$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The LHS is the posterior probability where B is the parameter of interest, A is the evidence/data, and $B|A$ is the targeted estimation. On the RHS, $P(A|B)$ is the likelihood or probability of data/effect and $P(B)$ is a prior probability, a prior model or theory.

Finding $P(B|A)$ using A(data) and applying it to $P(B)$ is called Bayesian conditionalism.

Definition 0.1. Law of Total Probability: Let $B_1, \ldots, B_k$ be mutually exclusive events and collectively exhaustive. Then

$$P(A) = \sum_{i=1}^{k} P(A, B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Theorem 0.1. Baye's Theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$

Definition 0.2. Bayesian Conditionalism is taking $P(B)$, adding A, or data, to it, to find $P(B|A)$

Another way to think about probability of A is: Odds$(A) := \frac{P(A)}{P(A^C)} = \frac{P(A)}{1-P(A)}$.

Example 0.1. Let's say an event has an odds of 4, or "4 to 1" odds. Then the event has a probability of occurring of 0.8 since for each 4 +1, or 5, chances, the odds of it occurring is 4.

Note: To get odds against,

$$\text{Odd}(A)^{-1} = \frac{P(A^C)}{P(A)} = \frac{1-P(A)}{P(A)}$$

Example 0.2. Let A represent the event of a person being a smoker and B be the event that a person has lung cancer.

$$P(A) = 0.2, P(B) = 0.0.06, P(A, B) = 0.036$$

Then $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{0.36}{0.06} = 0.06$. That's easy.

$$P(A|B^C) = \frac{P(A, B^C)}{P(B^C)} = \frac{P(A) - P(A, B)}{1 - P(B)} = \frac{0.2 - 0.036}{1 - 0.06} = 0.174$$

What's the ratio of $\frac{P(B|A)}{P(B^C)|A}$? Well we know, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ and $P(B^C|A) = \frac{P(A|B^C)P(B^C)}{P(A)}$. Thus,

$$\underbrace{\frac{P(B|A)}{P(B^C|A)}}_{\text{posterior odds}} = \overbrace{\frac{P(A|B)}{P(A|B^C)}}^{\text{likelihood ratio}} \Big(\overbrace{\frac{P(B)}{P(B^C)}}^{\text{prior odds}}\Big)$$

Plugging in the numbers, that gives us

$$\frac{P(B|A)}{P(B^C|A)} = \frac{0.6}{0.174}\Big(\frac{0.06}{0.94}\Big) = 0.22$$

This tells us that the odds of getting lung cancer given that a person smokes is 0.22.

Let X, Y be two random variables. We can represent the joint probability mass function as follows:

$P(X=x, Y=y)$				Supp(Y)		
		1	2	3	4	5
	1					
Supp(X)	2					
	3					
	4					
	5					

Then

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

This is the shorthand form of

$$P(X=x|Y=y) = \frac{P(Y=y|X=x)P(X=x)}{P(Y=y)}$$

For this specific joint PMF,

$$P(Y=y) = P(Y=1|X=1) + \cdots + P(Y=1|X=5)$$

In general,

$$P(Y=y) = \sum_{x \in \text{ Supp}(X)} P(Y=y|X=x) = \sum_{x \in \text{ Supp}(X)} P(Y=y|X=x)P(X=x)$$

This is called marginalization, where we are margining out x.

For a probability density function,

$$f_Y(y) = \int_{x \in \text{ Supp}(X)} f(x,y)\,dx = \int_{x \in \text{ Supp(X)}} f_{y|x} f(x)\,dx$$

Consider $P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$ where x is the data and θ is the parameter of a model where $\mathcal{L}(\theta; X) = P(X;\theta)$. The LHS is the probability of cause given effect whereas $P(X|\theta)$ is the probability of effect given cause. We say $P(\theta) = \text{Deg}(\theta_0) = \{0,1\}$. We don't know what θ is exactly so $P(\theta)$ is degenerate. Also, for $P(X)$, we can't find the probability of the data values X without knowing θ. If we did, then $P(X) = \sum_{\theta \in \Theta} P(X|\theta_0)P(\theta_0)$. But $P(\theta_0)$ can only be zero or one (in the case $\theta_0 = \theta$). Thus $P(X) = P(X|\theta)$. This problem began when we assumed $P(\theta)$ is 0 or 1. There was only one true value of θ, call it θ_0.

In the frequentist approach, $P(\theta)$ is degenerate, In the Bayesian approach, we allow $P(\theta)$ to repress our prior knowledge, or prior information.

In the Bayesian approach,

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\sum_{\theta_i \in \Theta} P(X|\theta_i)P(\theta_i)} = \frac{P(X|\theta)P(\theta)}{\int_{\theta_i \in \Theta} P(X|\theta_i)P(\theta_i)\, d\theta_i}$$

Example 0.3. Let's assume $\mathcal{F}$ is a Bernoulli model where $X = \langle 0, 1, 1 \rangle$ and assume IID. If we estimate θ to be 0.75,

$$P(X|\theta = 0.75) = 0.25 \times 0.75^2 = 0.141$$

If we estimate θ to be 0.25,

$$P(X|\theta = 0.25) = 0.75 \times 0.25^2 = 0.047$$

Here we assumed $\Theta = \{0.25, 0.75\}$. But what's $P(\theta = 0.75|X)$?

$$P(\theta = 0.75|X) = \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X)}$$

We know that $P(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.25 \\ 0.5 & \text{if } \theta = 0.75 \end{cases}$. This is the principle of inference; we take all models to be equally likely. Then

$$\begin{aligned} P(\theta = 0.75|X) &= \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X)} \\ &= \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X|\theta = 0.75) + P(X|\theta = 0.25)} \\ &= \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X|\theta = 0.75)P(\theta = 0.75) + P(X|\theta = 0.25)P(\theta = 0.25)} \\ &= \frac{0.141 \times 0.5}{0.141 \times 0.5 + 0.047 \times 0.5} \\ &= 0.75 \end{aligned}$$

If we know this, what is $P(\theta = 0.25|X)$?

$$P(\theta = 0.25|X) = 1 - P(\theta = 0.75|X) = 1 - 0.75 = 0.25$$

Let X and θ be two random variables having a joint distribution. The "dim space" (of all possible realizations) if X can be 0 or 1 and there's three trials is:

$$x \in X = \{\langle 0,0,0 \rangle, \langle 0,0,1 \rangle, \langle 0,1,0 \rangle, \langle 1,0,0 \rangle, \langle 0,1,1 \rangle, \langle 1,0,1 \rangle, \langle 1,1,0 \rangle, \langle 1,1,1 \rangle\}$$

Then

$$\begin{aligned} P(x = \langle 0,0,0\rangle, \theta = 0.25) &= P(x = \langle 0,0,0\rangle | \theta = 0.25) P(\theta = 0.25) \\ &= 0.75^3 \times 0.5 = 0.211 \\ P(x = \langle 1,0,0\rangle, \theta = 0.25) &= 0.25 \times 0.75^2 \times 0.5 = 0.070 \\ P(x = \langle 1,1,0\rangle, \theta = 0.25) &= 0.25^2 \times 0.75 \times 0.5 = 0.023 \\ P(x = \langle 1,1,1\rangle, \theta = 0.25) &= 0.25^3 \times 0.5 = 0.008 \end{aligned}$$

What if we want to do it for the case where $\theta = 0.75$? Then $P(\langle 0,0,0\rangle, \theta = 0.75) = 0.008$. In fact, it'll be all the above probabilities, but reversed.

Is θ independent of X? No. Knowing θ tells you something about X and known x tells you something about θ.

Let's look at the case where $\Theta = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Then $P(\theta) = \begin{cases} 0.2 & \text{if } \theta = 0.1 \\ 0.2 & \text{if } \theta = 0.25 \\ 0.2 & \text{if } \theta = 0.5 \\ 0.2 & \text{if } \theta = 0.75 \\ 0.2 & \text{if } \theta = 0.9 \end{cases}$.

Let $X = \langle 0,1,1\rangle$. Then

$$\begin{aligned} P(X|\theta = 0.1) &= 0.09 \\ P(X|\theta = 0.25) &= 0.047 \\ P(X|\theta = 0.5) &= 0.125 \\ P(X|\theta = 0.75) &= 0.141 \\ P(X|\theta = 0.9) &= 0.061 \end{aligned}$$

What we have found that is that

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \left(\frac{1}{P(X)}\right) P(X|\theta)P(\theta) \propto P(X|\theta)P(\theta) \propto P(X|\theta)$$

We have previously calculated that $\hat{\theta}_{MLE} = 0.66$ for $x = \langle 0,1,1\rangle$ using the point estimate. But according to our best guess here, it is 0.75.