Review of basic probability

We use the following Notation:

- Sample space – Ω
  Elements of Ω are the outcomes of the random experiment
  We write $\Omega = \{\omega_1, \omega_2, \cdots\}$ when it is countable
- An event is, by definition, a subset of Ω
- Set of all possible events – $\mathcal{F} \subset 2^{\Omega}$ (power set of Ω)
  Each event is a subset of Ω
  For now, we take $\mathcal{F} = 2^{\Omega}$ (power set of Ω)

Probability axioms

Probability (or probability measure) is a function that assigns a number to each event and satisfies some properties.
$P : \mathcal{F} \to \Re, \;\; \mathcal{F} \subset 2^{\Omega}$

- A1 $P(A) \geq 0, \, \forall A \in \mathcal{F}$
- A2 $P(\Omega) = 1$
- A3 If $A_i \cap A_j = \phi, \forall i \neq j$ then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Some consequences of the axioms

- $0 \leq P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- If $A \subset B$ then, $P(A) \leq P(B)$
- If $A \subset B$ then, $P(B - A) = P(B) - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Case of finite Ω – Example

- Let $\Omega = \{\omega_1, \cdots, \omega_n\}$, $\mathcal{F} = 2^\Omega$, and P is specified through 'equally likely' assumption.
- That is, $P(\{\omega_i\}) = \frac{1}{n}$. (Note the notation)
- Suppose $A = \{\omega_1, \omega_2, \omega_3\}$. Then

$$P(A) = P(\{\omega_1\} \cup \{\omega_2\} \cup \{\omega_3\}) = \sum_{i=1}^{3} P(\{\omega_i\}) = \frac{3}{n} = \frac{|A|}{|\Omega|}$$

- We can easily see this to be true for any event, A.
- This is the usual familiar formula: number of favourable outcomes by total number of outcomes.
- Thus, 'equally likely' is one way of specifying the probability function (in case of finite Ω).
- An obvious point worth remembering: specifying P for singleton events fixes it for all other events.

Case of Countably infinite Ω

- Let $\Omega = \{\omega_1, \omega_2, \cdots\}$.
- Once again, any $A \subset \Omega$ can be written as mutually exclusive union of singleton sets.
- Let $q_i, i = 1, 2, \cdots$ be numbers such that $q_i \geq 0$ and $\sum_i q_i = 1$.
- We can now set $P(\{\omega_i\}) = q_i, i = 1, 2, \cdots$. (Assumptions on q_i needed to satisfy $P(A) \geq 0$ and $P(\Omega) = 1$).
- This fixes P for all events: $P(A) = \sum_{\omega \in A} P(\{\omega\})$
- This is how we normally define a probability measure on countably infinite Ω.
- This can be done for finite Ω too.

Example: countably infinite Ω

- Consider a random experiment of tossing a biased coin repeatedly till we get a head. We take the outcome of the experiment to be the number of tails we had before the first head.
- Here we have $\Omega = \{0, 1, 2, \cdots\}$.
- A (reasonable) probability assignment is:

$$P(\{k\}) = (1-p)^k p, k = 0, 1, \cdots$$

  where p is the probability of head and $0 < p < 1$.
  (We assume you understand the idea of 'independent' tosses here).
- In the notation of previous slide, $q_i = (1-p)^i p$
- Easy to see we have $q_i \geq 0$ and $\sum_{i=0}^{\infty} q_i = 1$.

Case of uncountably infinite Ω

- We would mostly be considering only the cases where Ω is a subset of $\Re^d$ for some d.
- Note that now an event need not be a countable union of singleton sets.
- For now we would only consider a simple intuitive extension of the 'equally likely' idea.
- Suppose Ω is a finite interval of $\Re$. Then we will take $P(A) = \frac{m(A)}{m(\Omega)}$ where $m(A)$ is length of the set A.
- We can use this in higher dimensions also by taking $m(\cdot)$ to be an appropriate 'measure' of a set.
- For example, in $\Re^2$, $m(A)$ denotes area of A, in $\Re^3$ it would be volume and so on.

  (There are many issues that need more attention here).

Example: Uncountably infinite Ω

Problem: A rod of unit length is broken at two random points. What is the probability that the three pieces so formed would make a triangle.

- Let us take left end of the rod as origin and let x, y denote the two successive points where the rod is broken.
- Then the random experiment is picking two numbers x, y with $0 < x < y < 1$.
- We can take $\Omega = \{(x, y) \ : \ 0 < x < y < 1\} \subset \Re^2$.
- For the pieces to make a triangle, sum of lengths of any two should be more than the third.

- The lengths are: $x, (y-x), (1-y)$. So we need

$$x + (y-x) > (1-y) \;\Rightarrow\; y > 0.5$$

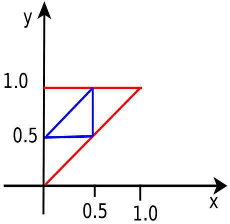
$$x + (1-y) > (y-x) \;\Rightarrow\; y < x + 0.5;$$

$$(y-x) + 1 - y > x \;\Rightarrow\; x < 0.5$$

- So the event of interest is:
$A = \{(x,y) \;:\; y > 0.5;\; x < 0.5;\; y < x + 0.5\}$

- We have

$$\begin{aligned} \Omega &= \{(x, y) \ : \ 0 < x < y < 1\} \\ A &= \{(x, y) \ : \ y > 0.5; \ x < 0.5; \ y < x + 0.5\} \end{aligned}$$

- We can visualize it as follows 

- The required probability is area of A divided by area of Ω which gives the answer as 0.25

- Everything we do in probability theory is always in reference to an underlying probability space: $(\Omega, \mathcal{F}, P)$ where
  - Ω is the sample space
  - $\mathcal{F} \subset 2^{\Omega}$ set of events; each event is a subset of Ω
  - $P : \mathcal{F} \to [0, 1]$ is a probability (measure) that assigns a number between 0 and 1 to every event (satisfying the three axioms).

Conditional Probability

- Let B be an event with $P(B) > 0$. We define conditional probability, conditioned on B, of any event, A, as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}$$

- The above is a notation. "$A \mid$ B" does not represent any set operation! (This is an abuse of notation!)
- Given a B, conditional probability is a new probability assignment to any event.
- That is, given B with $P(B) > 0$, we define a new probability $P_B : \mathcal{F} \to [0, 1]$ by

$$P_B(A) = \frac{P(AB)}{P(B)}$$

- Conditional probability is a probability. What does this mean?
- The new function we defined, $P_B : \mathcal{F} \to [0, 1]$, $P_B(A) = \frac{P(AB)}{P(B)}$, satisfies the three axioms of probability.
- $P_B(A) \geq 0$ and $P_B(\Omega) = 1$.
- If A_1, A_2 are mutually exclusive then $A_1 B$ and $A_2 B$ are also mutually exclusive and hence

$$\begin{aligned} P_B(A_1 + A_2) &= \frac{P\left((A_1 + A_2)B\right)}{P(B)} = \frac{P(A_1 B + A_2 B)}{P(B)} \\ &= \frac{P(A_1 B) + P(A_2 B)}{P(B)} = P_B(A_1) + P_B(A_2) \end{aligned}$$

- Once we understand condional probability is a new probability assignment, we go back to the 'standard notation'

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

- Note $P(B|B) = 1$ and $P(A|B) > 0$ only if $P(AB) > 0$.
- Now the 'new' probability of each event is determined by what it has in common with B.
- If we know the event B has occurred, then based on this knowledge we can readjust probabilities of all events and that is given by the conditional probability.
- Intuitively it is as if the sample space is now reduced to B because we are given the information that B has occurred.
- This is a useful intuition as long as we understand it properly.
- It is not as if we talk about conditional probability only for subsets of B. Conditional probability is also with respect to the original probability space. Every element of $\mathcal{F}$ has conditional probability defined.

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

- Suppose $P(A \mid B) > P(A)$
  Does it mean "B **causes** A"?

$$\begin{aligned} P(A \mid B) > P(A) &\Rightarrow P(AB) > P(A)P(B) \\ &\Rightarrow \frac{P(AB)}{P(A)} > P(B) \\ &\Rightarrow P(B \mid A) > P(B) \end{aligned}$$

- Hence, conditional probabilities cannot actually capture causal influences.
- There are probabilistic methods to capture causation (but far beyond the scope of this course!)

- In a conditional probability, the conditioning event can be any event (with positive probability)
- In particular, it could be intersection of events.
- We think of that as conditioning on multiple events.

$$P(A \mid B, C) = P(A \mid BC) = \frac{P(ABC)}{P(BC)}$$

- The conditional probability is defined by

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

- This gives us a useful identity

$$P(AB) = P(A \mid B)P(B)$$

- We can iterate this for multiple events

$$P(ABC) = P(A \mid BC)P(BC) = P(A \mid BC)P(B \mid C)P(C)$$

- Let $B_1, \cdots, B_m$ be events such that $\cup_{i=1}^{m} B_i = \Omega$ and $B_i B_j = \phi, \forall i \neq j$.
- Such a collection of events is said to be a partition of Ω. (They are also sometimes said to be mutually exclusive and collectively exhaustive).
- Given this partition, any other event can be represented as a mutually exclusive union as

$$A = AB_1 + \cdots + AB_m$$

To explain the notation again

$$A = A \cap \Omega = A \cap (B_1 \cup \cdots \cup B_m) = (A \cap B_1) \cup \cdots \cup (A \cap B_m)$$

Hence, $A = AB_1 + \cdots + AB_m$

Total Probability rule

- Let $B_1, \cdots, B_m$ be a partition of Ω.
- Then, for any event A, we have

$$\begin{aligned} P(A) &= P(AB_1 + \cdots + AB_m) \\ &= P(AB_1) + \cdots + P(AB_m) \\ &= P(A\,|B_1)P(B_1) + \cdots + P(A\,|B_m)P(B_m) \end{aligned}$$

- The formula (where B_i form a partition)

$$P(A) = \sum_i P(A \mid B_i)P(B_i)$$

  is known as **total probability rule** or total probability law or total probability formula.
- This is a very useful in many situations. ("arguing by cases")

Example: Polya's Urn

An urn contains r red balls and b black balls. We draw a ball at random, note its color, and put back that ball along with c balls of the same color. We keep repeating this process. Let R_n (B_n) denote the event of drawing a red (black) ball at the n^{th} draw. We want to calculate the probabilities of all these events.

- It is easy to see that $P(R_1) = \frac{r}{r+b}$ and $P(B_1) = \frac{b}{r+b}$.
- For R_2 we have, using total probability rule,

$$\begin{aligned} P(R_2) &= P(R_2 \mid R_1)P(R_1) + P(R_2 \mid B_1)P(B_1) \\ &= \frac{r+c}{r+c+b}\,\frac{r}{r+b} + \frac{r}{r+b+c}\,\frac{b}{r+b} \\ &= \frac{r(r+c+b)}{(r+c+b)(r+b)} = \frac{r}{r+b} = P(R_1) \end{aligned}$$

- Similarly we can show that $P(B_2) = P(B_1)$.
- One can show by mathematical induction that $P(R_n) = P(R_1)$ and $P(B_n) = P(B_1)$ forall n. (Left as an exercise for you!)
- This does not depend on the value of c!

Bayes Rule

- Another important formula based on conditional probability is Bayes Rule:

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

- This allows one to calculate $P(A \mid B)$ if we know $P(B \mid A)$.
- Useful in many applications because one conditional probability may be more easier to obtain (or estimate) than the other.
- Often one uses total probability rule to calculate the denominator in the RHS above:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}$$

Example: Bayes Rule

Let D and D^c denote someone being diagnosed as having a disease or not having it. Let T_+ and T_- denote the events of a test for it being positive or negative. (Note that $T_+^c = T_-$). We want to calculate $P(D|T_+)$.

- We have, by Bayes rule,

$$P(D|T_+) = \frac{P(T_+|D)P(D)}{P(T_+|D)P(D) + P(T_+|D^c)P(D^c)}$$

- The probabilities $P(T_+|D)$ and $P(T_+|D^c)$ can be obtained through, for example, laboratory experiments.
- $P(T_+|D)$ is called the true positive rate and $P(T_+|D^c)$ is called false positive rate.
- We also need $P(D)$, the probability of a random person having the disease.

- Let us take some specific numbers
- Let: $P(D) = 0.5$, $P(T_+|D) = 0.99$, $P(T_+|D^c) = 0.05$.

$$P(D|T_+) = \frac{0.99 * 0.5}{0.99 * 0.5 + 0.05 * 0.5} = 0.95$$

That is pretty good.

- But taking $P(D) = 0.5$ is not realistic. Let us take $P(D) = 0.1$.

$$P(D|T_+) = \frac{0.99 * 0.1}{0.99 * 0.1 + 0.05 * 0.9} = 0.69$$

- Now suppose we can improve the test so that $P(T_+|D^c) = 0.01$

$$P(D|T_+) = \frac{0.99 * 0.1}{0.99 * 0.1 + 0.01 * 0.9} = 0.92$$

- These different cases are important in understanding the role of false positives rate.

- $P(D)$ is the probability that a random person has the disease. We call it the prior probability.
- $P(D|T_+)$ is the probability of the random person having disease once we do a test and it came positive. We call it the posterior probability.
- Bayes rule essentially transforms the prior probability to posterior probability.

- In many applications of Bayes rule the same generic situation exists
- Based on a measurement we want to predict (what may be called) the state of nature.
- For another example, take a simple communication system.
  - D can represent the event that the transmitter sent bit 1.
  - T_+ can represent an event about the measurement we made at the receiver.
  - We want the probability that bit 1 is sent based on the measurement.
  - The knowledge we need is $P(T_+|D)$, $P(T_+|D^c)$ which can be determined through experiment or modelling of channel.

$$P(D|T_+) = \frac{P(T_+|D)P(D)}{P(T_+|D)P(D) + P(T_+|D^c)P(D^c)}$$

- Not all applications of Bayes rule involve a 'binary' situation
- Suppose D_1, D_2, D_3 are the (exclusive) possibilities and T is an event about a measurement.

$$\begin{aligned} P(D_1|T) &= \frac{P(T|D_1)P(D_1)}{P(T)} \\ &= \frac{P(T|D_1)P(D_1)}{P(T|D_1)P(D_1) + P(T|D_2)P(D_2) + P(T|D_3)P(D_3)} \\ &= \frac{P(T|D_1)P(D_1)}{\sum_i P(T|D_i)P(D_i)} \end{aligned}$$

$$P(D|T_+) = \frac{P(T_+|D)P(D)}{P(T_+|D)P(D) + P(T_+|D^c)P(D^c)}$$

- In the binary situation we can think of Bayes rule in a slightly modified form too.

$$\frac{P(D|T_+)}{P(D^c|T_+)} = \frac{P(T_+|D)}{P(T_+|D^c)} \frac{P(D)}{P(D^c)}$$

- This is called the odds-likelihood form of Bayes rule (The ratio of $P(A)$ to $P(A^c)$ is called odds for A)

Independent Events

- Two events A, B are said to be independent if

$$P(AB) = P(A)P(B)$$

- Note that this is a definition. Two events are independent if and only if they satisfy the above.
- Suppose $P(A), P(B) > 0$. Then, if they are independent

$$P(A|B) = \frac{P(AB)}{P(B)} = P(A); \text{ similarly } P(B|A) = P(B)$$

- This gives an intuitive feel for independence.
- Independence is an important (often confusing!) concept.

Example: Independence

A class has 20 female and 30 male course (MTech) students and 6 female and 9 male research (PhD) students. Are gender and degree independent?

- Let F, M, C, R denote events of female, male, course, research students
- From the given numbers, we can easily calculate the following:

$$P(F) = \frac{26}{65} = \frac{2}{5};\ P(C) = \frac{50}{65} = \frac{10}{13};\ P(FC) = \frac{20}{65} = \frac{4}{13}$$

- Hence we can verify

$$P(F)P(C) = \frac{2}{5}\,\frac{10}{13} = \frac{4}{13} = P(FC)$$

and conclude that F and C are independent.
Similarly we can show for others.

- In this example, if we keep all other numbers same but change the number of male research students to, say, 12 then the independence no longer holds.
$(\frac{26}{68}\ \frac{50}{68} \neq \frac{20}{68})$
- One needs to be careful about independence!
- We always have an underlying probability space $(\Omega, \mathcal{F}, P)$
- Once that is given, the probabilities of all events are fixed.
- Hence whether or not two events are independent is a matter of 'calculation'

- If A and B are independent then so are A and B^c.
- Using $A = AB + AB^c$, we have

$$P(AB^c) = P(A) - P(AB) = P(A)(1 - P(B)) = P(A)P(B^c)$$

- This also shows that A^c and B are independent and so are A^c and B^c.
- For example, in the previous problem, once we saw that F and C are independent, we can conclude M and C are also independent (because in this example we are taking $F^c = M$).

- Consider the random experiment of tossing two fair coins (or tossing a coin twice).
- $\Omega = \{HH, HT, TH, TT\}$.
  Suppose we employ 'equally likely idea'.
- That is, $P(\{HH\}) = \frac{1}{4}$, $P(\{HT\}) = \frac{1}{4}$ and so on
- Let $A =$ 'H on 1st toss' $= \{HH, HT\}$ $(P(A) = \frac{1}{2})$
  Let $B =$ 'T on second toss' $= \{HT, TT\}$ $(P(B) = \frac{1}{2})$
- We have $P(AB) = P(\{HT\}) = 0.25$
- Since $P(A)P(B) = \frac{1}{2}\,\frac{1}{2} = \frac{1}{4} = P(AB)$,
  A, B are independent.
- Hence, in multiple tosses, assuming all outcomes are equally likely implies outcome of one toss is independent of another.

- In multiple tosses, assuming all outcomes are equally likely is alright if the coin is fair.
- Suppose we toss a biased coin two times.
- Then the four outcomes are, obviously, not ‘equally likely’
- How should we then assign these probabilities?
- If we assume tosses are independent then we can assign probabilities easily.

- Consider toss of a biased coin:
  $\Omega^1 = \{H, T\}$, $P(\{H\}) = p$ and $P(\{T\}) = 1 - p$.
- If we toss this twice then $\Omega^2 = \{HH, HT, TH, TT\}$ and we assign
  $P(\{HH\}) = p^2$, $P(\{HT\}) = p(1-p)$,
  $P(\{TH\}) = (1-p)p$, $P(\{TT\}) = (1-p)^2$.
- $P(\{HH, HT\}) = p^2 + p(1-p) = p$
- This assignment ensures that $P(\{HH\})$ equals product of probability of H on 1st toss and H on second toss.
- Essentially, Ω^2 is a cartesian product of Ω^1 with itself and essentially we used products of the corresponding probabilities.
- For any independent repetitions of a random experiment we follow this.
  (We will look at it more formally when we consider multiple random variables).

- In many situations calculating probabilities of intersection of events is difficult.
- One often **assumes** A and B are independent to calculate $P(AB)$.
- As we saw, if A and B are independent, then $P(A|B) = P(A)$
- This is often used, at an intuitive level, to justify assumption of independence.

Independence of multiple events

- Events $A_1, A_2, \cdots, A_n$ are said to be (totally) independent if for any k, $1 \leq k \leq n$, and any indices $i_1, \cdots, i_k$, we have

$$P(A_{i_1} \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

- For example, A, B, C are independent if

$$P(AB) = P(A)P(B);\ P(AC) = P(A)P(C);$$

$$P(BC) = P(B)P(C);\ P(ABC) = P(A)P(B)P(C)$$

Pair-wise independence

- Events $A_1, A_2, \cdots, A_n$ are said to be pair-wise independent if

$$P(A_i A_j) = P(A_i)P(A_j), \ \forall i \neq j$$

- Events may be pair-wise independent but not (totally) independent.
- Example: Four balls in a box inscribed with '1', '2', '3' and '123'. Let E_i be the event that number 'i' appears on a radomly drawn ball, $i = 1, 2, 3$.
- Easy to see: $P(E_i) = 0.5, \ i = 1, 2, 3$.
- $P(E_i E_j) = 0.25 \ (i \neq j) \Rightarrow$ pairwise independent
- But, $P(E_1 E_2 E_3) = 0.25 \neq (0.5)^3$