

Week 1

13 May 2021

We start this class at the very beginning: what is probability? When we think about probability, we normally think that it's the study of chance. Such as, what are the chances that I'll win the lotto or that our favourite team will win the tournament. But we can also use probability to figure out whether the information we are looking at is accurate, how different events can interact with one another, and so much more.

Since, in mathematics, we like to be precise when we talk about the world, we will be working towards taking the word "chance" and making it as precise as possible.

1 Events

Normally when we think of chance we think of flipping a coin, rolling a dice or playing the lotto. All of these are examples of chance which come from equally likely outcomes. For example, if we look at flipping a coin, I have $\frac{1}{2}$ chance that it will turn out to be heads and $\frac{1}{2}$ that it will be tails. An equal 50% – 50% for both outcomes. We start our exploration of probability with examples such as the above; examples where the outcome is equally probable.

So say I want to roll a dice and I want to figure out my chances of rolling a number. We inherently know the chance, but how do we do this mathematically? **First**, we need to take all possible outcomes and put them into a set. This set is called the *sample space*, and we (usually) denote it with an Ω . In our example of rolling a six-sided die, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The elements in our sample space Ω are normally called *outcomes*. Something that might or might not occur based on the outcome are called *events*. Using the dice, we might have events such as:

- The dice rolls a six.
- The dice rolls an even number.

Wikipedia: [Sample Space](#)

Note that in the book by Pitman, "sample space" is called an *outcome space*. The two are the same thing.

Wikipedia: [Outcome](#)

- The dice rolls a number less than 3.
- The dice doesn't roll a 4.

Our aim is to study what the probability or chance of certain events occurring is.

Example 1.1 Let's do an example. Say we're about to do a raffle in class and there are 200 students. We draw one name from a hat for who will get an A+ in the class, no questions asked. What is the sample space of the raffle? $\Omega = \underline{\hspace{2cm}}$.

Try and list some events for this space:

-
-
-

So what is an "event" mathematically? An *event* is a subset of our sample space Ω . For example, "the dice rolls a six" is represented by the subset $\{6\}$. The event "the dice rolls an even number" is represented by the subset $\{2, 4, 6\}$. How is the event "the dice rolls a number less than 3" represented? $\underline{\hspace{2cm}}$.

Sometimes, language actually helps us create these subsets!

Wikipedia: [Event](#)

"A or B" means $A \cup B$.

Or: Say I take the event: "The dice rolls a six or a two". This is really the combination of two potential events: "the dice rolls a six" and the event "the dice rolls a two". Since both are possible, we can represent this as: $\{6\} \cup \{2\} = \{2, 6\}$. In other words, for most purposes, the word "or" can be thought of as a union of two sets.

"A and B" means $A \cap B$. Note that the book uses AB to mean $A \cap B$.

And: What about "and"? Let's test it out with our event above by switching the word "or" with "and" - "The dice rolls a six *and* a two". Inherently, if we're rolling a die one time, this is impossible. You can't have two numbers show up by rolling a die so we expect the sample space to be empty. In terms of sets, we can think of "and" by intersecting sets: $\{6\} \cap \{2\} = \emptyset$. When we see that an event is empty (aka equal to \emptyset) then we say that the event is *impossible*. The opposite of an impossible event is a *certain event* and is when the event is Ω .

"not A" means $\Omega \setminus A$.

This is sometimes known as the "complementary event".

Wikipedia: [Complementary event](#)

Not: The final word we'll quickly go over is the word "not"; in other words, what is the chance of something *not* happening. An example of this is the event "the dice doesn't roll a 4". What we really want here is to take our set to be $\{4\}$ and remove it from our sample space in order to be left with the actual possible outcomes. In other words, we take the complement of our sample space: $\Omega \setminus \{4\} = \{1, 2, 3, 5, 6\}$.

Example 1.2 Let's try the above with the stereotypical example of flipping a coin. The sample space here is given by $\Omega = \underline{\hspace{2cm}}$. What subset of Ω is represented by the following events:

- The outcome is heads $\underline{\hspace{2cm}}$
- The outcome is heads or tails $\underline{\hspace{2cm}}$
- The outcome is heads and tails $\underline{\hspace{2cm}}$
- The outcome is not head $\underline{\hspace{2cm}}$

Here is a table of some useful conversions from English to Set theory/Probability.

Human language	Notation	Set language
Sample space	Ω	Collection of outcomes
Event that some outcome in A occurs	A	Subset of Ω
Not A	$A^c = \Omega \setminus A$	Complement of A
A and B	$A \cap B$	Intersection
A or B	$A \cup B$	Union
A , but not B	$A \setminus B$	Difference
Either A or B , but not both	$A \Delta B$	Symmetric difference
If A then B	$A \subseteq B$	Inclusion
Impossible event	\emptyset	Empty set
Certain event	Ω	Whole sample space.

2 Probability

Now that we've discussed events, we want to know what the probability of a certain event occurring is. If all outcomes in a finite set Ω are equally likely then the probability of an event $A \subseteq \Omega$ to occur is:

Wikipedia: [Probability](#)

Example 2.1 Let's look at the four events we had at the very beginning with rolling a dice. We know that $\Omega = \{1, 2, 3, 4, 5, 6\}$ and so $|\Omega| = 6$.

- "The dice rolls a six." – First we need to calculate our event A . In this case $A = \{6\}$. So we have

$$P(A) = \underline{\hspace{2cm}}$$

- “The dice rolls an even number.” – This statement is equivalent to “The dice rolls a 2 or a 4 or a 6”. Recalling that “or” means union, we have $A = \{2\} \cup \{4\} \cup \{6\} = \{2, 4, 6\}$. Therefore

$$P(A) = \underline{\hspace{2cm}}$$

- “The dice rolls a number less than 3.” –

- “The dice doesn’t roll a 4.” – We saw before that

$$A = \{1, 2, 3, 5, 6\} = \{4\}^c = \Omega \setminus \{4\}.$$

Therefore

$$P(A) = \underline{\hspace{2cm}}$$

Example 2.2 What happens if we flip multiple coins instead of one? Let’s say we have 3 coins and we flip them one at a time. What is our sample space?

Each coin can be heads or tails, and since we are flipping them one at a time we can put them in the order of when we flipped them. In other words we can think of each outcome as a list of three outcomes:

Notice how each outcome has three values and they are ordered based on which coin it is referring to. In other words (H, H, T) is not the same as (T, H, H) even though they both represent that you got two heads and one tails. *The _____ matters in this case since we are flipping them one at a time.* Therefore our sample space Ω contains the 8 elements above (*i.e.*, $|\Omega| = 8$).

Let’s look at some events. Say our event is “Flip heads twice.” First, we must notice that this statement is ambiguous! Does the statement mean we get heads exactly two times or does it mean we flip heads at least two times? (So three heads would be ok?) *Every word is important in probability.* Let’s try both and see the difference.

Flip heads exactly twice: In this case we have

$$A =$$

which are all outcomes in our sample space Ω where heads appears two times. Therefore our probability is given by:

$$P(A) = \underline{\quad}$$

Flip heads at least twice: In this case we are really asking “Flips heads twice or flips heads three times”. So our event is given by

$$A =$$

which are all possible outcomes. Therefore our probability is given by:


$$P(A) = \underline{\quad}$$

Another event we might want to look at is “We flip heads first”. If you think about it for a second, you’ll realize we’re just asking what the first flip’s outcome is, but we’ll be pedantic and do things slowly. Our set A should contain every outcome whose first entry is H :

$$A =$$

Therefore our probability is given by:

$$P(A) = \underline{\quad}$$

Example 2.3  The following example is a common student mistake! Let’s see what happens when you think about these flips differently. If we have 3 coins and we flip them all at once. We might be tempted to say that we can just look at the sets themselves in order to determine the probabilities. In this case we would set our sample space to be:

$$\Omega =$$

In other words, we only have 4 potential outcomes instead of 8. What happens when we look at the events from the previous example?

“Flip heads at least twice”: Our new event is given by

$$A =$$

This gives us a probability $P(A) = \underline{\quad}$ which is the same as before! We might be tempted to say that order doesn’t matter, until we look at one of the other events.

“Flip heads exactly twice”: Our new event is given by

$$A =$$

giving us a probability $P(A) = \underline{\quad}$. This is definitely not the same as before!

So what are we doing wrong? First, we need to realize that flipping coins are all independent of one another. In other words, the result on one coin does not influence the result on another coin. This means they are “independent” which is something we will cover in more depth later on. Because our coins are independent, looking at coin tosses as sets is problematic because it doesn’t take into account variations within the set. In other words, the set $\{H, H, T\}$ doesn’t take into account that the tail flip could have come from the first coin or the second coin or the third coin! Notice how I said “the first coin or the second coin or the third coin”. What this set actually represents is a *union* of our three lists from the previous example, but we randomly decided to combine them into a set where order doesn’t matter.

This is probably the hardest part about probability: making sure that your sample space and your events are accurately describing the problem. It is very easy to merge things together because you didn’t think about all possible variations beforehand.

3 Counting

We’re going to take a small detour and talk about counting since we are noticing that set theory and probability are intimately related. Most of these seem trivial, but it’s good to be mathematically precise when we talk about them.

We let $|A|$ denote the number of elements in a set A . Recall that a 1 – 1 *correspondence* is _____

The correspondence rule: If we have two sets A and B and we can make a 1 – 1 correspondence between the elements then

We say that two sets A and B are *disjoint* if $A \cap B = \emptyset$. If we have multiple sets A_1, A_2, \dots, A_n then we say they are *mutually disjoint* if every pair of sets are disjoint, *i.e.*, $A_i \cap A_j = \emptyset$ for all i, j .

The addition rule: If we have a set A that can be split into a collection of mutually disjoint sets A_1, A_2, \dots, A_n then

The multiplication rule: Suppose we have k choices to be made one after another with exactly n_j options at each choice with $j \leq k$. Then the total number of options which can be made are

Sequences: A *sequence of length k of elements of S* is an ordered list which has k elements each coming from the set S . Each component is chosen without consideration of the other ones and so there are $|S| = n$ options to fill in each component giving us _____ total sequences possible.

Example 3.1 For example if $S = \{A, B, C\}$ and we look at $k = 4$ sequences we'll have $3^4 = 81$ different sequences. Here are some:

Orderings: An *ordering of k out of n elements* is a sequence of k elements where every entry is different. This means that each time we chose an element, it can no longer be selected! So we have n options for the first entry, $n - 1$ options for the second entry, etc. This means there are

total options. Recall that n factorial is given by $n! = n(n - 1)(n - 2) \dots 2 \cdot 1$.

Then we can represent the above as _____ . $0! = 1$

Example 3.2 For example if $S = \{A, B, X, Y\}$ and we let $k = 2$ then we have the following $\frac{4!}{2!} = 4 \cdot 3 = 12$ options

Notice how the order matters!

Permutation: A *permutation of an ordering of k out of n elements* is a reshuffling of the entries of an ordering.

Example 3.3 For example if I have the ordering (A, B, C) then we have six permutations:

Notice how each permutation has the same content, but the order in the orderings is different.

In this case we have $3 \cdot 2 \cdot 1$ total options. Normally when we see something like this, we use the factorial symbol. In this case $3! = 3 \cdot 2 \cdot 1$.

For an ordering with k elements we have ____ permutations.

Combinations: Say that we don't actually care about order. So (A, B) and (B, A) are the same thing. In this case we can look at our ordering and divide by the number of permutations each ordering has. A *combination of k elements out of an n element set S* is a subset of S with k elements. We end up with $\frac{n!}{(n-k)!k!}$ which we normally denote by _____. This is called the *choice function* or the *binomial coefficient*. You can read it as " n choose k " because we are taking a set of n elements and we are choosing k of them.

Wikipedia: [Binomial coefficient](#)

The term "binomial coefficient" is coming from algebra where $(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$.

Example 3.4 For example if $S = \{tree, fern, sprig, grass\}$ then we have the combinations of 3 elements given by:

$\{tree, fern, sprig\}, \{tree, fern, grass\}, \{tree, sprig, grass\}, \{fern, sprig, grass\}$

Notice that all we did was we *chose three out of four* elements: $\binom{4}{3} = \frac{4!}{3!1!} = 4$.

Subsets: Finally, suppose I have an n element set S and I want to look at all possible subsets of S . Then for each element I need to decide whether that element is in our subset or not. In other words, I have 2 options for each element. This gives ____ different possibilities.

Example 3.5 For example if $S = \{A, B\}$ then my subset options are:

$\emptyset, \{A\}, \{B\}, S$

giving us $2^2 = 4$ total subsets.

We normally denote by 2^S the _____. In other words:

4 Odds

Another way to discuss an event is to describe the odds of an event happening which is often used in gambling. (Think of when people say “the odds are 3 to 2 that the person will win”.)

Odds are normally just another representation of probability that compares the chance of something occurring vs the chance of something not occurring. In mathematical terms, the *odds in favour* of something happening are $|A|$ to $|A^c|$. Alternatively, the *odds against* something happening are $|A^c|$ to $|A|$. Usually these are denoted as $|A| : |A^c|$

Wikipedia: [Odds](#)

Example 4.1 For example, if we roll a six sided dice, and we want to find the odds of rolling a 4.

We use odds super rarely in this class as it's just a different representation of probability itself. It is something that you should know though as this is an extremely common way of representing probability in the real world.

Example 4.2 Say you're watching a horse race and they say horse A has odds 7 : 1 of winning. What is the probability that the horse will win?

5 Interpreting probability

So we talked about the (extremely super) basics of probability, but a question we ask next is how should we interpret a probability? In other words, when we know the probability of something occurring, how do we want to use/interpret that probability? There are two main ways of interpreting probability called “frequency” and “subjective” interpretations. There are other interpretations such as “propensity”, “logical” and “predictive”, but those you can look up on Wikipedia or another book.

Wikipedia: [Interpretations](#)

Frequency interpretations ask _____

_____. When we look at probability from this standpoint, we call it *frequency probability*.

Wikipedia: [Frequency probability](#)

Usually we use this interpretation when something happens more than once. If you find yourself asking “how often will x occur”, then you’re thinking in frequency terms.

In real life, we normally test probability by running examples to see how often things actually happen in order to make sure we didn’t forget something. For this, we make a prediction using probability and then we run simulations to record the relative frequency. The *relative frequency* is a proportion measuring how often/frequently something occurs in a sequence of observations. We normally record these observations over time and see what value they trend to.

Example 5.1 Say we want to verify that the chance of rolling a 5 on a six-sided die is $\frac{1}{6}$. What we would do is roll the die once and record if it was a 5 or not. We then roll it again and record if it was a 5 or not. We repeat this process for an extremely long amount of time, always recording whether 5 occurred or not. This gives us a list of how often 5 occurred up to a point. For example say we were super lucky and saw 5 appear the 3rd, 4th and 7th time we rolled a dice out of ten times. Our list would look like:

Notice that the numerator increased by one each time I had a success (*i.e.*, I rolled a 5) and the denominator increased for each time an attempt was made.

After our simulations are done, we can plot these results on a graph to see what they look like. Our simulation gives us points $(i, \frac{x_i}{i})$ where i is the i th roll and x_i is the number of cases up to that point. Using our data from above we would have the points:

6 Distributions

We're now going to take probability and put it neck deep in mathematics. So far what we've been doing is starting with some set Ω of possible outcomes which we called the sample space. We then took events and we represented them by subsets A of Ω . Finally, we created a probability function P which took A and assigned it a number.

We'll now formalize this and try and make it more mathematically precise. We let Ω be some set of outcomes. Note that we don't assume Ω to be finite nor do we assume that every outcome is equally likely to occur. Let 2^Ω denote the set of all subsets of Ω and let $A \in 2^\Omega$, *i.e.*, A is a subset of Ω . We define the probability map to be the map $P : 2^\Omega \rightarrow \mathbb{R}$ which sends every subset to some real number.

Although all of this makes mathematical sense in set theory, we want to ask, how do we know if this represents probability or not? For this, we need to define some set of rules which ensure that P behaves "nicely". One easy example is that for every A we want $P(A)$ to be positive! (What does negative probability even mean?) What other rules do we need?

Let's consider multiple events happening. Since we're thinking of events as sets, let's break down what happens when sets interact with one another. We've already seen union, intersection, etc. but now we'll introduce something called a partitioning. We say that two sets A and B are *disjoint* if $A \cap B = \emptyset$. We use the notation $A \sqcup B$ to mean $A \cup B$ where A and B are disjoint. An event A is *partitioned* into n smaller events if:

$$A = A_1 \sqcup A_2 \sqcup \dots \sqcup A_n.$$

Since we are breaking apart A into smaller chunks, one thing we want is that the sum of the little chunks should equal the probability of A itself. This gives us a second rule:

$$P(A) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Another rule we might want is that if $A = \Omega$ then our probability should be 100%.

In particular, we end up having three rules, which together we call the *rules of proportion and probability*.

- **Non-negative:** $P(A) \geq 0$
- **Addition:** If A_1, A_2, \dots, A_n is a partition of A then

$$P(A) = P(A_1) + P(A_2) + \dots + P(A_n)$$

- **Total one:** $P(\Omega) = 1$.

For us, event and subset are synonymous. Also, remember that \cup just means "or".

We can also technically partition A into an infinite number of subsets! We'll look at this in later classes.

A *distribution over* Ω is a function of subsets of Ω satisfying these three rules.

With this we can be a little more stringent with our language translations from before.

Theorem 6.1 (Complement Rule) *The probability of the complement of an event A is*

$$P(\text{not } A) =$$

Proof.

Wikipedia: [Distribution](#)

□

Theorem 6.2 (Difference Rule) *The probability that B will occur and A will not is given by:*

$$P(B \text{ and not } A) =$$

Proof.

Note that the book assumes $A \subseteq B$, but we don't make that assumption.

□

Theorem 6.3 (Inclusion-Exclusion) *Let A and B be two subsets of Ω . Then*

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

Proof.



Example 6.4 Say I have a bag of marbles and I tell you that 15% of the marbles are blue, 12% are red and 5% are both red and blue. If I pull out a marble, what are the chances that the marble is not blue?

If I pull out a marble, what are the chances that the marble is red, but not blue?

If I pull out a marble, what are the chances that the marble is either

blue or red?

Example 6.5 Say I roll a 6-sided die twice. What is the distribution of the sum of the values?

Recall that the sample space for rolling a die twice can be represented by ordered pairs (i, j) and that we end up with 36 possible outcomes. To compute the distribution of the sums, we need to figure out how many times each sum appears.

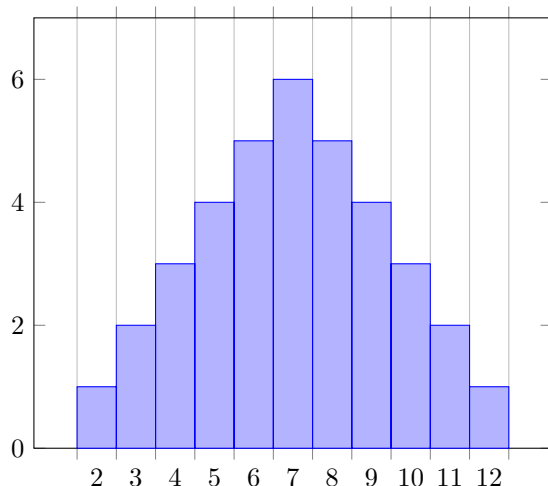
We know that to get a 2 we need to roll a 1 followed by a 1 (represented by $(1, 1)$). Since this is the only way we know $P(2) = \frac{1}{36}$.

Similarly, to get a 7 we have the following 6 options

$$(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$$

Therefore, we get the following distribution:

$$\begin{aligned} P(2) &= \frac{1}{36}, P(3) = \frac{2}{36}, P(4) = \frac{3}{36}, P(5) = \frac{4}{36}, \\ P(6) &= \frac{5}{36}, P(7) = \frac{6}{36}, P(8) = \frac{5}{36}, P(9) = \frac{4}{36} \\ P(10) &= \frac{3}{36}, P(11) = \frac{2}{36}, P(12) = \frac{1}{36} \end{aligned}$$



Wikipedia: [Bernoulli \$p\$ distribution](#)

6.1 Bernoulli p distribution

Probably the easiest distribution to understand is the *Bernoulli distribution*. It is basically a generalization of _____ . We suppose that Ω has exactly two outcomes $\{A, B\}$. We let $P(A) = p \in [0, 1]$ and we let $P(B) = 1 - p$, which also is contained in $[0, 1]$. Notice that $A^c = B$ and $B^c = A$.

Let's verify that this actually gives a distribution. First, notice that every subset gives a probability between 0 and 1. Next, notice that the only set that can be partitioned is Ω with the partition A and B . Also notice that

$$P(\Omega) = P(A) + P(B) = p + 1 - p = 1.$$

Since all three points of the rules proportion and probability are satisfied, this is a distribution!

Wikipedia: [Discrete uniform distribution](#)

Note that the book uses “uniform distribution on a finite set” to mean discrete uniform distribution.

If $n = 2$, then the uniform distribution is the same as the Bernoulli $\frac{1}{2}$ distribution.

6.2 Discrete uniform distribution

A *discrete uniform distribution* is a generalization of _____ .

In essence we let $\Omega = \{1, 2, \dots, n\}$ be a set of n elements. We suppose that the probability of choosing any one of them is given by $\frac{1}{n}$, *i.e.*, each outcome is equally likely.

To verify this is a distribution, we first notice that $P(A) = \frac{|A|}{n}$ for every $A \subseteq \Omega$ and that $P(A) \in [0, 1]$ always. Partitioning can be proved inductively, and the final rule can be deduced from above.

6.3 Continuous uniform distribution

A *continuous uniform distribution* is a generalization of the _____ to the set of real numbers. Here the idea is that we take an interval (a, b) (where $a < b$) and we pick a point randomly in the interval. (We assume that every point has an equal chance of getting picked). In this case we can't ask questions like “What is the probability that we pick the point x ” because the chance is 0. (The reasoning for this is discussed in future lectures). Instead we ask questions like “What is the probability that the point is between x and y where $a < x < y < b$ ”. The probability in this case is given by $\frac{y-x}{b-a}$. If $a = 0$ and $b = 1$ then we call this the *standard uniform distribution*.

Wikipedia: [Continuous uniform distribution](#)

Note that the book uses “uniform (a, b) distribution” to mean continuous uniform distribution.