathcal{L}^1 , then for $a > 0$ we have

1. $P(\{|X| \geq a\}) \leq \frac{EX^2}{a^2}$
2. $P(\{|X - E(X)| \geq a\}) \leq \frac{\sigma_X^2}{a^2}$

Definition 1.13 (Binomial Distribution). Let n be a positive integer and $0 \leq p \leq 1$. A random variable X has the binomial distribution with parameters n and p if the possible values of X are $\{0, 1, \dots, n\}$ and the probabilities are

$$P(\{X = k\}) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n.$$

This is denoted $X \sim \text{Bin}(n, p)$.

Definition 1.14 (Geometric Distribution). A random variable X follows a Geometric distribution with parameter p (success probability per trial) if the probability of k independent trials till a success on the k th trial is given by,

$$P(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \dots$$

Definition 1.15 (Hypergeometric Distribution). A hypergeometric random variable represents the number of successes of size n , drawn without replacement from a population of size N that contains K successes. The PMF is given by

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad \max(0, n - (N - K)) \leq k \leq \min(n, K).$$

Definition 1.16 (Poisson Distribution). A Poisson random variable models the number of events occurring in a fixed interval of time or space, under the assumption that events occur independently and at a constant average rate λ . A random variable X follows a Poisson distribution with rate parameter $\lambda > 0$ if

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

Definition 1.17 (Normal Distribution). A random variable X follows a Normal distribution with mean μ and variance σ^2 , written as $X \sim \mathcal{N}(\mu, \sigma^2)$, if its probability density function (PDF) is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$