LECTURE 6

Probability

Motivation

Remember when we went over the statistics lifecycle:

We've discussed Steps 1 and 2 (Producing Data & EDA), and we're now moving on to Step 3, Probability. Step 4 is the last 3rd of the course.

Why do we use probability?

Example – Who Prefers Fall to Summer?

Group 1:

Group 2:

From this, we can see that there is variation due to the random sample.

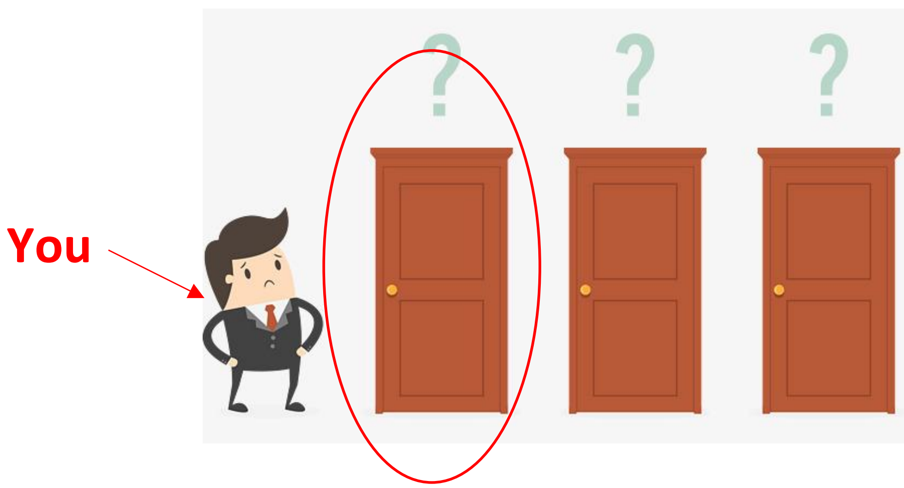
Probability is not always intuitive

Example – Let's Make a Deal
In the 1970's show "Let's Make a Deal," a contestant was first given a choice of 3 doors to select the one they believed the prize lay behind (a goat was behind the 2 other). After the contestant chose which door they thought the prize was behind, the host would then reveal one of the other two doors that contained a goat instead of the prize. Finally, the host would ask if the contestant would like to switch doors to the other remaining door.

The question is: Should the contestant switch? Are the odds of winning higher if he/she switches?

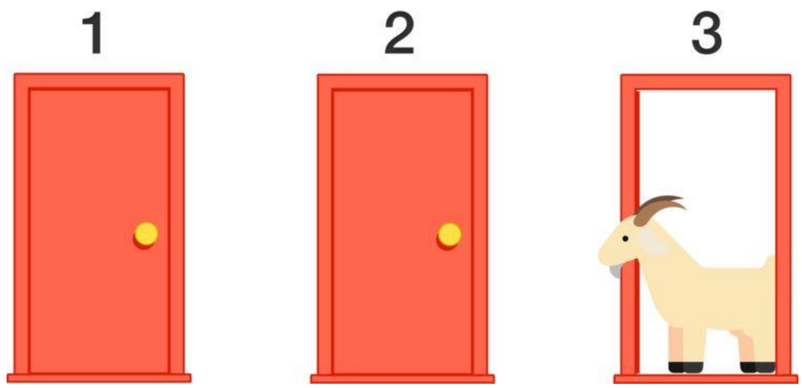
Let's take this step by step…

Step 1: One door has the prize, and the other two have goats. Whatever is behind your door is what you take home. Suppose that you, the contestant, select door 1.



Right now at this step, the chance of the contestant selecting the door with the prize behind it is

Step 2: Now, the Monty Hall (the host) opens one of the other two doors that he knows has a goat behind it , say Door 3, and then asks if you'd like to switch your door selection from Door 1 to Door 2:



However, the odds of Door 2 containing the prize are

Comparing this to your odds at the start, you should

Terminology:

Probability → quantification of randomness and uncertainty

Random Experiment → An experiment that produces an outcome that cannot be determined in advance (i.e. involves uncertainty)

Example 1: Toss a coin once
Example 2: Roll a die once
Example 3: A couple decides to have children until they have one boy and one girl, but no more than 3 children.

Sample space, S, → the list of all possible outcomes of a random experiment

Event → a statement about the nature of the outcome that we're actually going to get once the experiment is conducted

→

→

- Sidenote – Subset →

Set A = {1,2,3,4}
Set B = {2,4}
Set C = {2,3,5}
Set D = {1,2,3,4}

B is a subset of A, but C isn't a subset of A because 5 is not in Set A

Outcome → a possible result of the experiment

So, now let's re-examine Example 1 – One Coin Toss

Looking at just one coin being flipped once, we have two possible events:

Event A:

Event B:

Now, let's consider an example with more event possibilities, Example 2 – One Die Roll

We could have the following events that we are interested in:

Event A:

Event B:

Event C:

Event D:

Example 3 takes a bit more work. Here, we are looking at a couple that decides to have children until they have one boy and one girl, but no more than 3 children.

Let's first look at the sample space for this example:

So, some possible events we could look at are:

Event A:

Event B:

Note how an event is always a subset of the sample space.

How Can We Visualize Probability (and make our lives easier…)

Venn Diagrams are your best friend! ALWAYS DRAW ONE.

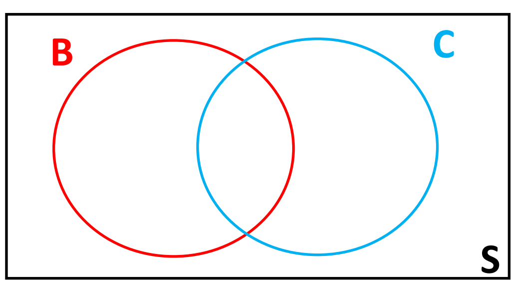
Let's look at Example 2 again:

2 events we were interested in were:

Event B: roll an even number

Event C: roll a number less than or equal to 4

We can visualize this using a Venn Diagram:



What does the overlap imply?

What if there was no overlap?

Now, how does this blend into probability? We can talk about the probability of an event occurring:

P(B) →

P(C) →

The probability associated with an event describes the likelihood of that event occurring, i.e. the outcome is one of the possible values dictated by the event.

If P(B) = 0, then

If P(B) = 1/2, then

If P(B) = 1, then

The Probability Lifecycle:

So, how do we calculate the probability that an event will occur?

I. Relative Frequency Approach

Let's start simple. The probability that you'll get a Head on a coin flip is 1/2. We all know that. How is this calculated, though? It ties into relative frequency.

Remember that we defined relative frequency in Lecture 1 as the percent of all subjects that fall into the category.

Formulaically, this was written as

In the case of probability, we can tweak the definition to write it as

Example – Coin Flips

Trial 1: ___ , ___ , ___ , ___ , ___ , ___ r.f. of Heads =

Trial 2: ___ , ___ , ___ , ___ , ___ , ___ r.f. of Heads =

In theory, we could flip the coin an infinite number of times. The relative frequency of heads will be half the number of flips.

In practice, however, we can't flip a coin infinitely many times. So, we must instead choose what is considered a very large number, 40,000 times for example like a couple of Berkeley undergrads did (https://www.stat.berkeley.edu/~aldous/Real-World/coin_tosses.html) . Out of 40,000, theoretically 20,000 should be heads.

However, the relative frequency approach only provides an estimate for P(A). As you can imagine, if we tossed a coin 40,000 times (1 hour a day for an entire semester according to the Berkeley undergrads…) we wouldn't get exactly 20,000 heads. It would vary due to variability, but the average would be 20,000 if you repeated the experiment a number of times.

Note: We control how good this estimate is by the number of times we repeat the random experiment (increase from 40,000 to 100,000). The more repetitions performed, the closer this relative frequency gets to the true probability P(Head).

II. Equally Likely Outcomes

If all the outcomes in the sample space S are equally likely, then we can find P(A) using the following formula:

Example – Roll a Die
What is the probability that when you roll a die, the number will be less than 3?

Note: It's NOT always the case that all outcomes are equally likely.

Example - Gender of babies
Re-examining Example 3 from above, we said the sample space for this random experiment was

Not each outcome has the same probability!

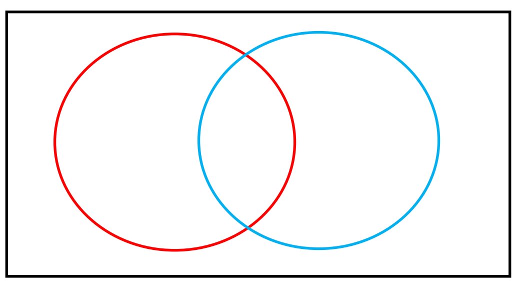
Note: if looking at pairs, order may or may not matter depending on the context.

Example – Job Selection
4 people have applied to work as a Data Scientist at Facebook, 2 males (M1, M2) and 2 females (F1, F2). There are only 2 positions, but their HR department has told those involved with the hiring process that they can't hire 2 people of the same sex. Assuming all candidates are equally qualified and thus equally likely to secure the opening, what is the probability that the HR department's mandate would be violated?

III. Probability Principles (Rules)

1. ***The Complement Principle***

Let's look at the Complement Principle visually first

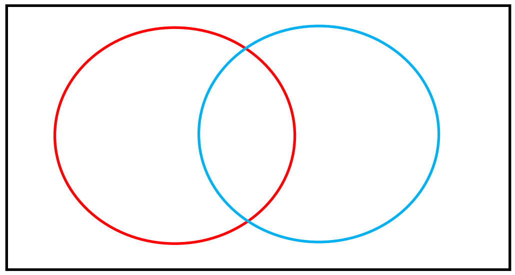


Formulaically:

Note: this rule is especially helpful when finding P(A) is quite complicated or laborious.

2. ***The Intersection of 2 Events***: (which outcomes are in both events)

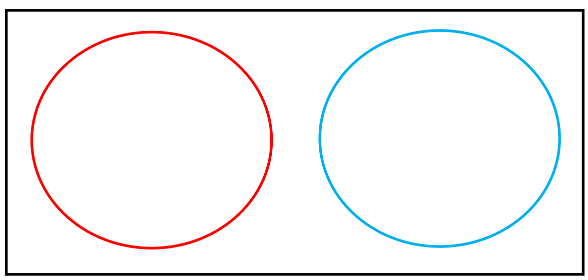
Visually:



Notation:

Note: 2 events are called ***disjoint*** if they don't have any outcomes in common

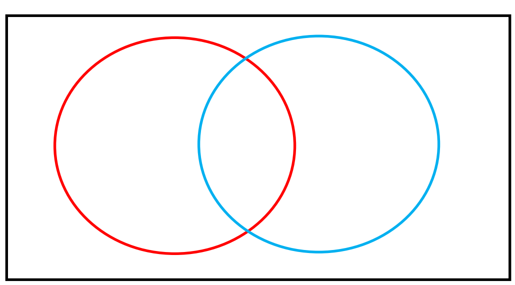
Visually:



Mathematically:

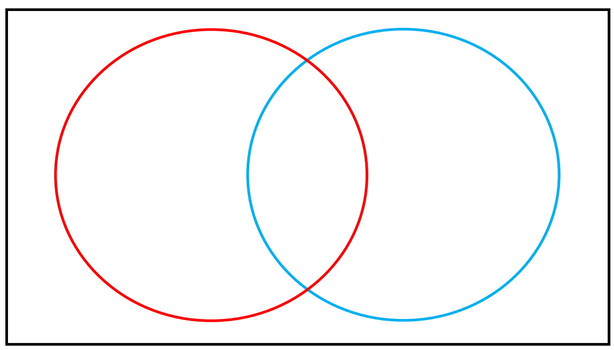
3. ***The Union of 2 Events*** (all the outcomes that are in A or B (or both) → all the outcomes that are in at least one of the 2 events)

Visually:



Notation:

So, let's summarize by labelling each part of the Venn Diagram:



Example – Throwing a Die
Suppose we are only throwing 1 die once. Event A is the event that the roll is less than or equal to 2. Event B is that the roll is odd.

a. Are all the possible outcomes equally likely? What is the sample space?

b. What is P(A)? P(B)?

c. What is the $P(A \cap B)$?

d. What is the $P(A \cup B)$?

e. What is the $P(A^c)$? What does the event A^c actually mean in context?

Example – Symptoms

Suppose that we are looking at a drug to treat high blood pressure. We know that there is a 14% chance of headache, an 11% chance of indigestion, and a 6% chance of getting both.

a. Write down the 3 pieces of information given in terms of probabilities of H and I

b. Summarize the info using a Venn Diagram

c. Are the two events H and I disjoint?

d. What is the probability of experiencing ***only*** Indigestion?

e. What is the probability of experiencing ***only*** headache?

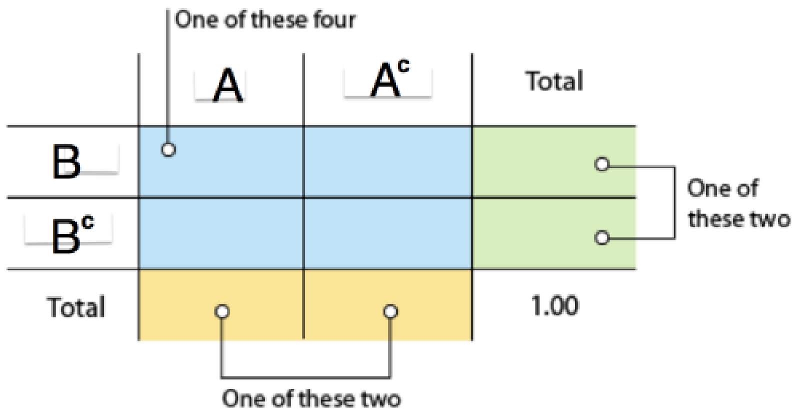
f. What is the probability of experiencing neither?

Another way to dissect and summarize this information is to use something called a probability table:

A fundamental principle used to fill in the probability table is that when there are only 2 events A,B:

To put this in context from our previous example, 14% of all people who use the drug get headaches. Of that 14%, 6% also get indigestion and the other 8% experience only headaches.

The information needed to be able to completely fill in a probability table is as follows:



Example – Package Delivery

As he does with most things, Austin waited until the day before his taxes were due to mail them in. It's vital that the documents reach the destination within one day. To maximize the chances of on-time delivery, two copies of the document are sent using 2 services, Fed-Ex (F) and Amazon's new delivery service Prime Delivery. We know that there is a 92% chance of on-time delivery by Fed-Ex and a 94% chance of on-time delivery by Prime Delivery. We also know there is a 2% chance of on-time delivery by neither Fed-Ex or Prime Delivery.

a) Write down the 3 pieces of given information in terms of probabilities involving event F and event P

b) Summarize the info using a probability table

c) What is the probability that only Prime Delivery is on time? (put this in symbols and then answer using the table)

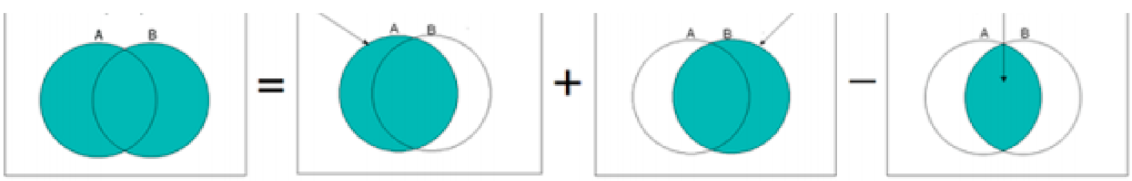
d) What is the probability that both services render on-time delivery? (put this in symbols and then answer using the table)

e) What is the probability that only Fed-Ex renders on-time delivery? (put this in symbols and then answer using the table)

4. The Addition Principle (for finding $P(A \cup B)$)

First, what is $P(A \cup B)$ in words?

Now, visually what is $P(A \cup B)$ using a Venn Diagram?



We therefore get:

P(A or B) = P(A) + P(B) − P(A and B)

So, the formula is

What happens when the 2 events are disjoint?

First off, what does it mean to be disjoint?

So, to find $P(A \cup B)$ in the case where the 2 events are disjoint, we have

Example – Car Ownership
For this example, we are looking at the town of Greenwich, CT and studying the types of cars families own. We know that 15% own a Jeep Wrangler and 21% own an Audi. 3% of families own both a Jeep Wrangler and an Audi (must be nice…).

a. Write down the 3 pieces of info given in terms of the probabilities of events J and A

b. Construct a probability table

c. What is the probability that a family living in Greenwich, CT owns a Jeep Wrangler or an Audi?

Note: There is another way to word the same question: What is the probability that a family living in Greenwich, CT owns at least one of the two types of cars?

d. Now, what is the probability that a family living in Greenwich, CT owns exactly one of the two car types mentioned?

Visually:

Formulaically: