1. Conditional probability and independence

Definition 1. Let A, B be two events in the same probability space.

(1) If $\mathbf{P}(B) > 0$, we define the *conditional probability of A given B* as

$$\mathbf{P}\left(A \,\middle|\, B\right) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

(2) We say that A and B are *independent* if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. If $\mathbf{P}(B) \neq 0$, then A and B are independent if and only if $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ (and similarly with the roles of A and B reversed). If $\mathbf{P}(B) = 0$, then A and B are necessarily independent since $\mathbf{P}(A \cap B)$ must also be 0.

What do these notions mean intuitively? In real life, we keep updating probabilities based on information that we get. For example, when playing cards, the chance that a randomly chosen card is an ace is $1/13$, but having drawn a card, the probability for the next card may not be the same - if the first card was seen to be an ace, then the chance of the second being an ace falls to $3/51$. This updated probability is called a conditional probability. Independence of two events A and B means that knowing whether or not A occurred does not change the chance of occurrence of B. In other words, the conditional probability of A given B is the same as the unconditional (original) probability of A.

Example 2. Let $\Omega = \{(i,j) : 1 \leq i, j \leq 6\}$ with $p_{(i,j)} = \frac{1}{36}$. This is the probability space corresponding to a throw of two fair dice. Let $A = \{(i,j) : i \text{ is odd}\}$ and $B = \{(i,j) : j \text{ is } 1 \text{ or } 6\}$ and $C = \{(i,j) : i + j = 4\}$. Then, $A \cap B = \{(i,j) : i = 1, 3, \text{ or } 5, \text{ and } j = 1 \text{ or } 6\}$. It is easy to see that

$$\mathbf{P}(A \cap B) = \frac{6}{36} = \frac{1}{6}, \quad \mathbf{P}(A) = \frac{18}{36} = \frac{1}{2}, \quad \mathbf{P}(B) = \frac{12}{36} = \frac{1}{3}.$$

In this case, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ and hence A and B are independent. On the other hand,

$$\mathbf{P}(A \cap C) = \mathbf{P}\{(1,3),(3,1)\} = \frac{1}{18}, \quad \mathbf{P}(C) = \mathbf{P}\{(1,3),(2,2),(3,1)\} = \frac{1}{12}.$$

Thus, $\mathbf{P}(A \cap C) \neq \mathbf{P}(A)\mathbf{P}(C)$ and hence A and C are not independent.

This agrees with the intuitive understanding of independence since A is an event that depends only on the first toss and B is an event that depends only on the second toss. Therefore, A and B ought to be independent. However, C depends on both tosses, and hence cannot be expected to be independent of A. Indeed, it is easy to see that $\mathbf{P}(C \mid A) = \frac{1}{9}$.

Caution: Independence should not be confused with disjointness! If A and B are disjoint, $\mathbf{P}(A \cap B) = 0$ and hence A and B can be independent if and only if one of $\mathbf{P}(A)$ or $\mathbf{P}(B)$ equals 0. Intuitively, if A and B are disjoint, then knowing that A occurred gives us a lot of information about B (that it did not occur!), so independence is not to be expected.

Exercise 3. If A and B are independent events, show that the following pairs of events are also independent.

(1) A and B^c.

(2) A^c and B.

(3) A^c and B^c.

Total probability rule and Bayes' rule: Let $A_1, \dots, A_n$ be pairwise disjoint and mutually exhaustive events in a probability space. Assume $\mathbf{P}(A_i) > 0$ for all i. This means that $A_i \cap A_j = \emptyset$ for any $i \neq j$ and $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$. We also refer to such a collection of events as a partition of the sample space.

Let B be any other event.

(1) (Total probability rule). $\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)$.

(2) (Bayes' rule). Assume that $\mathbf{P}(B) > 0$. Then, for each $k = 1, 2, \dots, n$, we have

$$\mathbf{P}(A_k \mid B) = \frac{\mathbf{P}(A_k)\mathbf{P}(B \mid A_k)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)}.$$

Proof. The proof is merely by following the definition.

(1) The right hand side is equal to

$$\mathbf{P}(A_1)\frac{\mathbf{P}(B \cap A_1)}{\mathbf{P}(A_1)} + \cdots + \mathbf{P}(A_n)\frac{\mathbf{P}(B \cap A_n)}{\mathbf{P}(A_n)} = \mathbf{P}(B \cap A_1) + \cdots + \mathbf{P}(B \cap A_n),$$

which is equal to $\mathbf{P}(B)$ since A_i are pairwise disjoint and exhaustive.

(2) Without loss of generality take $k = 1$. Note that $\mathbf{P}(A_1 \cap B) = \mathbf{P}(B \cap A_1) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1)$. Hence,

$$\begin{aligned}\mathbf{P}(A_1 \mid B) &= \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)},\end{aligned}$$

where we used the total probability rule to get the denominator. ■

Exercise 4. Suppose A_i are events such that $\mathbf{P}(A_1 \cap \cdots \cap A_n) > 0$. Then, show that

$$\mathbf{P}(A_1 \cap \cdots \cap A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1)\mathbf{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbf{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

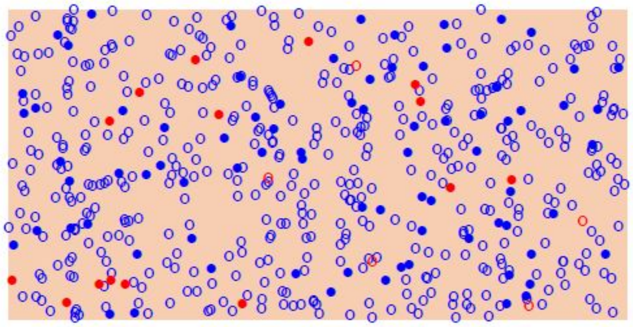

FIGURE 1. A population of healthy (blue) and diseased (red) individuals. Filled circle indicates those who tested positive and hollow circles indicate those who tested negative. The majority of those who tested positive are in fact healthy.

Example 5. Consider a rare disease X that affects one in a million people. A medical test is used to test for the presence of the disease. The test is 99% accurate in the sense that if a person has no disease, the chance that the test shows positive is 1% and if the person has disease, the chance that the test shows negative is also 1%.

Suppose a person is tested for the disease and the test result is positive. What is the chance that the person has the disease X?

Let A be the event that the person has the disease X. Let B be the event that the test shows positive. The given data may be summarized as follows.

(1) $\mathbf{P}(A) = 10^{-6}$. Of course $\mathbf{P}(A^c) = 1 - 10^{-6}$.

(2) $\mathbf{P}(B \mid A) = 0.99$ and $\mathbf{P}(B \mid A^c) = 0.01$.

What we want to find is $\mathbf{P}(A \mid B)$. By Bayes' rule (the relevant partition is $A_1 = A$ and $A_2 = A^c$),

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B \mid A)\mathbf{P}(A) + \mathbf{P}(B \mid A^c)\mathbf{P}(A^c)} = \frac{0.99 \times 10^{-6}}{0.99 \times 10^{-6} + 0.01 \times (1 - 10^{-6})} = 0.000099.$$

The test is quite an accurate one, but the person tested positive has a really low chance of actually having the disease! Of course, one should observe that the chance of having disease is now approximately 10^{-4} which is considerably higher than 10^{-6}.

A calculation-free understanding of this surprising looking phenomenon can be achieved as follows: Let everyone in the population undergo the test. If there are 10^9 people in the population, then there are only 10^3 people with the disease. The number of true positives is approximately $10^3 \times 0.99 \approx 10^3$ while the number of false positives is $(10^9 - 10^3) \times 0.01 \approx 10^7$. In other words, among all positives, the false positives are way more numerous than true positives.

The surprise here comes from not taking into account the relative sizes of the sub-populations with and without the disease. Here is another manifestation of exactly the same fallacious reasoning.

Question: A person X is introverted, very systematic in thinking and somewhat absent-minded. You are told that he is a doctor or a mathematician. What would be your guess - doctor or mathematician?

Most people answer "mathematician". Even accepting the stereotype that a mathematician is more likely to have all these qualities than a doctor, this answer ignores the fact that there are perhaps a hundred times more doctors in the world than mathematicians! In fact, the situation is identical to the one in the example above, and the mistake is in confusing $\mathbf{P}(A \big| B)$ and $\mathbf{P}(B \big| A)$.

Medical diagnosis: Several different physiological problems can give rise to the same symptoms in a person. When a patient goes to a doctor and tells his/her symptoms, the doctor tries to guess the underlying disease that is causing the symptoms. This is Bayes' rule at work (or ought to be at work). Even though one may not be to write down all the probabilities, there is a lesson from the previous examples, which is that a priori chances of different diseases must be taken into account. In other words, suppose a rare but serious lung problem P always causes some symptom X to show. Suppose that common cold Q (rather common) also causes the symptom X in 1% of the cases.

If you are a doctor and you encounter a patient with symptom X, what would be your first guess - that it is caused by P, or by Q? Is such reasoning really used by doctors? It has been observed from various experiments that people do not naturally/intuitively do the right reasoning in such cases - they tend to overestimate the chance of the cause being the rare disease P.

2. Independence of three or more events

Definition 6. Events $A_1, \ldots, A_n$ in a common probability space are said to be independent if

$$\mathbf{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}\right) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \cdots \mathbf{P}(A_{i_m})$$

for every choice of $m \leq n$ and every choice of $1 \leq i_1 < i_2 < \ldots < i_m \leq n$.

The independence of n events requires us to check 2^n equations (that many choices of $i_1, i_2, \ldots$). Should it not suffice to check that each pair of A_i and A_j are independent? The following example shows that this is not the case!

Example 7. Let $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega} \in \Omega$. Define the events $A = \{\underline{\omega} : \omega_1 = 0\}$, $B = \{\underline{\omega} : \omega_2 = 0\}$ and $C = \{\underline{\omega} : \omega_1 + \omega_2 = 0 \text{ or } 2\}$. In words, we toss a fair coin n times and A denotes the event that the first toss is a tail, B denotes the event that the second toss is a tail and C denotes the event that out of the first two tosses are both heads or both tails. Then $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = \frac{1}{4}$. Further,

$$\mathbf{P}(A \cap B) = \frac{1}{4},\ \mathbf{P}(B \cap C) = \frac{1}{4},\ P(A \cap C) = \frac{1}{4},\ \mathbf{P}(A \cap B \cap C) = \frac{1}{4}.$$

Thus, A, B, C are independent *pairwise*, but not independent by our definition because $\mathbf{P}(A \cap B \cap C) \neq \frac{1}{8} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$.

Intuitively this is right. Knowing A does not give any information about C (similarly with A and B, or B and C), but knowing A and B tells us completely whether or not C occurred! Thus it is right that the definition should not declare them to be independent.

Exercise 8. Let $A_1, \ldots, A_n$ be events in a common probability space. Then, $A_1, A_2, \ldots, A_n$ are independent if and only if the following equalities hold: For each i, define B_i as A_i or A_i^c. Then

$$\mathbf{P}(B_1 \cap B_2 \cap \cdots \cap B_n) = \mathbf{P}(B_1)\mathbf{P}(B_2) \cdots \mathbf{P}(B_n).$$

Note: This should hold for any possible choice of B_is. In other words, the system of 2^n equalities in the definition of independence may be replaced by this new set of 2^n equalities. The latter system has the advantage that it immediately tells us that if $A_1, \ldots, A_n$ are independent, then $A_1, A_2^c, A_3, \ldots$ (for each i choose A_i, or its complement) are independent.

3. Discrete probability distributions

Let (Ω, p) be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable. We define two objects associated to X.

Probability mass function (pmf). The range of X is a countable subset of $\mathbb{R}$, denote it by $\text{Range}(X) = \{t_1, t_2, \ldots\}$. Then, define $f_X : \mathbb{R} \to [0,1]$ as the function

$$f_X(t) = \begin{cases} \mathbf{P}\{\omega : X(\omega) = t\} & \text{if } t \in \text{Range}(X), \\ 0 & \text{if } t \notin \text{Range}(X). \end{cases}$$

One obvious property is that $\sum_{t \in \mathbb{R}} f_X(t) = 1$. Conversely, any non-negative function f that is non-zero on a countable set S and such that $\sum_{t \in \mathbb{R}} f(t) = 1$ is a pmf of some random variable.

Cumulative distribution function (CDF). Define $F_X : \mathbb{R} \to [0,1]$ by

$$F_X(t) = \mathbf{P}\{\omega : X(\omega) \le t\} \text{ for } t \in \mathbb{R}.$$

Example 9. Let $\Omega = \{(i,j) : 1 \le i, j \le 6\}$ with $p_{(i,j)} = \frac{1}{36}$ for all $(i,j) \in \Omega$. Let $X : \Omega \to \mathbb{R}$ be the random variable defined by $X(i,j) = i + j$. Then, $\text{Range}(X) = \{2, 3, \ldots, 12\}$. The pmf of X is given by

k	2	3	4	5	6	7	8	9	10	11	12
$f_X(k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and the CDF is given by

t	< 2	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,7)$	$[7,8)$	$[8,9)$	$[9,10)$	$[10,11)$	$[11,12)$	≥ 12
$F_X(t)$	0	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	1

Basic properties of a CDF: The following observations are easy to make:

(1) F is an increasing function on $\mathbb{R}$.

(2) $\lim\limits_{t \to +\infty} F(t) = 1$ and $\lim\limits_{t \to -\infty} F(t) = 0$.

(3) F is right continuous, that is, $\lim\limits_{h \to 0^+} F(t+h) = F(t+) = F(t)$ for all $t \in \mathbb{R}$.

(4) F increases only in jumps. This means that if F has no jump discontinuities (an increasing function has no other kind of discontinuity anyway) in an interval $[a,b]$, then $F(a) = F(b)$.

Since $F(t)$ is the probability of a certain event, these statements can be proved using the basic rules of probability that we saw earlier.

Proof. Let $t < s$. Define two events, $A = \{\omega : X(\omega) \le t\}$ and $B = \{\omega : X(\omega) \le s\}$. Clearly, $A \subseteq B$ and hence $F(t) = \mathbf{P}(A) \le \mathbf{P}(B) = F(s)$. This proves the first property.

To prove the second property, let $A_n = \{\omega : X(\omega) \le n\}$ for $n \ge 1$. Then, A_n are increasing in n and $\bigcup_{n=1}^{\infty} A_n = \Omega$. Hence, $F(n) = \mathbf{P}(A_n) \to \mathbf{P}(\Omega) = 1$ as $n \to \infty$. Since F is increasing, it follows that $\lim_{t\to+\infty} F(t) = 1$. Similarly, one can prove that $\lim_{t\to-\infty} F(t) = 0$.

Right continuity of F is also proved the same way, by considering the events $B_n = \{\omega : X(\omega) \le t + \frac{1}{n}\}$. We omit details. ■

Remark 10. It is easy to see that one can recover the pmf from the CDF and vice versa. For example, given the pmf f, we can write the CDF as $F(t) = \sum_{u:u\le t} f(u)$. Conversely, given the CDF, by looking at the locations of the jumps and the sizes of the jumps, we can recover the pmf.

The point is that *probabilistic questions about X can be answered by knowing its CDF F_X*. Therefore, in a sense, the probability space becomes irrelevant. For example, the expected value of a random variable can be computed using its CDF only. Hence, we shall often make statements like "X is a random variable with pmf f" or "X is a random variable with CDF F", without bothering to indicate the probability space.

Some distributions (i.e., CDF or the associated pmf) occur frequently enough to merit a name.

Example 11. Let f and F be the pmf, CDF pair

$$f_X(t) = \begin{cases} p & \text{if } t = 1, \\ q & \text{if } t = 0, \end{cases} \qquad F_X(t) = \begin{cases} 1 & \text{if } t \ge 1, \\ q & \text{if } t \in [0,1), \\ 0 & \text{if } t < 0. \end{cases}$$

A random variable X having this pmf (or, equivalently the CDF) is said to have *Bernoulli distribution* with parameter p (with $q = 1-p$) and write $X \sim \text{Ber}(p)$. For example, if $\Omega = \{1, 2, \dots, 10\}$ with $p_i = 1/10$, and $X(\omega) = \mathbf{1}_{\omega\le 3}$, then $X \sim \text{Ber}(0.3)$. Any random variable taking only the values 0 and 1, has Bernoulli distribution.

Example 12. Fix $n \ge 1$ and $p \in [0,1]$. The pmf defined by $f(k) = \binom{n}{k} p^k q^{n-k}$ for $0 \le k \le n$ is called the *Binomial distribution* with parameters n and p, and is denoted Bin(n,p). The CDF is as usual defined by $F(t) = \sum_{\mathbf{u}:\mathbf{u}\le t} f(u)$, but it does not have any particularly nice expression.

For example, if $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n-\sum_i \omega_i}$ and $X(\underline{\omega}) = \omega_1 + \cdots + \omega_n$, then $X \sim \text{Bin}(n,p)$. In words, the number of heads in n tosses of a p-coin has Bin(n,p) distribution.

Example 13. Fix $p \in (0,1]$ and let $f(k) = q^{k-1}p$ for $k \in \mathbb{N}_+$. This is called the *Geometric distribution* with parameter p and is denoted Geo(p). The CDF is

$$F(t) = \begin{cases} 0 & \text{if } t < 1, \\ 1 - q^k & \text{if } k \le t < k+1, \text{ for some } k \ge 1. \end{cases}$$

For example, the number of tosses of a p-coin till the first head turns up, is a random variable with Geo(p) distribution.

Example 14. Fix $\lambda > 0$ and define the pmf $f(k) = e^{-\lambda}\frac{\lambda^k}{k!}$. This is called the *Poisson distribution* with parameter λ and is denoted Pois(λ).

Example 15. Fix positive integers b, w and $m \leq b + w$. Define the pmf $f(k) = \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}}$ where the binomial coefficient $\binom{x}{y}$ is interpreted to be zero if $y > x$ (thus $f(k) > 0$ only for $\max\{m - w, 0\} \leq k \leq b$). This is called the *Hypergeometric distribution* with parameters b, w, m and we shall denote it by Hypergeo(b, w, m).
Consider a population with b men and w women. The number of men in a random sample (without replacement) of size m, is a random variable with the Hypergeo(b, w, m) distribution.

4. General probability distributions

We take the first three of the four properties of CDF proved in the previous section as the *definition* of a CDF or distribution function, in general.

Definition 16. A (cumulative) distribution function (or, CDF for short) is any function $F : \mathbb{R} \to [0,1]$ be a non-decreasing, right continuous function such that $F(t) \to 0$ as $t \to -\infty$ and $F(t) \to 1$ as $t \to +\infty$.

If (Ω, p) is a discrete probability space and $X : \Omega \mapsto \mathbb{R}$ is any random variable, then the function $F(t) = \mathbf{P}\{\omega : X(\omega) \leq t\}$ is a CDF, as discussed in the previous section. However, there are distribution functions that do not arise in this manner.

Example 17. Let

$$F(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ t & \text{if } 0 < t < 1, \\ 1 & \text{if } t \geq 1. \end{cases}$$

Then, it is easy to see that F is a distribution function. However, it has no jumps and hence it does not arise as the CDF of any random variable on a discrete probability space.

There are two ways to rectify this issue:

(1) The first way is to learn the notion of uncountable probability spaces, which poses many subtleties. It requires a semester or so of real analysis and measure theory. But, after that one can define random variables on uncountable probability spaces and the above example will turn out to be the CDF of some random variable on some (uncountable) probability space.

(2) Just regard CDFs such as in the above example as reasonable approximations to CDFs of some discrete random variables. For example, if $\Omega = \{\omega_0, \omega_1, \dots, \omega_N\}$ and $p(\omega_k) = 1/(N+1)$ for all $0 \leq k \leq N$, and $X : \Omega \mapsto \mathbb{R}$ is defined by $X(\omega_k) = k/N$, then it is easy to check that the CDF of X is the function G given by

$$G(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ \frac{k}{N+1} & \text{if } \frac{k-1}{N} \leq t < \frac{k}{N} \text{ for some } k = 1, 2, \dots, N, \\ 1 & \text{if } t \geq 1. \end{cases}$$

Now, if N is very large, then the function G looks approximately like the function F. Just as it is convenient to regard water as a continuous medium in some problems (although water is made up of molecules and is discrete at small scales), it is convenient to use the continuous function F as a reasonable approximation to the step function G.

We shall take the second option. Whenever we write continuous distribution functions such as in the above example, at the back of our mind we have a discrete random variable (taking a large number of closely placed values) whose CDF is approximated by our distribution function. The advantage of using continuous objects instead of discrete ones is that the powerful tools of Calculus become available to us.

5. Uncountable probability spaces - conceptual difficulties

The following two "random experiments" are easy to imagine, but difficult to fit into the framework of probability spaces.

(1) Toss a p-coin infinitely many times: Clearly the sample space is $\Omega = \{0,1\}^{\mathbb{N}}$. But what is $p_{\underline{\omega}}$ for any $\underline{\omega} \in \Omega$? The only reasonable answer is $p_{\underline{\omega}} = 0$ for all ω. But then how to define $\mathbf{P}(A)$ for any A? For example, if $A = \{\underline{\omega} : \omega_1 = 0, \omega_2 = 0, \omega_3 = 1\}$, then everyone agrees that $\mathbf{P}(A)$ "ought to be" $q^2 p$, but how does that come about? The basic problem is that Ω is uncountable, and probabilities of events cannot be obtained by summing probabilities of singletons.

(2) Draw a number at random from $[0,1]$: Again, it is clear that $\Omega = [0,1]$, but it also seems reasonable that $p_x = 0$ for all x. Again, Ω is uncountable, and probabilities of events cannot be obtained by summing probabilities of singletons. It is "clear" that if $A = [0.1, 0.4]$, then $\mathbf{P}(A)$ "ought to be" 0.3, but it gets confusing when one tries to derive this from something more basic!

The resolution: Let Ω be uncountable. There is a class of *basic subsets* (usually NOT singletons) of Ω for which we take the probabilities as given. We also take the rules of probability, namely, countable additivity, as axioms. Then, we use the rules to compute the probabilities of more complex events (subsets of Ω) by expressing those events in terms of the basic sets using countable intersections, unions and complements and applying the rules of probability.

Example 18. In the example of infinite sequence of tosses, $\Omega = \{0,1\}^{\mathbb{N}}$. Any set of the form $A = \{\underline{\omega} : \omega_1 = \epsilon_1, \ldots, \omega_k = \epsilon_k\}$, where $k \geq 1$ and $\epsilon_i \in \{0,1\}$ will be called a basic set and its probability is defined to be $\mathbf{P}(A) = \prod_{j=1}^{k} p^{\epsilon_j} q^{1-\epsilon_j}$, where we assume that $p > 0$. Now, consider a more complex event, for example, $B = \{\underline{\omega} : \omega_k = 1 \text{ for some } k\}$. We can write $B = A_1 \cup A_2 \cup A_3 \cup \cdots$, where $A_k = \{\underline{\omega} : \omega_1 = 0, \ldots, \omega_{k-1} = 0, \omega_k = 1\}$. Since A_k are pairwise disjoint, the rules of probability demand that $\mathbf{P}(B)$ should be $\sum_k \mathbf{P}(A_k) = \sum_k q^{k-1} p$ which is in fact equal to 1.

Example 19. In the example of drawing a number at random from $[0,1]$, $\Omega = [0,1]$. Any interval (a,b) with $0 \leq a < b \leq 1$ is called a basic set and its probability is defined as $\mathbf{P}(a,b) = b - a$. Now, consider a non-basic event $B = [a,b]$. We can write $B = A_1 \cup A_2 \cup A_3 \ldots$, where $A_k = (a + (1/k), b - (1/k))$. Then A_k is an increasing sequence of events and the rules of probability say that $\mathbf{P}(B)$ must be equal to $\lim_{k\to\infty} \mathbf{P}(A_k) = \lim_{k\to\infty}(b - a - (2/k)) = b - a$. Another example could be $C = [0.1, 0.2) \cup (0.3, 0.7]$. Similarly, argue that $\mathbf{P}(\{x\}) = 0$ for any $x \in [0,1]$. A more interesting one is $D = \mathbb{Q} \cap [0,1]$. Since it is a countable union of singletons, it must have zero probability! Even more interesting is the $1/3$-Cantor set. Although uncountable, it has zero probability! We discuss an example (related to the Cantor set) later.

Consistency: Is this truly a solution to the question of uncountable spaces? Are we assured of never running into inconsistencies? NOT always!

Example 20. Let $\Omega = [0,1]$ and let intervals (a,b) be open sets with their probabilities defined as $\mathbf{P}(a,b) = \sqrt{b-a}$. This quickly leads to problems. For example, $\mathbf{P}(0,1) = 1$ by definition. But $(0,1) = (0,0.5) \cup (0.5,1) \cup \{1/2\}$ from which the rules of probability would imply that $\mathbf{P}(0,1)$ must be at least $\mathbf{P}(0,1/2) + \mathbf{P}(1/2,1) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$ which is greater than 1. Inconsistency!

Exercise 21. Show that we run into inconsistencies if we define $\mathbf{P}(a,b) = (b-a)^2$ for $0 \le a < b \le 1$.