1 Basic definitions:

- **Random Variable**: A variable with a stochastic outcome/value.
- **Event**: A set of one or more possible outcomes/values.
- **Probability**: Chance of an event occurring.
- **Distribution**: A mapping from outcomes to probability masses or densities (more below).

**Example**

Let’s consider the scenario of rolling a dice. The (future) outcome of the dice is a random variable as it is currently unknown. Numbers \{1, 2, 3, 4, 5, 6\} are the possible outcomes. Note that outcomes are **mutually exclusive**; we cannot get a 2 and a 4 in a single throw. An example of an event is rolling a number greater than 3, or rolling an even number. Note that events are not necessarily mutually exclusive; in a single row we can get a number that is both even and greater than 3.

Most scenarios in real life involve more than one random variable. Consider the scenario of rolling a dice again \((X)\), and tossing a coin \((Y)\) at the same time. In this scenario, we can compute the probability of e.g. getting a 4 and Head: \(P(X = 4, Y = H)\). This is known as the **joint probability** of \(X\) and \(Y\): \(P(X = x, Y = y)\)

2 Probability Rules:

When it comes to probability, there are basically two main rules that you should learn by heart: **product rule** and **sum rule**. The product rule states that the probability of two events occurring is the probability of one of the events occurring times the conditional probability of the other event happening given the first event happened:

\[
P(A, B) = P(A|B)P(B) = P(B|A)P(A) \tag{1}
\]

Where \(P(A,B)\) denotes the probability that both the events \(A\) and \(B\) occur. We can write the product rule in another format known as the **conditional probability**. In product rule, if you divide both sides by the conditional probability, you get:

\[
P(A|B) = \frac{P(A, B)}{P(B)} \tag{2}
\]
Now we get to sum rule. The most general form of sum rule is defined below:

\[ P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A, B) \] (3)

A corollary of the sum rule (which we will be using in this course a lot) is that the marginal marginal \( P(Y = y) \) is related to the joint \( P(Y = y, X = x) \) as:

\[ P(Y = y) = \sum_{x \in X} P(Y = y, X = x) \] (4)

\[ P(Y = y) \approx 0.7 \] (5)

**Example:**

Let's say the probability of someone random being male is 0.5 and the probability of them having short hair is 0.6 (since majority of people have short hair). Let's also say the probability of someone random both being male and having short hair is 0.4. Using sum rule, we can compute the probability of being male or having short hair:

\[ P(\text{male } \cup \text{ shorthair}) = P(\text{male}) + P(\text{shorthair}) - P(\text{male } \cap \text{ shorthair}) \]
\[ = 0.5 + 0.6 - 0.4 \]
\[ = 0.7 \]

**Bayes Rule:** This is central formula in Bayesian Machine Learning. If you combine the product rule and conditional probability rule, you get:

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \] (6)

**Example: Bayesian Trap**

Let's say you wake up one day and you feel sick, so you go to the doctor. The doctor is also unsure about what is going on so they suggest you run a series of tests. The results come back and it turns out you tested positive for a very rare disease that affects only 0.01% of the population. You ask the doctor about how accurate this test really is, and the doctor says this test can correctly identify 99% of the people who have the disease, and only incorrectly identify 1% of the people who don‘t have the disease. Now, before you panic, you remember Bayes Rule, and you use to find out how likely it is to have this disease:

Let's denote \( A \) and \( B \) as the events the we have the disease and the test was positive respectively. Given the information above, we have

\( P(A) = 0.0001, P(B|A) = 0.99, P(B|\neg A) = 0.01 \)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]
\[ = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)} \]
\[ = \frac{0.99 \times 0.0001}{(0.99 \times 0.0001) + (0.01 \times 0.9999)} \]
\[ \approx 0.01 \]
Probability Density Functions: The examples above are all discrete. Unfortunately, most real life scenarios are in continuous space. Consider the scenario of a runner doing a Marathon. What is the probability of the runner finishing after exactly π hours. The answer is obviously 0. In fact, this probability for any value is 0, since there are infinite possible outcomes. However, remember that we can define the probabilities of events: say the probability of the runner finishing between 3 to 4 hours. This probability is clearly not 0. As a result of this, probability in continuous space has been characterized with functions known as probability density functions (pdf) which reflect the relative probability of areas of the space of values. The pdf is the density of probability rather than the probability mass. We can informally define it by considering the probability of being some small area of the space of size $2\delta x$. We can assume that probability density is roughly constant across within $2\delta$ since $\delta$ is very small:

$$p_X(x) = \lim_{\delta \to \infty} \frac{P((x-\delta) < X < (x+\delta))}{2\delta} \quad (7)$$

Note that a probability density, unlike probability, can take positive values greater than 1.

3 Measure Theory:

A probability measure is defined as a triple $(\Omega, F, P)$. $\Omega$ is a set of possible outcomes (in continuous case any value between $-\infty$ and $+\infty$). $F$ is the set of possible events, informally the set of possible subsets of outcomes. Finally, $P$ is a probability measure.

A measure can be thought of as something that assigns a size to a set of objects. A probability measure is a measure that assigns probabilities to events; mapping an event $E$ to $[0, 1]$. A probability measure $P(E)$ must satisfies three conditions:

- Probability of the event of entire sample space $\Omega$ is equal to 1: $P(\Omega) = 1$
- Probability of an empty set is equal to 0: $P(\emptyset) = 0$
- The sum rule should be satisfied: $P(\bigcup_{i=1}^{N} E_i) = \sum_{i=1}^{n} P(E_i)$, for any set of disjoint events $E_1, \ldots, E_n$.

A probability measure is defined as an integral with respect to a reference measure $\mu$:

$$P(A) := \int_{x \in A} p_X(x)d\mu(x) \quad \text{(Notation in statistics)}$$

$$P(A) := \int_{A} p(x)dx \quad \text{(Informal notation, reference measure is implied)}$$

There are the 3 common measures you might need to know:

- Lebesgue measure: Most common measure when you are working with real numbers. If you define a measure on closed interval $[a, b]$, its probability measure is width of the interval: $\mu([a, b]) = b - a$
- Counting measure: Say we a finite set of events. A counting measure simply counts the number of elements in the set $\mu(\{x_1, x_2, \ldots, x_n\}) = n$
- Product measure: Suppose we have an event $E = (E_1, E_2) \in \Omega_1 \times \Omega_2$ in the Cartesian product of two sample spaces $\Omega_1$ and $\Omega_2$. Product measure is defined as: $\mu(E) = \mu_1(E_1)\mu_2(E_2)$. 


4 Expectation and Variance:

The expected value $E[X]$, or mean, of a random variable $X$ is the average value that the variable will take if an infinite number of independent draws are made. We can define it as:

$$E[X] = \sum_{x \in \text{dom}(X)} p(x)x$$  \hspace{1cm} \text{(Discrete case)}

$$= \int_{x \in \text{dom}(X)} p(x)x \, dx$$  \hspace{1cm} \text{(Continuous case)}

Normally, what we are interested in is not the expectation of $X$ itself, but some quantity of interest $f$ which depends on $X$:

$$E[f(X)] = \int_{x \in \text{dom}(X)} p(x)f(x) \, dx$$  \hspace{1cm} \text{(8)}$$

We can also define expectation over more than one random variables:

$$E[f(X,Y)] = \int_{x \in \text{dom}(X)} \int_{y \in \text{dom}(Y)} p(x,y)f(x,y) \, dy \, dx$$  \hspace{1cm} \text{(9)}$$

If we wish to average only with respect to part of the randomness in a system, we can instead use a conditional expectation for example:

$$E[f(X,Y) | Y = y] = \int_{x \in \text{dom}(X)} p(x|y)f(x,y) \, dx$$  \hspace{1cm} \text{(10)}$$

Note: Soon, you’ll find out we can (and should) consider more than one distributions for the same random variable (because we are machine learning people!) For example, we can assume the random variable $X$ has a probability distribution $P_X(x)$ or $Q_X(x)$ (under two different models of course). It is important to understand that expectation of some function of random variable under different distribution is not the same:

$$E_{X \sim p(x)}[f(X)] \neq E_{X \sim q(x)}[f(X)]$$

Therefore, when writing expectations, I would encourage you all to write the distribution under which you are computing the expectation. You can use the notation $E_{X \sim p(x)}[f(X)]$ or $E_{p(x)}[f(x)]$

The central problem that we face many times in this course is computing the conditional expectation $E_{p(x|y)}[f(x,y)]$, essentially, the framework for this problem is consisted via three things:

1. Some unknown variable(s) $X$
2. Some observed variable(s) $Y$
3. Some quantity of interest that we find to find out $f$, which may depend on both $X$ and $Y$

Example:

Consider the application of self-driving cars. In this case, $Y$ can be the past trajectories of the car itself, the cars around, and the people around the street. $X$ can be the future trajectories. As mentioned above, $f$ is what we are trying to find out in order to help us make a decision. For example, it can indicate: "will a pedestrian cross the street now?" which can be used to determine whether the car needs to hit break or not.
Example:
Let's say someone gives you a coin and asks you to test whether the coin is fair. What you should do is first define a prior, representing what you believe about the fairness of the coin. Tossing a coin gives you either heads or tails, so the distribution to use is a Bernoulli distribution:

\[ p(y) = \begin{cases} 
  x & \text{if } y = 1 \\
  1 - x & \text{if } y = 0 
\end{cases} \]

Where 1 means heads and 0 means tails. Note that we can also write this in the following way:

\[ P(y) = x^y(1-x)^{1-y} \]

The parameter of the Bernoulli distribution lies in the interval [0, 1]. It is very common to use Beta distribution for such cases. Beta distribution domain also lies between 0 and 1 (you can think about it in this way: it represents a distribution of probabilities). It has to parameters \((\alpha, \beta)\), and the following pdf:

\[ p(x) = \text{Beta}(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)} \]

Where \(B(\alpha, \beta)\) is the normalization constant defined as \(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\). You then toss the coin, say \(N\) times (independently) to collect data: \(\{y_1, y_2, \ldots, y_n\}\). Finally, in order to reason about the fairness of the coin given your observations and your prior belief, you apply Bayes rule:

\[
p(x) = \frac{p(\{y_i\}_{i=1}^n | x) p(x)}{p(\{y_i\}_{i=1}^n)} \\
= \prod_{i=1}^n p(x_i | x) p(x) \\
= \prod_{i=1}^n x^{y_i}(1-x)^{1-y_i} \frac{1}{B(\alpha, \beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)} \\
= \frac{1}{B(\alpha, \beta)} x^{(\alpha-1)+\sum_i y_i} (1-x)^{(\beta-1)+(n-\sum_i y_i)}
\]

Note that the resulting distribution has the form of beta distribution, with parameters \(\alpha' = \alpha + \sum_i y_i\) and \(\beta' = \beta + (n - \sum_i y_i)\). Since we defined \(y = 1\) as head and \(y = 0\) as tail, the sum of all observations \(\sum_i y_i\) is the total number of heads we got in our experiment. Similarly, \((n-\sum_i y_i)\) is the total number of tails. Denoting \(n_H\) and \(n_T\) as the total number of heads and tails respectively, we can define the posterior distribution \(p(x | \{y_i\}_{i=1}^n)\) as:

\[ p(x | \{y_i\}_{i=1}^n) = \text{Beta}(x; \alpha + n_H, \beta + n_T) \]

This is pretty much how it works when you are working in a Bayesian framework. You have some initial belief about the unknown, you observe some data, you then update your belief based on those observations.

**Bayesian Prediction:** A lot of the times, we are interested in reasoning about future data. For example, we would like to know the distribution of a future toss \(y^*\) given the observations and our initial belief: \(p(y^* | \{y_i\}_{i=1}^n, \alpha, \beta)\). This is known as posterior predictive distribution:
\[ p(y^*|\{y_i\}_{i=1}^n) = \int_x p(y^*, x|\{y_i\}_{i=1}^n) \, dx \quad \text{(Sum rule)} \]

\[ = \int_x p(y^*|x)p(x|\{y_i\}_{i=1}^n) \, dx \quad \text{(Product rule)} \]

\[ = \int_x y^*(1-x)(1-y^*) \text{Beta}(x; \alpha + n_H, \beta + n_T) \, dx \]

Since \( y^* \) can only take 2 values (0 or 1), we can just compute the probability for one of them say \( p(y^* = 1) \) and compute the probability of other case by \( 1 - p(y^* = 1) \). If we substitute \( y^* \) by 0 in the expression above we get:

\[ p(y^* = 1|\{y_i\}_{i=1}^n) = \int_x x^1(1-x)^{(1-1)} \text{Beta}(x; \alpha + n_H, \beta + n_T) \, dx \]

\[ = \int_x \text{Beta}(x; \alpha + n_H, \beta + n_T) \, dx \]

\[ = E_{X\sim\text{Beta}(\alpha+n_H,\beta+n_T)}[X] \]

You do not need to memorize the expectation and variance of all distributions (use Wikipedia!), but you should know how to derive them. The expected value of a random variable under the distribution \( \text{Beta}(\alpha, \beta) \) is \( \frac{\alpha}{\alpha + \beta} \). So the probability of the next toss being Head is \( \frac{\alpha + n_H}{\alpha + \beta + n_H + n_T} \).

\[ p(y^* = 1|\{y_i\}_{i=1}^n) = \begin{cases} \frac{\alpha + n_H}{\alpha + \beta + n_H} & \text{if } y = 1 \\ \frac{\beta + n_T}{\alpha + \beta + n_H + n_T} & \text{if } y = 0 \end{cases} \]

We can also compare that with prior predictive distribution \( p(y^*|\alpha, \beta) \), which is the distribution of a next toss over the prior (before observing any data):

\[ p(y^*|\{y_i\}_{i=1}^n) = \begin{cases} \frac{\alpha}{\alpha + \beta} & \text{if } x = 1 \\ \frac{\beta}{\alpha + \beta} & \text{if } x = 0 \end{cases} \]

You can notice that how we updated our belief based on the number of Heads and Tails we got in our experiment.

There are a couple of important concepts/vocabulary that you should keep in mind during the course:

**Conjugacy:** Conjugacy means that the prior-likelihood combination gives a posterior that is of the same form as the prior distribution. In the coin example: \( (\text{Beta} \times \text{Bernoulli} \rightarrow \text{Beta}) \). This is quite a desirable property (you will appreciate it a lot more when we reach sampling). Not only you get a closed-form expression for the posterior distribution, but you can also visualize how observing the data updated the prior distribution.

**Sufficient statistics:** You might have noticed that when doing inference about the outcome of a future toss in the coin example, we did not need to know the outcome of every single toss individually. All we needed to know in the end was either the total number of heads or the total...
number of tails (we could calculate the other one subtracting \( n \)). We say that \( n_H \) or \( n_T \) is the sufficient statistics for \( x \) in this model.

**Posterior, Likelihood, Prior, Model Evidence:** This was already discussed above, but just to make it clear, I will re-write them again. In general, if you have observed data \( y \), and you want to inference on your model parameter \( \theta \), Bays rule gives you:

\[
p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}
\]

(11)

\[
\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Model Evidence}}
\]

(12)

**Likelihood:** How likely is the observed data given a certain model parameter

**Prior:** Represent your initial belief about what the value of your model parameters should be

**Model Evidence/Marginal Likelihood:** The is the distribution of the data marginalized over the model parameters. Expanding the expression gives us: \( p(y) = \int \theta p(y \mid \theta)p(\theta) d\theta \). This integral is almost always quite hard to compute. Fortunately, we do not necessarily need to know this in order perform inference, because model evidence is constant. All it does is basically ensures that the posterior is a probability (the integral sums up to 1). This is why in a lot of papers you just see:

\[
\text{Posterior} \propto \text{Likelihood} \times \text{Prior}
\]

(13)