* INTRODUCTION TO PROBABILITY
* Games of chance have been around for centuries. Knowing how to compute probabilities gives you an edge in these games. As a result, throughout history, many, many smart people, including famous mathematicians such as Cardano, Fermat, Pascal spent time and energy thinking about the math of games of chances.
* As a result, probability theory was born. Probability continues to be highly useful in modern games of chance.
* For example, in poker, we can compute the probability of winning a hand based on the cards on the table. Casinos rely on probability theory to develop games that guarantee a profit.
* It turns out probability theory is useful in many other contexts and, in particular, those that depend on data affected by chance in some way. As a result, knowledge of probability is indispensable for data science.
* In games of chance, probability has a very intuitive definition. For example, we know what it means that the chance of a pair of dice coming out 7 is one in six. However, this is not the case in other contexts. Today, probability is being used much more broadly with the word “probability” commonly used in everyday language. For example, Google’s auto-complete of “what are the chances of” give us having twins, rain today, getting struck by lightning, and getting cancer.
* We also see the word probability used by election forecasters. In 2008, Nate Silver gave Obama a 94% chance of wining. In 2012, it was 90%. Obama won both elections. In 2016, however, Mr. Silver was not as certain and gave Hilary Clinton only a 71% chance of wining. She lost. But 71% is still more than 50%. Was Mr. Silver wrong ? And what does probably mean in this context anyway? Are there dice being tossed somewhere?
* To answer this questions, we will need to learn, among other things, some probability theory.
* We cover election forecasting in detail in the next module, since we also need to understand the concept of statistical inference which we explain there. But note that **statistical inference** builds upon probability theory.
* The motivation for this module are the circumstances surrounding the financial crisis of 2007 to 2008. Part of what caused this financial crisis was that the risk of certain securities sold by financial institutions was underestimated. Specifically, the risk of mortgage backed securities and Collateralized Debt Obligations, or CDOs, were grossly underestimated.
* These two were sold at prices that assume most home owners would make their monthly payments, and that the probability of this not happening was low . A combination of factors resulted in many more defaults than were expected. This resulted in a price crash of these securities. As a consequence, banks lost so much money that they needed government bailouts to avoid closing down completely.
* To begin to understand this very complicated event, we need to understand the basics of probability. We will introduce important concepts such as random variables, independence, Monte Carlo simulation, expected values, standard errors, margin of errors and the central limit theorem.
* Before we start to try to understand the financial crisis using probability, we will use several examples related to games of chance, as these are simple and illustrative.
* 1.1 INTRODUCTION TO DISCRETE PROBABILITY
* We start by covering some basic principles related to categorical data. This subset of probability is referred to as **discrete probability**.
* It will help us understand the probability theory we will later introduce for numeric and continuous data, which is more common in data science applications.
* Discrete probability is more useful in card games and we use these as examples.
* The word probability is used in everyday language. Here, we discuss a mathematical definition of probability that does permit use to give precise answers to certain questions. For example, if I have two red beads and three blue beads inside an urn and I pick one at random, what is the probability of picking a red one?
* our intuition tells us the answer is 2/5, or 40%.
* A precise definition can be given by noting there are five possible outcomes of which two satisfy the condition necessary for the event “pick a red bead.”
* Because each of the five outcomes has the same chance of occurring, we conclude that the probability is 0.4 for red and 0.6 for blue.
* A more tangible way to think about the probability of an event is as a proportion of times the event occurs when we repeat the experiment over and over independently and under the same conditions.
* Before we continue, let’s introduce some notation.
* We use the notation Pr(A) to denote the probability of an event A happening.
* We use the very general term event to refer to things that can happen when something happens by chance. For example, in our previous example, the event was “picking a red bead”.
* In a political poll, in which we call 100 likely voters at random, an example of an event is “calling 48 Democrats and 52 Republicans”.
* in data science applications, we will often deal with continuous variables.
* In these cases, events will often be things like, is that person taller than 6 feet?
* in this case we write events in a more mathematical form.
* For example, x>=6.
* We’ll see more of these examples later. Here, we focus on categorical data and discrete probability.
* 1.2 MONTE CARLO SIMULATIONS
* Computers provide a way to actually perform the simple random experiments, such as the one we did before.
* Pick a bead at random from a bag or an urn with 3 blue beads and 2 red beads.
* Random number generators permit us to mimic the process of picking at random. An example in R is the sample() function.
* We demonstrate its use showing you some code :
* beads <- rep(c(“red”,”blue”), times = c(2,3))
* First, use the rep() function to generate an urn.
* We create an urn with 2 red and 3 blues. You can see when we type beads we see this :
* [1] "red" "red" "blue" "blue" "blue"
* Now we can use a sample() function to pick one at random.
* sample(beads, 1)
* This line of code produces one random outcome.
* Now, we want to repeat this experiment over and over. However, it is, of course, impossible to repeat forever. Instead, we repeat the experiment a large enough number of times to make the results practically equivalent to doing it over and over forever. This is an example of a Monte Carlo simulation. Note that much of what mathematical and theoretical statisticians study –something we do not cover in this course – relates to providing rigorous definitions of “practically equivalent”, as well as studying how close a large number of experiment gets us to what happens in the limit, the limit meaning if we did it forever.
* later in this module, we provide a practical approach to decident what is large enough.
* To perform our first Monte Carlo simulation, we use the replicate() function. This permits us to repeat the same task any number of times we want.
* numevents <- 10000
* events <- replicate(numevents, sample(beads,1))
* We set numevents to be 10 000, then we use the replicate function to sample from the beads 10,000 times.
* We can now see if, in fact, our definition is in agreement with this Monte Carlo simulation approximation.
* We can use table(), for example to see the distribution.
* > tab<-table(events)
* > tab
* events
* blue red
* 6025 3975
* And then we can use prop.table to give us the proportions.
* > prop.table(tab)
* events
* blue red
* 0.6025 0.3975
* And we see that in fact, the Monte Carlo simulation gives a very good approximation with 0.5962 for blue and 0.4038 for the red.
* We didn’t exactly get 0.6 and 0.4, but statistical theory tells us that, if numevents is large enough, we can get as close as we want to those numbers.
* We just covered a simple and not very useful example of Monte Carlo simulations. But we will use Monte Carlo simulation to estimate probabilities in cases in which it is harder to compute exact ones.
* Before we go into more complex examples, we still use simple ones to demonstrate the computing tools available in R.
* Let’s start by noting that we don’t actually have to use replicate() in this particular example.
* This is because the function sample() has an argument that permits us to pick more than one element from the urn.
* However, by default, this selection occurs without replacement. After a bead is selected, it is not put back in the urn.
* Note what happens when we ask to randomly select 5 beads. Let’s do it over and over again. let’s do it three times.
* > sample(beads)
* [1] "blue" "blue" "red" "blue" "red"

> sample(beads)

* [1] "blue" "blue" "red" "blue" "red"
* > sample(beads)
* [1] "blue" "blue" "blue" "red" "red"
* This results in a rearrangement that always has three blue and two reds. But if we asked for six beads, then we get an error.
* > sample(beads,6)
* Error in sample.int(length(x), size, replace, prob) :
* cannot take a sample larger than the population when 'replace = FALSE'
* This is because it’s doing it without replacement. However, this function can be used directly –again, without the replicate—to repeat the same experiment of picking 1 out of 5 beads over and over under the same conditions. To do this, we sample with replacement. After we pick the bead we put it back in the urn.
* We can tell sample() to do this by changing the replace argument which defaults to false to true. We do it like this :
* events <- sample(beads,numevents,replace=TRUE)
* > prop.table(table(events))
* events
* blue red
* 0.6021 0.3979

And when we do this, we see that we get very similar answers to what we got using the replicate function.

PROBABILITY DISTRIBUTIONS

Defining a distribution for categorical outcomes is relatively straight forward.

We simply assign a probability to each category. In cases that can be thought of as beads in an urn, for each bead type, the proportion defines the distribution.

Another example comes from polling.

if you are randomly calling likely voters from a population that has 44% Democrat, 44% Republican, 10% undecided and 2% green, these proportions define the probability for each group.

For this example, the probability distribution is simply these four proportions.

Again categorical data makes it easy to define probability distributions. However, later in applications that more common in data science, we will learn about probability distributions for continuous variables.

in this case, it will get a little bit more complex. But for now, we are going to stick to discrete probabilities before we move on.

INDEPENDENCE

We say that two events are independent if the outcome of one does not affect the other. The classic example are coin tosses. Every time we toss a fair coin, the probability of seeing heads is ½ regardless of what previous tosses have revealed.

The same is true when we pick beads from an urn with replacement. In the example we saw earlier, the probability of red was 0.4 regardless of previous draws. Many examples of events that are not independent come from card games. When we deal the first card, the probability of getting a king is 1 in 13. This is because there are 13 possibilities. Now if we deal a king for the first card and I don’t replace it, then the problem of getting a king in the second card is less because there only three kings left. The probability is 3 out of not 52 –because we already dealt one card – but out of 51. These events are therefore not independent. The first outcome affects the second.

To see an extreme case of non-independent events, consider an example of drawing five beads at random without replacement from an urn. Three are blue, two are red. I’m going to generate data like this using the sample() function and assign it to x. You can’t see the outcomes. Now, if I ask you to guess the color of the first bead, what do you guess ? Since there’s more blues beads, there’s actually 0.6 chance of seeing blue. That’s probably what you guess.

But now, I am going to show you the outcomes of the other four. The second, third, fourth and fifth outcome you can see here. You can see that the three blue beads have already come out. This affects the probability of the first. They are not independent. So would you still guess blue? Of course not. Now, you know that the probability of red is 1. These events are not independent : the probabilities change once you see the other outcomes.

When events are not independent, conditional probabilities are useful and necessary to make correct calculations. We already saw an example of a conditional probability. We computed the probability that a second card dealt is a king given that the first card is already a king. In probability, we use the following notation :

***Pr(Card 2 is a king | Card 1 is a king) = 3/51***

We use the | as a shorthand for “given that” or “conditional on”.

Note that, when two events, A and B are independent, we have the following equation :

**Pr(A|B) = Pr(A)**

It doesn’t matter what B is. The probability of A is unchanged. This is the mathematical way of saying. And in fact, this can be considered the mathematical definition of independence.

Now, if we want to know the probability of two events, say A and B, occurring, we can use the multiplication rule. So :

Pr(A and B) = Pr(A) \* Pr(B|A)

Let’s use blackjack as an example. in blackjack, you get assigned two random cards without replacement. Then, you can ask for more. the goal is to get closer to 21 than the dealer without going over. Face cards are worth 10 points , so is the 10 card. And aces are worth either 11 or 1. So if you get an ace and a face card you win automatically.

So, in blackjack, to calculate the chances of getting 21 in the following way, first we get an ace and we get a face card or a 10, we compute the problem of the first being an ace and then multiply by the probability of a face card or a 10 given that the first card was an ace.

The calculation is 1/13—chance of getting an ace—times the chance of getting a card with the value 10 or a face card, given that we already saw an ace, which is 16 out of 51. This is approximately 2%.

The multiplicative rule also applies to more than two events. We can use induction to expand for more than two. So the probability of A and B and C is equal to the probability of B given that A happened times the probability of C given that A and B happened.

**Pr(A and B and C) = Pr(A)\*Pr(B|A)\*P(C|A and B)**

**When we have independent events**, the multiplication rule becomes simpler. **We simply multiply the three probabilities.**

But we have to be very careful when we use a multiplicative rule in practice. We’re assuming independence, and this can result in very different and incorrect probability calculations when we don’t actually have independence. This could have dire consequences, for example, in a trial if an expert doesn’t really know the multiplication rule and how to use it. So let’s use an example.

This is loosely based on something that actually happened. Imagine a court case in which the suspect was described to have a mustache and a beard, and the prosecution brings in an expert to argue that, because 1 in 10 men have beards and 1 in 5 men have mustaches, using the multiplication rule, this means that only 2% of men have both beards and mustaches—1/10 times 1/5. 2% is a pretty unlikely event. However, to multiply like this, we need to assume independence. And in this case, it is clearly not true. The conditional probability of a man having a mustache conditional on them having a beard is quite high. it’s about 95%. So the correct calculation actually gives us a much higher probability. It’s 9%. So there’s definitely a reasonable doubt.

1.2 COMBINATIONS AND PERMUTATIONS