

CIS 2033 Lectures 2 and 3, Spring 2017¹

Instructor: David Dobor

Updated January 19, 2017

¹ Readings from the textbook:
Chapter 1, Sec 1..?

In the previous class we discussed the the basic rules of probability theory - its axioms. They are surprisingly few, but they imply many interesting properties that we will explore in this class.

Implications of the Axioms

The probability axioms are surprisingly few, but they imply many interesting properties that we will now explore. For example, we have an axiom that says that probabilities are non-negative. We will show that probabilities are also less than or equal to 1. We have another axiom that says that the probability of the entire sample space Ω is 1. We will show a kind of counterpart to this axiom, one that says that the empty set, that is, Ω 's *complement*, has the probability equal to 0. This makes perfect sense. The empty set has no elements, the event corresponding to it is impossible. There is 0 probability that the outcome of the experiment would lie in the empty set.

Some simple consequences of the axioms

Axioms

$$P(A) \geq 0$$

$$P(\Omega) = 1$$

Consequences

$$P(A) \leq 1$$

$$P(\emptyset) = 0$$

We also have another intuitive property. The probability that an event happens plus the probability that the event does not happen exhaust all possibilities. And these two probabilities together should add to 1. For instance, if the probability of heads is 0.6, then the probability of tails should be 0.4.

Finally, we can generalize the additivity axiom, which was originally given for the case of two disjoint events to the case where we're dealing with the union of several disjoint events. By disjoint here we mean that the intersection of any two of these events is the empty set. We will prove this for the case of three events and then the argument generalizes for the case where we're taking the union of k disjoint events, where k is any finite number.

Axioms

$$P(A) \geq 0$$

$$P(\Omega) = 1$$

For disjoint events:

$$P(A \cup B) = P(A) + P(B)$$

Consequences

$$P(A) \leq 1$$

$$P(\emptyset) = 0$$

$$P(A) + P(A^c) = 1$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

So the intuition of this result is the same as for the case of two events. But we will derive it formally and we will also use it to come up with a way of calculating the probability of a finite set by simply adding the probabilities of its individual elements.

All of these statements that we just presented are intuitive. And you do not need to be really convinced about their validity. *Nevertheless*, it is instructive to see how these statements follow from the axioms that we have put in place.

SO WE WILL NOW PRESENT THE ARGUMENTS based only on the three axioms that we have available. And in order to be able to refer to these axioms, let us give them some names, call them axioms (a), (b), and (c):

We start as follows. Let us look at the sample space and a subset of that sample space - call it A - and consider the complement of that subset. The complement is the set of all elements that do not belong to the set A . So a set together with its complement make up everything, which is the entire sample space. On the other hand, if an element belongs to a set A , it does not belong to its complement. So the intersection of a set with its complement is the empty set.

Now we argue as follows. We have that the probability of the entire sample space is equal to 1. This is true by our second axiom. Now the sample space, as we just discussed, can be written as the union of an event and the complement of that event. This is just a set theoretic relation. Next, since a set and its complement are disjoint, this means that we can apply the additivity axiom and write this probability as the sum of the probability of event A with the probability of the complement of A . This is one of the relations that we had claimed and which we have now established.

Based on this relation, we can also write that the probability of an event A is equal to 1 minus the probability of the complement of that event. And because, by the non-negativity axiom the quantity $P(A^c)$ is non-negative, 1 minus something non-negative is less than or equal to 1. Thus, by using the nonnegativity axiom, we have established

Axioms

$$(a) \quad P(A) \geq 0$$

$$(b) \quad P(\Omega) = 1$$

For disjoint events:

$$(c) \quad P(A \cup B) = P(A) + P(B)$$

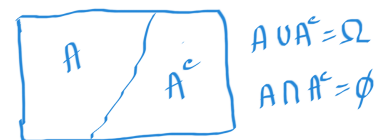


Figure 1: Complementary events make up the entire sample space Ω .

$$\begin{aligned} 1 &\stackrel{(b)}{=} P(\Omega) = P(A \cup A^c) \\ &\stackrel{(c)}{=} P(A) + P(A^c) \end{aligned}$$

$$P(A) = 1 - P(A^c) \stackrel{(a)}{\leq} 1$$

Figure 2: Thus we establish another property - namely, that probabilities are always ≤ 1 .

another property, namely that probabilities are always less than or equal to 1.

Finally, let us note that 1 is the probability, always, of a set plus the probability of a complement of that set. And let us use this property for the case where the set of interest is the entire sample space. Now, the probability of the entire sample space is itself equal to 1. And what is the complement of the entire sample space?

The complement of the entire sample space consists of all elements that do not belong to the sample space. But since the sample space is supposed to contain all possible elements, its complement is just the empty set. And from this relation we get the implication that the probability of the empty set is equal to 0. This establishes yet one more of the properties that we had just claimed a little earlier.

We finally come to the proof of the generalization of our additivity axiom from the case of two disjoint events to the case of three disjoint events. We look at our sample space in Figure 4, and within that sample space we have three events, three subsets. These subsets are disjoint in the sense that any two of those subsets have no elements in common. We're interested in the probability of the union of A , B , and C .

How do we make progress? We have an additivity axiom in our hands, which applies to the case of the union of two disjoint sets. Here we have three of them. But we can do the following trick. We can think of the union of A , B , and C as consisting of the union of the set shaded in blue with that set shaded in green in Figure 5. Formally, what we're doing is that we're expressing the union of these three sets as follows. We form one set by taking the union of A with B . And we have the other set C . And the overall union can be thought of as the union of these two sets.

Now since the three sets are disjoint, this implies that the blue set is disjoint from the green set and so we can use the additivity axiom here to write this probability as the probability of A union B plus the probability of C . And now we can use the additivity axiom once more since the sets A and B are disjoint to write the first term as probability of A plus probability of B . We carry over the last term and we have the relation that we wanted to prove.

$$A, B, C \text{ disjoint: } P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

This actually was the proof for the case of three events. You should be able to follow this line of proof to write an argument for the case of four events and so on. And you might want to continue by induction. Eventually you should be able to prove that if the sets A_1 up to A_k are disjoint then the probability of the union of those sets is going

$$\begin{aligned} 1 &= P(\Omega) + P(\Omega^c) \\ 1 &= 1 + P(\emptyset) \\ \Rightarrow P(\emptyset) &= 0 \end{aligned}$$

Figure 3: Thus we establish yet another property – namely, that the probability of the empty set is 0.

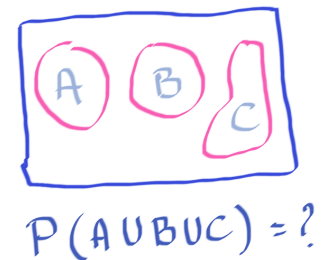


Figure 4: Can we extend our additivity axiom (c) to three events? How about any finite number of disjoint events. Infinite number of disjoint events?

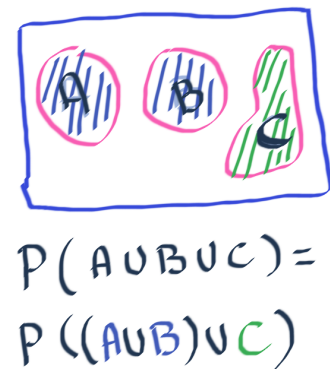


Figure 5: We can think of the union of events A and B as a single event $A \cup B$. We can then apply additivity axiom to $A \cup B$ and C .

to be equal to the sum of their individual probabilities. So this is the generalization to the case where we're dealing with the union of finitely many disjoint events:

$$\text{If } A_1, \dots, A_K \text{ are disjoint} \implies P(A_1 \cup \dots \cup A_K) = \sum_{i=1}^K P(A_i)$$

Check Your Understanding: Let A , B , and C be disjoint subsets of the sample space. For each one of the following statements, determine whether it is true or false. Note: "False" means "not guaranteed to be true."

Please don't read the answers before attempting the questions.

$$P(A) + P(A^c) + P(B) = P(A \cup A^c \cup B) \quad (1)$$

$$P(A) + P(B) \leq 1 \quad (2)$$

$$P(A^c) + P(B) \leq 1 \quad (3)$$

$$P(A \cup B \cup C) \geq P(A \cup B) \quad (4)$$

- (1) False. For a counterexample, let $A = \emptyset$, $B = \Omega$, and $C = \emptyset$. In that case, the left-hand side of the equation equals 2, whereas the right-hand side equals 1.
- (2) True. Since A and B are disjoint, we have $P(A) + P(B) = P(A \cup B) \leq 1$.
- (3) False. For a counterexample, let $A = \emptyset$, $B = \Omega$, and $C = \emptyset$. In that case, $P(A^c) + P(B) = 2$.
- (4) True. Since A , B , and C are disjoint, we have $P(A \cup B \cup C) = P(A) + P(B) + P(C) \geq P(A) + P(B) = P(A \cup B)$.

ANSWERS:

More Consequences of Axioms

EXAMPLE 1. We will now continue and derive some additional properties of probability laws which are, again, consequences of the axioms that we have introduced. The first property is the following.

- If $A \subset B$, then $P(A) \leq P(B)$

We could argue why this is true as follows. If we have two sets and one set is smaller than the other, we have a picture similar to the one shown on the right. We have our sample space. And we have some set, A . We also have this set B , shown in figure 6, which is even bigger. So if B is a set which is larger than A , then, naturally, the probability that the outcome falls inside B should be at least as big as the probability that the outcome falls inside A .

How do we prove this formally? The set B can be expressed as a union of two pieces. One piece is the set A itself. The second piece is whatever elements of B there are that do not belong to A , which means that they belong to the complement of A .

So we have expressed the set B as the union of two pieces. One piece is A . The other piece is outside A - it is the one shaded in green in figure 6 and is that part of B that does not include anything that belongs to A also. Or, in set notation, $B \cap A^c$. This is sometimes also denoted as $B \setminus A$. Either way, these two pieces are disjoint. And so we can apply the additivity axiom, and write that the probability of B is equal to the probability of A plus the probability of the other set.

Here is the exact same argument expressed more compactly using a mathematical notation. First write set B as follows:

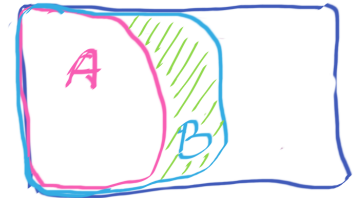
$$B = A \cup (B \cap A^c)$$

The above line expresses B as the union of two non-intersecting sets. So we can apply the additivity axiom, our axiom (c):

$$P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c) \geq P(A)$$

Since probabilities are non-negative, that is, since $P(B \cap A^c) \geq 0$, then indeed we conclude that the probability of A is less than or equal to the probability of B . \square

EXAMPLE 2. The next property we will show allows us to write the probability of the union of two sets for the case now, where the two sets are not necessarily disjoint.



$$A, B, \quad B \setminus A = B \cap A^c$$

Figure 6: Here A is fully contained in B . The shaded part in green is that part of B that does not contain any elements of A . This is denoted by $B \setminus A$. Another way to write this set: $B \cap A^c$.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

So the picture is as follows. We have our two sets, A and B. These sets are not necessarily disjoint. And we want to say something about the probability of the union of A and B.

Can you prove that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$? Use the picture shown on the right and try to use either the mathematical notation or argue it verbally, as we did in example 1.

Here is a hint if you get stuck: $A \cup B$ can be expressed as the union of non intersecting sets $A \cup B^c$, $A \cap B$ and $B \cup A^c$. Denote the probabilities of these sets by a, b, c , respectively. What is then $P(A \cup B)$ in terms of a, b and c ? It is $a + b + c$. On the other hand, can you express $P(A) + P(B) - P(A \cap B)$ in terms of a, b, c as well? If the answer to this is also $a + b + c$, then you've got your proof. *Further hint:* note that $P(A) = a + b$ and $P(B) = b + c$. \square

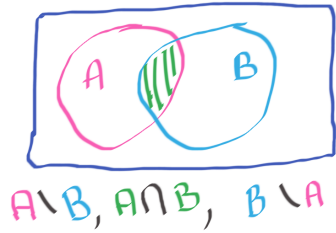


Figure 7: Here A and B do intersect. But the union of A and B can be represented as the union of the three non-intersecting sets $A \setminus B$, $A \cap B$ and $B \setminus A$.

EXAMPLE 3. Here's yet another consequence of the three axioms:

- $P(A \cup B) \leq P(A) + P(B)$

Can you see why this is true? This should follow right from the previous example. Recall that the probabilities are non-negative and thus, looking at what we proved in example 2, since $P(A \cap B) \geq 0$, it must follow that $P(A \cup B) \geq P(A) + P(B)$.

In fact, this inequality even has a name. It's called the *union bound inequality*. \square

Check Your Understanding: Let A , B , and C be subsets of the sample space, not necessarily disjoint. For each one of the following statements, determine whether it is true or false. Note: "False" means "not guaranteed to be true."

Please don't read the answers before attempting the questions.

$$P((A \cap B) \cup (C \cap A^c)) \leq P(A \cup B \cup C) \quad (1)$$

$$P(A \cup B \cup C) = P(A \cap C^c) + P(C) + P(B \cap A^c \cap C^c) \quad (2)$$

- (1) True. This is because the set $(A \cap B) \cup (C \cap A^c)$ is a subset of $A \cup B \cup C$.
 (2) True. This is the same property shown in the last segment, with the three sets appearing in a different order.

Answers:

We now move from the abstract to the concrete :)

Recall the example that we discussed earlier where we have two rolls of a tetrahedral die. So there are 16 possible outcomes shown in figure 8. To continue, we need to specify a probability law, some kind of probability assignment.

To keep things simple, we're going to make the assumption that the 16 possible outcomes are all equally likely. And each outcome has a probability of 1 over 16. Given this assumption, we will now proceed to calculate certain probabilities.

Let us look first at the probability that X , which stands for "the result of the first roll", is equal to 1. The way to calculate this probability is to identify what exactly that event is in our picture of the sample space, and then calculate. The event that $X = 1$ can happen in four different ways that correspond to these four particular outcomes.

Each one of these outcomes has a probability of 1 over 16. The probability of this event is the sum of the probabilities of the outcomes that it contains. So it is $4 \times \frac{1}{16} = \frac{1}{4}$.

Let now Z stand for the smaller of the two numbers that came up in our two rolls. So for example, if X is 2 and Y is equal to 3, then Z is equal to 2, which is the smaller of the two.

Let us try to calculate the probability that the smaller of the two outcomes is equal to 4.

Now, for the smaller of the two outcomes to be equal to 4, we must have that both X and Y are equal to 4. So this top rightmost square in figure 10 is the only way that this particular event can happen. Since there's only one outcome that makes the event happen, the probability of this event is the probability of that outcome and is equal to $\frac{1}{16}$.

	1	2	3	4
4				
3				
2				
1				

Figure 8: The tetrahedral die we saw earlier. Let's now *make an assumption* that the 16 possible outcomes are all equally likely.

	1	2	3	4
4	✓			
3	✓			
2	✓			
1	✓			

Figure 9: The checkmarked boxes correspond to the two rolls of the die that resulted in the first roll showing face 1. The probability of this event works out to be $\frac{1}{4}$.

	1	2	3	4
4				✓
3				
2				
1				

Figure 10: The checkmark now corresponds to the event that the smaller of the two faces of the die came up equal to 4. There is only one such outcome – namely, when both rolls result in a 4. Its probability is then $\frac{1}{16}$.

For another example, let's calculate the probability that the minimum is equal to 2. What does it mean that the minimum is equal to 2? It means that one of the dice resulted in a 2, and the other die resulted in a number that's 2 or larger. So we could have both equal to 2. We could have X equal to 2, but Y larger. Or we could have Y equal to 2 and X something larger.

The checkmarked boxes on this figure to the right correspond to all outcomes for which the minimum of the two rolls is equal to 2. There's a total of five such outcomes. Each one of them has probability 1 over 16 . And we have discussed that for finite sets, the probability of a finite set is the sum of the probabilities of the elements of that set. So we have five elements here, each one with probability 1 over 16 . And the answer to this problem is then $\frac{5}{16}$.

Y = Second roll	4		✓		
	3		✓		
	2		✓	✓	✓
	1				
		1	2	3	4
		X = First roll			

Figure 11: The checkmark now corresponds to the event that the smaller of the two faces of the die came up equal to 2. There are 5 such outcomes – namely, the ones checkmarked in this figure. The probability of this event is $\frac{5}{16}$.

THIS PARTICULAR EXAMPLE that we just saw is a special case of what is called a discrete uniform law. In a discrete uniform law, we have a sample space which is finite. And it has n elements, and we assume that these n elements are *equally likely*. Now since the probability of the entire sample space Ω is equal to 1 , this means that each one of these elements must have probability $1/n$. That's the only way that the sum of the probabilities of the different outcomes would be equal to 1 as required by the normalization axiom.

Consider now some subset of the sample space, an event A that, exactly k elements. What is the probability of the set A ? It's the sum of the probabilities of its elements. There are k elements. And each one of them has a probability of 1 over n . And this way we can find the probability of the set A . It is $k \times \frac{1}{n}$.

So when we have a discrete uniform probability law, we can calculate probabilities by simply counting the number of elements of Ω , which is n , finding the number n , and counting the number of elements of the set A . *That's the reason why counting will turn out to be an important skill. And there will be a whole lecture devoted to this particular topic.*

Check Your Understanding: Consider the same model of two rolls of a tetrahedral die, with all 16 outcomes equally likely. Find the probability of the following events:

1. The value in the first roll is strictly larger than the value in the second roll.
2. The sum of the values obtained in the two rolls is an even number.

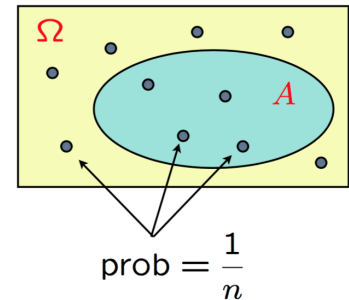


Figure 12: **Discrete Uniform Law.** Here Ω has finite number of elements, n , and each is equally likely to occur, that is, each has probability of $\frac{1}{n}$ of occurring. Set A contains k elements. Then $P(A) = k \cdot \frac{1}{n} = \frac{k}{n}$.

Please don't read the answers before attempting the questions.

ANSWERS:

(1) The event of interest is $\{(2,1), (3,1), (4,1), (2,2), (3,2), (4,2), (2,3), (3,3), (4,3)\}$. It consists of 6 elements (outcomes), each of which has probability $1/16$, for an overall probability of $6/16 = 3/8$.

(2) The event of interest is $\{(1,1), (2,2), (3,3), (4,4), (1,3), (3,1), (2,4), (4,2)\}$. It consists of 8 elements (outcomes), each of which has probability $1/16$, for an overall probability of $8/16 = 1/2$.

Probability calculation, a continuous example

We will now go through a probability calculation for the case where we have a continuous sample space. We revisit our earlier example in which we were throwing a dart into a square target, the square target being the unit square. And we were guaranteed that our dart would fall somewhere inside this set. So our sample space is the unit square itself.

We have a description of the sample space, but we also need to specify a probability law. The choice of a probability law could be arbitrary. It's up to us to choose how to model a certain situation. And to keep things simple, we're going to assume that our probability law is a uniform one, which means that the probability of any particular subset of the sample space is going to be the area of that subset.

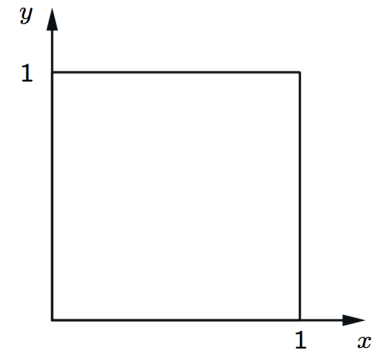


Figure 13: An example of a continuous sample space: Ω consists of all (x, y) such that $0 \leq x, y \leq 1$.

Uniform probability law: Probability = Area

So if we have some subset lying somewhere in the unit square and we ask what is the probability that we fall into that subset? The probability is exactly the area of that particular subset. Once more, this is an arbitrary choice of a probability law. There's nothing in our assumptions so far that would force us to make this particular choice. And we just use it for the purposes of this example.

So now let us calculate some probabilities. Let us look at this event. This is the event that the sum of the two numbers that we get in our experiment is less than or equal to $1/2$:

$$P(\{(x, y) \mid x + y \leq 1/2\}) = ?$$

It is always useful to work in terms of a picture and to depict that event in a picture of the sample space. So in terms of that sample space, the points that make this event to be true are just a triangle that lies below the red line shown in this picture, that's x plus y equals $1/2$. Below that line are the outcomes that make this event happen.

So we're trying to find the probability of the event shaded in red in Figure 14. We have assumed that probability is equal to area.

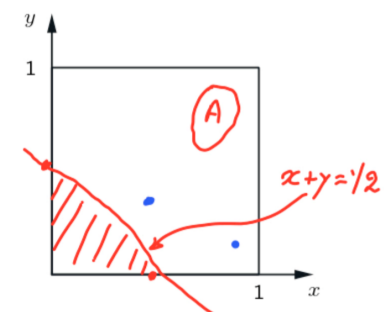


Figure 14: We are interested in (x, y) such that $x + y \leq 1/2$.

Therefore, the probability we're trying to calculate is the area of a triangle. And the area of a triangle is $1/2$ times the base of the triangle, which is $1/2$ in our case, times the height of the triangle, which is again $1/2$ in our case. And the end result is $1/8$:

$$P(\{(x, y) \mid x + y \leq 1/2\}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

Let us now calculate another probability - let's find the probability of an event that consists of only a single element. We take the point $0.5, 0.3$, which sits somewhere in the square. The event of interest is a set, but that set consists of a single point. So we're asking for the probability that our dart falls exactly on top of that point.

What is it? Well, it is the area of a set that consists of a single point. What is the area of a single point? It is 0 . And similarly for any other single point inside that sample space that we might have considered the answer is going to be 0 .

Check Your Understanding: Consider a sample space that is the rectangular region $[0, 1] \times [0, 2]$, i.e., the set of all pairs (x, y) that satisfy $0 \leq x \leq 1$ and $0 \leq y \leq 2$. Consider a "uniform" probability law, under which the probability of an event is half of the area of the event. Find the probability of the following events:

Please don't read the answers before attempting the questions.

- The two components x and y have the same values.
- The value, x , of the first component is larger than or equal to the value, y , of the second component.
- The value x^2 is larger than or equal to the value of y .

and therefore the corresponding probability is $1/6$.

$$\frac{3}{1} = \int_1^0 x^2 dx = \left. \frac{x^3}{3} \right|_1^0$$

(c) This event corresponds to the region below the curve $y = x^2$, where x ranges from 0 to 1 . The area of this region is

probability is $1/4$.

(b) This event is a triangle with vertices at $(0,0)$, $(1,0)$, $(1,1)$. Its area is $1/2$, and therefore the

(a) This event is a line, and since a line has zero area, the probability is zero.

ANSWERS:

Let us now abstract from this example, as well as the previous one, and note the following. Probability calculations involve a sequence of four steps. Starting with a word description of a problem, of a probabilistic experiment, we first write down the sample space.

Probability calculation steps

- Specify the sample space
- Specify a probability law
- Identify an event of interest
- Calculate...

Then we specify a probability law. Let me emphasize again here that this step has some arbitrariness in it. You can choose any probability law you like, although for your results to be useful it would be good if your probability law captures the real-world phenomenon you're trying to model.

Next, typically you're interested in calculating the probability of some event. That event may be described in some loose manner, so you need to describe it mathematically. And if possible, it's always good to describe it in terms of a picture. Pictures are immensely useful when going through this process.

And finally, the last step is to go ahead and calculate the probability of the event of interest. Now, a probability law in principle specifies the probability of every event, and there's nothing else to do. But quite often the probability law will be given in some implicit manner, for example, by specifying the probabilities of only some of the events. In that case, you may have to do some additional work to find the probability of the particular event that you care about.

This last step sometimes will be easy. Sometimes it may be complicated. But in either case, by following this four-step procedure and by being systematic you will always be able to come up with a single correct answer.

Probability calculation, discrete but infinite sample space

We have seen so far an example of a probability law on a discrete and finite sample space as well as an example with an infinite and continuous sample space. Let us now look at an example involving a discrete but infinite sample space.

We carry out an experiment whose outcome is an arbitrary positive integer. As an example of such an experiment, suppose that we keep tossing a coin and the outcome is the number of tosses until we observe heads for the first time. The first heads might appear in the first toss or the second or the third, and so on. So in this example, any positive integer is possible. And so our sample space is infinite.

Let us now specify a probability law. A probability law should determine the probability of every event, of every subset of the sample space. That is, the probability of every set of positive integers. But instead I will just tell you the probability of events that contain a single element. I'm going to tell you that there is probability $1/2^n$ that the outcome is equal to n :

$$P(n) = \frac{1}{2^n}$$

Is this information good enough to determine the probability of *any* subset?

Before we look into that question, let us first do a quick sanity check to see whether these numbers that we are given look like legitimate probabilities. Do they add to 1? Let's do a quick check. So the sum over all the possible values of n of the probabilities that we're given, which is an infinite sum starting from 1, all the way up to infinity, of $1/2^n$, is equal to the following:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{2^n}$$

And now on the right side we have a usual infinite geometric series and we have a formula for this: $\sum_{n=0}^{\infty} \frac{1}{2^n} = \frac{1}{1-1/2}$. Thus

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{2^n} &= \frac{1}{2} \cdot \frac{1}{1-1/2} \\ &= \frac{1}{2} \cdot 2 \\ &= 1 \end{aligned}$$

So indeed, *it appears* that we have the basic elements of what it would take to have a legitimate probability law.

But now let us look into how we might calculate the probability of some general event. For example, the probability that the outcome

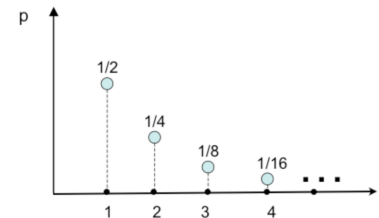


Figure 15: An example of a discrete but infinite sample space.

is even. We proceed as follows. The probability that the outcome is even is the probability of an infinite set that consists of all the even integers. We can write this set as the union of lots of little sets that contain a single element each. So it's the set containing the number 2, the set containing the number 4, the set containing the number 6, and so on:

$$\begin{aligned} P(\text{outcome is even}) &= P(\{2, 4, 6, \dots\}) \\ &= P(\{2\} \cup \{4\} \cup \{6\} \cup \dots) \\ &= P(\{2\}) + P(\{4\}) + P(\{6\}) \dots \end{aligned}$$

The last equality follows because we notice that we're talking about the probability of a union of sets and these sets are disjoint because they contain different elements. So we can use an additivity property and say that this is the probability of obtaining a 2, plus the probability of obtaining a 4, plus the probability of obtaining a 6 and so on.

If you're curious about doing this calculation and actually obtaining a numerical answer, you would proceed as follows. You notice that the last expression above equals to:

$$\begin{aligned} \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} \dots &= \frac{1}{4} \cdot (1 + \frac{1}{4} + \frac{1}{4^2} \dots) \\ &= \frac{1}{4} \cdot \frac{1}{1 - 1/4} \\ &= \frac{1}{3} \end{aligned}$$

Thus the probability that the outcome of this experiment is an even number is $1/3$. Does this result sound intuitive to you?

Leaving the details of the calculation aside, the more important question I want to address is the following. Is this calculation correct? We seem to have used an additivity property at this point. But the additivity properties that we have in our hands at this point only talk about disjoint unions of finitely many subsets.

Our initial axiom talked about a disjoint union of two subsets and then later on we established a similar property for a disjoint union of finitely many subsets. But here we're talking about the union of infinitely many subsets. So this step of taking an infinite union is not really allowed by what we have in our hands.

On the other hand, we would like our theory to allow this kind of calculation.

The way out of this dilemma is to introduce an additional axiom that will indeed allow this kind of calculation. The axiom that we introduce is the following. If we have an infinite sequence of disjoint

events, as for example in this picture. We have our sample space. We have a first event, A_1 . We have a second event, A_2 . The third event, A_3 . And so we keep continuing and we have an infinite *sequence of such events*. Then the probability of the union of these events, of these infinitely many events, is the sum of their individual probabilities.

Countable Additivity Axiom:

If A_1, A_2, A_3, \dots is an infinite **sequence** of **disjoint** events, then $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$

The key word here is the word **sequence**. Namely, these events, these sets that we're dealing with, can be arranged so that we can talk about the first event, A_1 , the second event, A_2 , the third one, A_3 , and so on.

To appreciate the issue that arises here and to see why the word sequence is so important, let us consider the following calculation. Our sample space is the unit square. And we consider a model where the probability of a set is its area, as in the examples that we considered earlier. Let us now look at the probability of the overall sample space. Our sample space is the unit square and the unit square can be thought of as the union of various sets that consist of single points. So it's the union of subsets with one element each. And it's a union taken over all the points in the unit square.

Then we think about additivity. We observe that these subsets are disjoint. If we're considering different points, then we get disjoint single element sets. And then an additivity property would tell us that the probability of this union is the sum of the probabilities of the different single element subsets.

Now, as we discussed before, single element subsets have 0 probability. So we have a sum of lots of 0s and the sum of 0s should be equal to 0. On the other hand, by the probability axioms, the probability of the entire sample space should be equal to 1.

$$\begin{aligned} 1 &= P(\Omega) = P(\cup \{x, y\}) \\ &= \sum P(\{x, y\}) = \sum 0 = 0 \end{aligned}$$

And so we have established that 1 is equal to 0.

This looks like a paradox. Is it?

The catch is that there is nothing in the axioms we have introduced so far or the properties we have established that would justify the

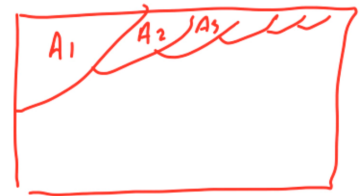


Figure 16: An infinite sequence of events.

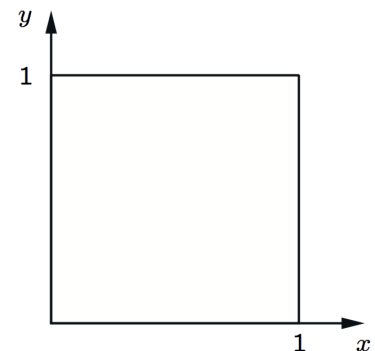


Figure 17: The unit square again. Each point has zero probability in our model. All points make up the entire square, which has probability 1. There is an infinite number of such points. Can the sum of these infinitely many zeros add up to one?

step that says $P(\cup\{x, y\}) = \sum P(\{x, y\})$ in these calculations. This is the step that's questionable!

You might argue that the unit square is the union of disjoint single element sets, which is the case that we have in additivity axioms. But the additivity axiom only applies when we have a sequence of events. And this is not what we have here. This is not a union of a sequence of single element sets. In fact, there is no way that the elements of the unit square can be arranged in a sequence. The unit square is said to be an uncountable set.

- Additivity holds only for "countable" sequences of events
- The unit square (similarly, the real line, etc.) is **not countable** (its elements cannot be arranged in a sequence)
- "Area" is a legitimate probability law on the unit square, as long as we do not try to assign probabilities/areas to "very strange" sets

This is a deep and fundamental mathematical fact. What it essentially says is that there are two kinds of infinite sets. Discrete ones, or, in formal terminology, countable. These are sets whose elements can be arranged in a sequence, like the integers.

And then there are uncountable sets, such as the unit square or the real line, whose elements cannot be arranged in a sequence. If you're curious, you can find the proof of this important fact in supplementary materials (ask me).

After all these discussions, you may now have legitimate suspicions about the models we have been looking at. Is area a legitimate probability law? Does it even satisfy countable additivity? This question takes us into deep waters and has to do with a deep subfield of mathematics called Measure Theory.

Fortunately, it turns out that all is well. Area is a legitimate probability law. It does indeed satisfy the countable additivity axiom as long as we only deal with nice subsets of the unit square. Fortunately, the subsets that arise in whatever we do in this course will be "nice". Subsets that are not nice are quite pathological and we will not encounter them.

At this stage we are not in a position to say anything more that would be meaningful about these issues because they're quite complicated and mathematically deep. We can only say that there are some serious mathematical subtleties. But fortunately, they can all be overcome in a rigorous manner. And for the rest of this class, you can just forget about these subtle issues.