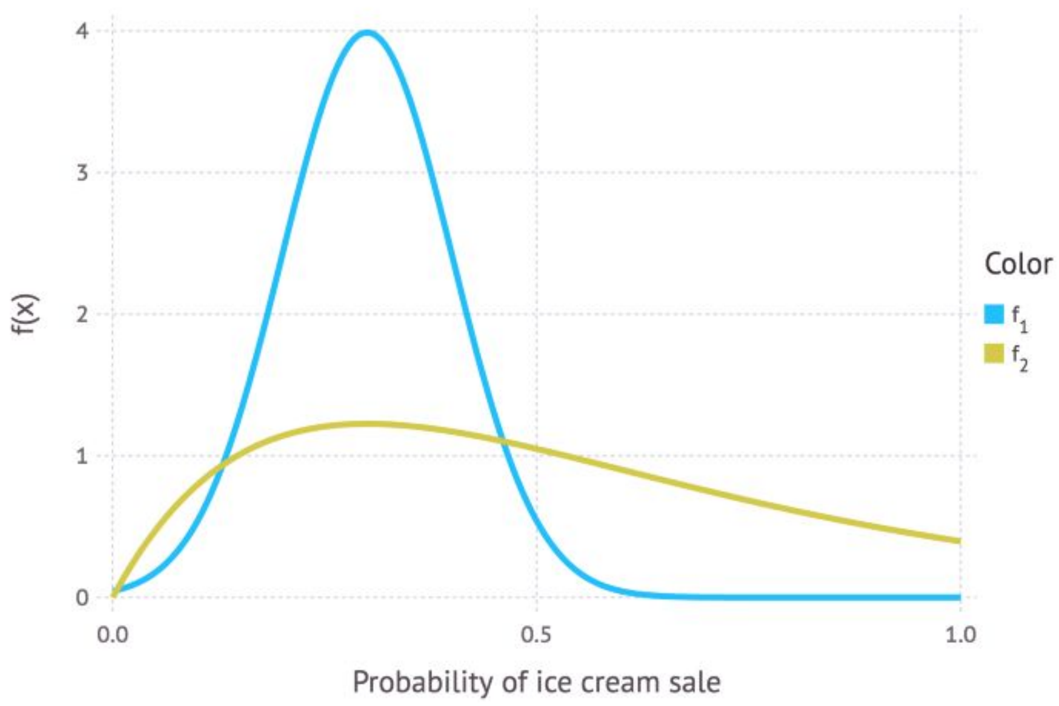
Bayesian inference

- **Bayesian inference** is therefore just the process of deducing properties about a population or probability distribution from data *using Bayes' theorem.*
- An important part of *bayesian inference* is the establishment of *parameters* and *models.*
- Models are the mathematical formulation of the observed events. Parameters are the factors in the models affecting the observed data. For example, in tossing a coin, **fairness of coin** may be defined as the parameter of coin denoted by θ. The outcome of the events may be denoted by D.

How does Bayes' Theorem allow us to incorporate prior beliefs?

- Let A represent the event that we sell ice cream and B be the event of the weather. Then we might ask *what is the probability of selling ice cream on any given day given the type of weather?* Mathematically this is written as P(A=ice cream sale | B = type of weather) which is equivalent to the left hand side of the equation.
- P(A) on the right hand side is the expression that is known as the **prior.** In our example this is P(A = ice cream sale), i.e. the (marginal) probability of selling ice cream regardless of the type of weather outside. P(A) is known as the prior because we might already know the marginal probability of the sale of ice cream.

- For example, 30 people out of a potential 100 actually bought ice cream at some shop somewhere. So my P(A = ice cream sale) = 30/100 = 0.3, *prior to me knowing anything about the weather*. This is how Bayes' Theorem allows us to incorporate prior information.
- The prior probability of selling ice cream was 0.3. However, what if 0.3 was just my best guess but I was a bit uncertain about this value. The probability could also be 0.25 or 0.4. In this case a distribution of our prior belief might be more appropriate (see figure below). This distribution is known as the **prior distribution**.



- 2 distributions that represent our prior probability of selling ice Cream on any given day. The peak value of both the blue and gold curves occur around the value of 0.3 which, as we said above, is our best guess of our prior probability of selling ice cream.
- The fact that f(x) is non-zero of other values of x shows that we’re not completely certain that 0.3 is the true value of selling ice cream. The blue curve shows that it’s likely to be anywhere between 0 and 0.5, whereas the gold curve shows that it’s likely to be anywhere between 0 and 1.

- Instead of event A, we’ll typically see Θ, this symbol is called Theta. Theta is what we’re interested in, it represents the set of parameters. So if we’re trying to estimate the parameter values of a Gaussian distribution then Θ represents both the mean, μ and the standard deviation, σ (written mathematically as $\Theta = \{\mu, \sigma\}$).
- Instead of event B, we’ll see *data* or $\boldsymbol{y} = \{y1, y2, ..., yn\}$. These represent the data, i.e. the set of observations that we have.
- So now Bayes’ theorem in model form is written as:
- $P(\theta|D)=(P(D|\theta) X P(\theta))/P(D)$

- We‘ve seen that P(Θ) is the prior distribution. It represents our beliefs about the true value of the parameters, just like we had distributions representing our belief about the probability of selling ice cream.
- *P(*Θ|*data)* on the left hand side is known as the **posterior distribution.** This is the distribution representing our belief about the parameter values after we have calculated everything on the right hand side taking the observed data into account.

- $P(data|\,\Theta)$ is something we've come across before. If you made it to on maximum likelihood then you'll remember that we said $L(data;\ \mu,\ \sigma)$ is the likelihood distribution (for a Gaussian distribution). Well $P(data|\,\Theta)$ is exactly this, it's the **likelihood distribution** in disguise. Sometimes it's written as $\mathcal{L}(\Theta;\ data)$ but it's the same thing here**.**
- Therefore we can calculate the *posterior distribution* of our parameters using our *prior beliefs* updated with our *likelihood.*

- We should be more interested in knowing : Given an outcome (D) what is the probbaility of coin being fair (θ=0.5)

- Lets represent it using Bayes Theorem:

- P(θ|D)=(P(D|θ) X P(θ))/P(D)

- Here, P(θ) is the prior i.e the strength of our belief in the fairness of coin before the toss. It is perfectly okay to believe that coin can have any degree of fairness between 0 and 1.

- P(D|θ) is the likelihood of observing our result given our distribution for θ. If we knew that coin was fair, this gives the probability of observing the number of heads in a particular number of flips.

- P(D) is the evidence. This is the probability of data as determined by summing (or integrating) across all possible values of θ, weighted by how strongly we believe in those particular values of θ.

- If we had multiple views of what the fairness of the coin is (but didn't know for sure), then this tells us the probability of seeing a certain sequence of flips for all possibilities of our belief in the coin's fairness.

- P(θ|D) is the posterior belief of our parameters after observing the evidence i.e the number of heads .

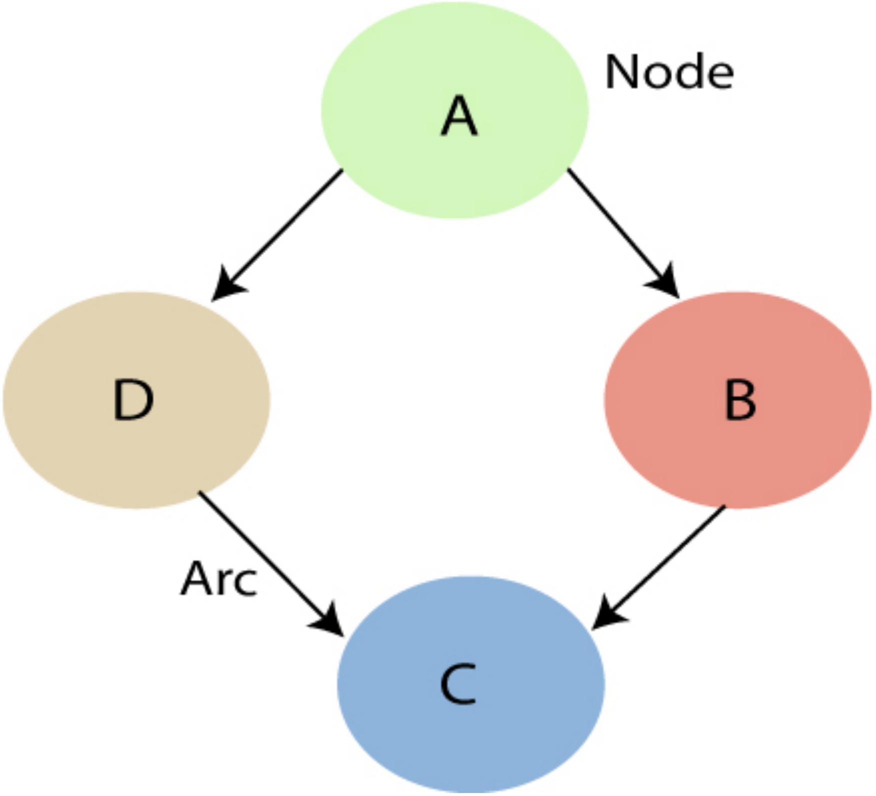
- From here, we’ll dive deeper into mathematical implications of this concept. Don’t worry. Once you understand them, getting to its *mathematics* is pretty easy.
- To define our model correctly , we need two mathematical models before hand. One to represent the ***likelihood function*** P(D|θ)   and the other for representing the distribution of ***prior beliefs*** **.** The product of these two gives the ***posterior belief*** P(θ|D) distribution.
- Since prior and posterior are both beliefs about the distribution of fairness of coin, intuition tells us that both should have the same mathematical form.

Bayesian Belief Network

- Bayesian belief network is key computer technology for dealing with probabilistic events and to solve a problem which has uncertainty. We can define a Bayesian network as:
- "A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph."
- It is also called a **Bayes network, belief network, decision network**, or **Bayesian model**.

- Bayesian networks are probabilistic, because these networks are built from a **probability distribution**, and also use probability theory for prediction and anomaly detection.
- Real world applications are probabilistic in nature, and to represent the relationship between multiple events, we need a Bayesian network. It can also be used in various tasks including **prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction**, and **decision making under uncertainty**.
- Bayesian Network can be used for building models from data and experts opinions, and it consists of two parts:
- **Directed Acyclic Graph**
- **Table of conditional probabilities.**

- The generalized form of Bayesian network that represents and solve decision problems under uncertain knowledge is known as an **Influence diagram**.
- **A Bayesian network graph is made up of nodes and Arcs (directed links), where:**



- Each **node** corresponds to the random variables, and a variable can be **continuous** or **discrete**.
- **Arc or directed arrows** represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph.
  These links represent that one node directly influence the other node, and if there is no directed link that means that nodes are independent with each other
  - **In the above diagram, A, B, C, and D are random variables represented by the nodes of the network graph.**
  - **If we are considering node B, which is connected with node A by a directed arrow, then node A is called the parent of Node B.**
  - **Node C is independent of node A.**

❖ Note: The Bayesian network graph does not contain any cyclic graph. Hence, it is known as a **directed acyclic graph or DAG**.

The Bayesian network has mainly two components:

- **Causal Component**
- **Actual numbers**

Each node in the Bayesian network has condition probability distribution $\mathbf{P(X_i \,|Parent(X_i)\,)}$, which determines the effect of the parent on that node.

Bayesian network is based on Joint probability distribution and conditional probability. So let's first understand the joint probability distribution:

Joint probability distribution:

If we have variables x1, x2, x3,....., xn, then the probabilities of a different combination of x1, x2, x3.. xn, are known as Joint probability distribution.

P[x1, x2, x3,....., xn], it can be written as the following way in terms of the joint probability distribution.

= P[x1| x2, x3,....., xn]P[x2, x3,....., xn]

= P[x1| x2, x3,....., xn]P[x2|x3,....., xn]....P[xn-1|xn]P[xn].

In general for each variable Xi, we can write the equation as:

$$P(X_i|X_{i-1},........., X_1) = P(X_i |Parents(X_i))$$

- Example :

Support Vector Machine (SVM)

*A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.*