Frequentist vs Bayesian approach

There are two approaches to statistical inference: Bayesian and Frequentist. The method of Maximum Likelihood you've seen so far falls in the Frequentist category.

Let's see what some differences between the two approaches:

Frequentist	**Bayesian**
Probabilities represent long term frequencies	Probabilities represent a degree of belief, or certainty
Parameters are fixed (but unknown) constants, so you can not make probability statements about them	You make probability statements about parameters, even if they are constants (beliefs)
Find the model that better explains the observed data	Update beliefs on the model based on the observed data
Statistical procedures have well-defined long run frequency properties	You make inferences about a parameter θ by producing a probability distribution for θ

The main difference between Frequentists and Bayesians is in the interpretation of probabilities. For Frequentists, probabilities represent long term relative frequencies, which is the frequency of appearance of a certain event in infinite repetitions of the experiment. This implies that probabilities are objective properties of the real world and that the parameters of the distribution are fixed constants; you might not know their value but the value is fixed. Since probabilities represent long term frequencies, Frequentists interpret observed data as samples from an unknown distribution, so it is natural to estimate models in a way that they explain the sampled data as best as possible.

On the other hand, Bayesians interpret probabilities as a degree of belief. This belief applies to models as well. When you are Bayesian, even though you know the parameters take on a fixed value, you are interested on your beliefs on those values. Here is where the concept of prior is introduced. A prior is your baseline belief, what you believe about the parameter before you get new evidence or data. The goal of Bayesians is to update this belief as you gather new data. Your result will be an updated probability distribution for the parameter you are trying to infer. Using this distribution you can later obtain different point estimates.

Let's see how this works with a simple example. Suppose you want to estimate the probability p of a coin landing heads. For that you toss the coin 10 times and record the result of each coin toss, which can be heads or tails. Imagine you get 8 heads out of the 10 coin tosses.

A Frequentist would say that the $p = 8/10$. What about Bayesians?

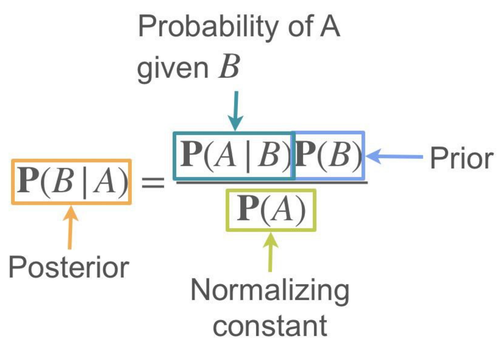
Bayesians will start with a belief. Suppose you know nothing about the coin and your initial belief is that both heads and tails are equally possible. However, after you observe 8 heads in 10 coin flips, you will update your beliefs. You should probably favor values of p around 0.8 rather than smaller values, but you still don't discard completely the possibility that p could have some other value, say 0.3.

How do you actually perform this update? The answer lies in Bayes theorem.

Remember from Week 1 of this course, that Bayes theorem states that given two events A and B

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

But how can you use this to update the beliefs?



Notice that if the event B represents the event that the parameter takes a particular value, and the event A represents your observations, then $\mathbf{P}(B)$ is your prior belief of the parameter, before observing the data. $\mathbf{P}(A|B)$ is the probability of your data given the particular value of the parameter. Finally, $\mathbf{P}(B|A)$ is the updated belief of the parameter, which is called the posterior. Note that $\mathbf{P}(A)$ is simply a normalizing constant, so that the posterior is well defined.

Formalizing Bayesian statistics

In Bayesian statistics, the parameter you want to estimate is considered a random variable, and as such it has a probability distribution. This distribution will represent your beliefs on the parameters.

First, let's introduce some notation. We will use the Greek letter Θ (uppercase theta) to represent any parameters we want to estimate. For example, consider the distributions you learnt in Week 1, Lesson 2:

- If the samples come from a population with a Bernoulli(p) distribution, then $\Theta = p = \mathbf{P}(\text{Success})$
- If the samples come from a population with a Gaussian(μ, σ) distribution, then Θ will be the vector $\Theta = (\mu, \sigma)$.
- If the samples come from a population with a Uniform($0, b$) distribution, then $\Theta = b$.

Now that both the parameters and the data are random variables, to update the posterior you will need the Bayes theorem formula for random variables, rather than events. You learnt about this in Week 2, Lesson 2, 'Conditional Distribution' video.

Bayesian parameter updating (Discrete case)

Following are the main ingredients of Bayes thorem for the discrete case:

- **Parameter** (Θ): the parameter that you want to estimate. Note that we distinguish Θ the random variable representing the parameter, from θ, a particular value that the parameter takes.
- **Sample vector** ($\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$) : your vector of observations
- **Prior distribution** ($f_\Theta(\theta)$): your initial beliefs on the parameter before having any samples. This tells about how you think the probabilities for $\Theta = \theta$ are distributed
- **Conditional distribution of the samples**: For each possible $\Theta = \theta$ you know the joint distribution of the samples. If samples come from a discrete population, you will use the conditional probability mass function (PMF) $p_{\boldsymbol{X}|\Theta=\theta}(x)$. Similarly, if the samples come from a continuous population, you will use the conditional probability density function (PDF) $f_{\boldsymbol{X}|\Theta=\theta}(x)$.
- **Posterior distribution**: your updated beliefs on the parameter Θ after having the data. Following Bayes theorem you have that in general the posterior can be obtained as

$$f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \frac{p_{\boldsymbol{X}|\Theta=\theta}(\boldsymbol{x}) f_\Theta(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \quad \text{if } X \text{ is a discrete random variable}$$

In general, Θ will be a considered a continuous random variable, so we use a PDF for its prior and posterior. In the case where Θ is discrete, just replace a PDF by a PMF, and everything else holds.

An example: parameter p for the Bernoulli distribution.

Let's go back to the coin example stated before. You are interested on the probability of the coin landing heads. Your data consists of 10 coin flips, 8 of which turned out to be heads. Let's interpret who each of the elements are:

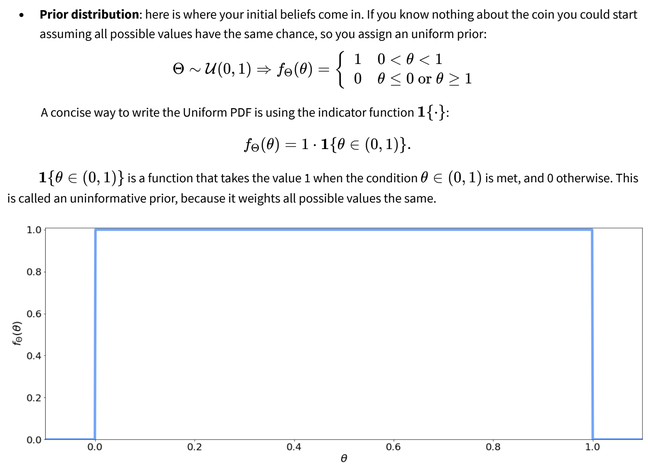
- **Parameters**: $\Theta = \mathbf{P}(H)$
- **Sample vector:** If heads are represented by 1 and tails by 0, then $\boldsymbol{x} = (1, 1, 1, 1, 1, 1, 1, 1, 0, 0)$
- **Prior distribution**: here is where your initial beliefs come in. If you know nothing about the coin you could start assuming all possible values have the same chance, so you assign an uniform prior:

$$\Theta \sim \mathcal{U}(0, 1) \Rightarrow f_\Theta(\theta) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \theta \leq 0 \text{ or } \theta \geq 1 \end{cases}$$

  A concise way to write the Uniform PDF is using the indicator function $\mathbf{1}\{\cdot\}$:

$$f_\Theta(\theta) = 1 \cdot \mathbf{1}\{\theta \in (0, 1)\}.$$

  $\mathbf{1}\{\theta \in (0, 1)\}$ is a function that takes the value 1 when the condition $\theta \in (0, 1)$ is met, and 0 otherwise. This is called an uninformative prior, because it weights all possible values the same.



- **Conditional distribution of the samples:** Each sample comes from a Bernoulli distribution, so $p_{X|\Theta=\theta}(x) = \theta^x(1-\theta)1 - x$, so that the joint conditional distribution can be written as:

$$p_{\boldsymbol{X}|\Theta=\theta}(\boldsymbol{x}) = \theta^{\sum_{i=1}^{10} x_i}(1-\theta)^{10-\sum_{i=1}^{10} x_i}$$

- **Posterior distribution**:

$$\begin{aligned} f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) &= \frac{p_{\boldsymbol{X}|\Theta=\theta}(\boldsymbol{x}) f_\Theta(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{\theta^{\sum_{i=1}^{10} x_i}(1-\theta)^{10-\sum_{i=1}^{10} x_i}\mathbf{1}\{\theta \in (0,1)\}}{p_{\boldsymbol{X}}(\boldsymbol{x})} \end{aligned}$$

Remember that $p_{\boldsymbol{X}}(\boldsymbol{x})$ is simply a normalizing constant.

This is a good time to use the information that $\sum_{i=1}^{10} x_i = 8$, so that

$$f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \frac{\theta^8(1-\theta)^2\mathbf{1}\{\theta \in (0,1)\}}{p_{\boldsymbol{X}}((1,1,1,1,1,1,1,1,0,0))}$$

If you were to do all the calculations, for this constant you would get that the posterior for the probability of heads looks like this:

$$f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \frac{11!}{8!2!}\theta^8(1-\theta)^2\mathbf{1}\{\theta \in (0,1)\}$$