Bayesian inference and MAP

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. Imagine you have four coins, three of which are fair and one biased with a probability of heads of 0.8. You choose randomly one of the four coins, and you want to know if you chose the biased one. You flip the coin 10 times and get 7 heads. Which coin did you choose?

A frequentist would say that you chose the biased one, because it has a higher likelihood:

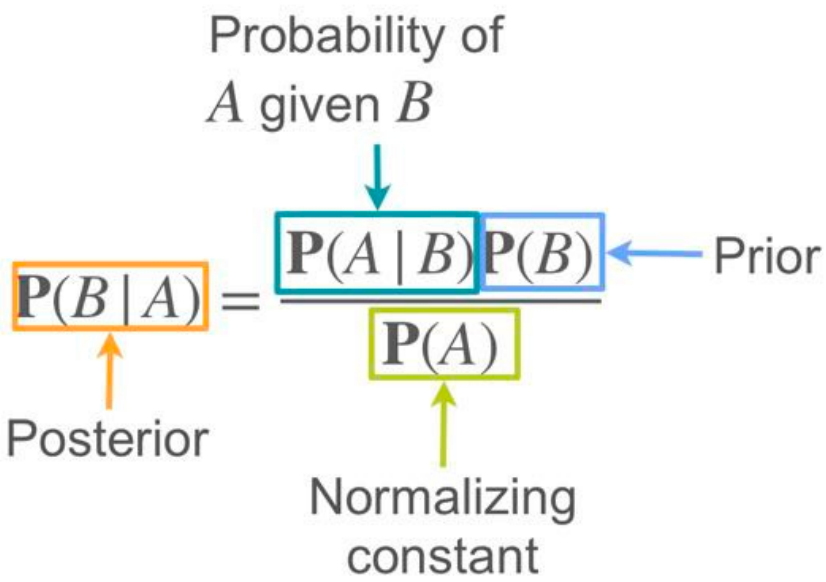
$$L(0.8; 7H3T) = 0.8^7 0.2^3 = 0.0018 \quad > \quad L(0.5; 7H3T) = 0.5^7 0.5^3 = 0.00098$$

What would a Bayesian say? Notice that that the frequentist didn't take into account the fact that the biased coin had 1 in 4 chances of being picked at the beginning. A Bayesian will try to exploit that information, it will be their initial belief: without having any other information (observations) the chances of picking the biased coin is 1/4. The goal of a Bayesian is to update this belief, or probability, based on the observations.

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 (7 heads and 3 tails), then $\mathbf{P}(B)$ is your prior belief of how likely is that you chose the biased coin before observing the data (0.25). $\mathbf{P}(A|B)$ is the probability of your data given the particular value of the parameter (probability of seeing 7 heads followed by 3 tails given that the probability of heads is 0.8). Finally, $\mathbf{P}(B|A)$ is the updated belief, which is called the **posterior**. Note that $\mathbf{P}(A)$ is simply a normalizing constant, so that the posterior is well defined.

$$\mathbf{P}(\text{Biased coin}|7H3T) = \frac{\mathbf{P}(7H3T|\text{Biased coin})\mathbf{P}(\text{Biased coin})}{\mathbf{P}(7H3T)}$$

$$= \frac{0.8^7 0.2^3 0.25}{0.8^7 0.2^3 0.25 + 0.5^7 0.5^3 0.75} = 0.364$$

Look how you went from a 0.25 confidence of choosing the biased coin to 0.364. Note that while your beliefs of having chosen the biased coin have increased, this still isn't the most likely explanation. This is because you are taking into account the fact that originally your chances of choosing the biased coin were much smaller than choosing a fair one.

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.

In the previous example, you are actually trying to estimate the probability of heads, which can be either 0.5, if the coin is fair, or 0.8 is the coin is biased. Before seeing the coin flips, your belief was that $\mathbf{P}(H) = 0.5$ with probability 0.75, and $\mathbf{P}(H) = 0.8$ with probability 0.25. After the 10 coin flips, you updated your beliefs to favor a little bit more the biased coin than initially, so that: $\mathbf{P}(H) = 0.5$ with probability 0.636, and $\mathbf{P}(H) = 0.8$ with probability 0.364.

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, depending on your unknowns, you could have $\Theta = \mu$, $\Theta = \sigma$ or even 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.

There are five components in Bayesian statistics:

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.

Observed sample vector ($\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$) : your vector of observations

Prior distribution: it represents 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. Since the observed values of the sample are fixed, this is turns out to be a function of θ, so the condition distribution of the samples actually represents a **likelihood!**

Posterior distribution: your updated beliefs on the parameter Θ after observing the data. You will update them using the Bayes rules based on the observed data.

There are four scenarios you need to consider, depending on the nature of each of the random variables:

Θ is a discrete random variable, and the data comes from a discrete distribution.

Θ is a continuous random variable, and the data comes from a discrete distribution

Θ is a discrete random variable, and the data comes from a continuous distribution

Θ is a continuous random variable, and the data comes from a continuous distribution

This will affect whether you need a probability mass function or a probability density function to describe the parameter and the data. Let's see how the posterior looks for each of these scenarios.

Discrete parameter, discrete data

This is the case you saw in the example you worked with earlier. In this case, both Θ and the data will be described by a probability mass function (PMF). While considering a discrete distribution for the parameters isn't the most common in practice, it is always good to first understand the simplest case.

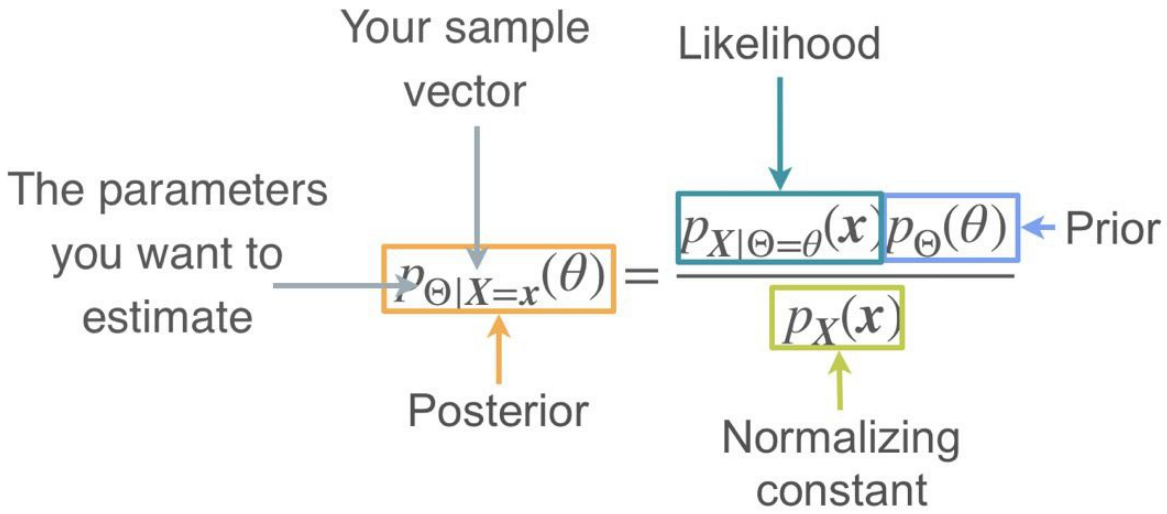
In this case:

Prior distribution will be a PMF ($p_\Theta(\theta)$)

Conditional distribution of the samples: For each possible $\Theta = \theta$ your samples will be described by a conditional PMF ($p_{X|\Theta=\theta}(x)$).

Posterior distribution: since Θ is still discrete, the posterior distribution will also be represented by the posterior PMF. Following Bayes rule you have that in general the posterior can be obtained as:

$$p_{\Theta|X=x}(\theta) = \frac{p_{X|\Theta=\theta}(\boldsymbol{x}) p_\Theta(\theta)}{p_X(\boldsymbol{x})}$$



Let's identify each of the five constitutive elements for the coin example

Parameters: $\Theta = \mathbf{P}(H)$

Sample vector: If heads are represented by 1 and tails by 0, then $\boldsymbol{x} = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$

Prior distribution: here is where your initial beliefs come in.

$$p_\Theta(\theta) = \begin{cases} 0.75 & \text{if } \theta = 0.5 \\ 0.25 & \text{if } \theta = 0.8 \end{cases}$$

Conditional distribution of the samples: Each sample comes from a Bernoulli distribution, so $p_{X|\Theta=\theta}(x) = \theta^x(1-\theta)^{1-x},\ x = \{0, 1\}$, so that the likelihood can be written as:

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

Note that this is the same expression you got for the likelihood when looking for the MLE for a Bernoulli(p) population. This is no coincidence, and is valid in general: whether you interpret the likelihood as coming from a joint distribution in the frequentist approach, or as coming from a conditional joint distribution, as is the case in Bayesian statistics, the final expression for the likelihood is always the same.

Posterior distribution:Continuous parameter, discrete data

$$p_{\Theta|X=x}(\theta) = \frac{p_{X|\Theta=\theta}(\boldsymbol{x})p_\Theta(\theta)}{p_X(\boldsymbol{x})}$$

$$= \begin{cases} \dfrac{0.5^{\sum_{i=1}^{10} x_i}(1-0.5)^{10-\sum_{i=1}^{10} x_i}0.75}{p_X(\boldsymbol{x})} & \text{if } \theta = 0.5 \\ \dfrac{0.8^{\sum_{i=1}^{10} x_i}(1-0.8)^{10-\sum_{i=1}^{10} x_i}0.25}{p_X(\boldsymbol{x})} & \text{if } \theta = 0.8 \end{cases}$$

Replacing for the observations, you get that $\sum_{i=1}^{10} x_i = 7$, and the denominator $p_X(\boldsymbol{x}) = 0.00115$, so that

$$p_{\Theta|X=x}(\theta) = \begin{cases} \dfrac{0.5^7(1-0.5)^3 0.75}{0.00115} & \text{if } \theta = 0.5 \\ \dfrac{0.8^7(1-0.8)^3 0.25}{0.00115} & \text{if } \theta = 0.8 \end{cases}$$

$$= \begin{cases} 0.636 & \text{if } \theta = 0.5 \\ 0.364 & \text{if } \theta = 0.8 \end{cases}$$

While this is written formally as a PMF, it is the same result you got when applying Bayes theorem directly to the events.

Continuous parameter, discrete data

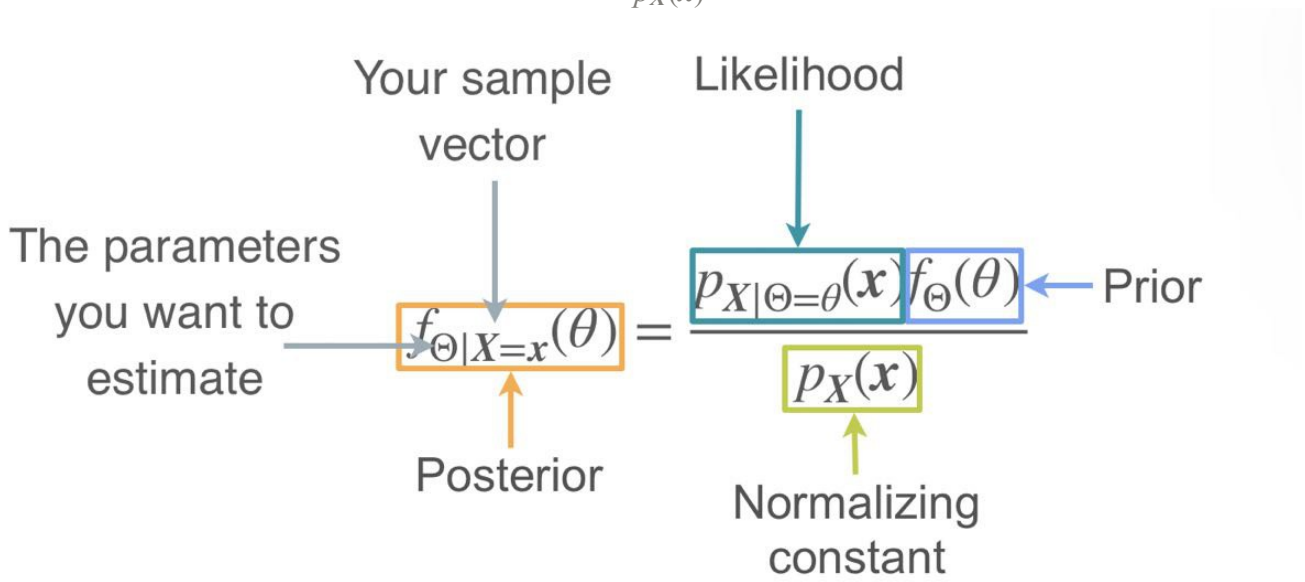
Imagine now that instead of having four particular coins to choose from, you have just one coin but you have no idea what its probability of landing heads is. It could be any value between 0 and 1. In this case, you have infinite possible values for the parameter, so it makes sense to model it as a continuous random variable, so it will be described with a probability density function (PDF). The observations are the same, and data will still be represented by a PMF.

In general, when you have a parameter that is a continuous random variable, and data which is discrete, you get:

Prior distribution: since Θ is continuous, this will be a PDF ($f_\Theta(\theta)$)

Conditional distribution of the samples: For each possible $\Theta = \theta$ the joint distribution of the samples will be described by a conditional PMF $p_{X|\Theta=\theta}(x)$. Remember this is still a **likelihood**, because the values of the observations x are fixed.
Posterior distribution: the parameter Θ after observing the data is still continuous, so the posterior distribution will be described also by a PDF.
Following Bayes rule you have that the posterior for the continuous-discrete case can be obtained as

$$f_{\Theta|X=x}(\theta) = \frac{p_{X|\Theta=\theta}(x) f_\Theta(\theta)}{p_X(x)}$$



Let's see how this works with an example. For reference, remember that now that there are no restrictions on the possible values of the probability of heads, a Frequentist would simply state that $\mathbf{P}(H) = \frac{7}{10}$.
Let's interpret who each of the elements in the Bayesian approach are:

Parameters: $\Theta = \mathbf{P}(H)$

Sample vector: If heads are represented by 1 and tails by 0, then $\boldsymbol{x} = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$

These two elements are the same as before, now what is going to change is the prior:

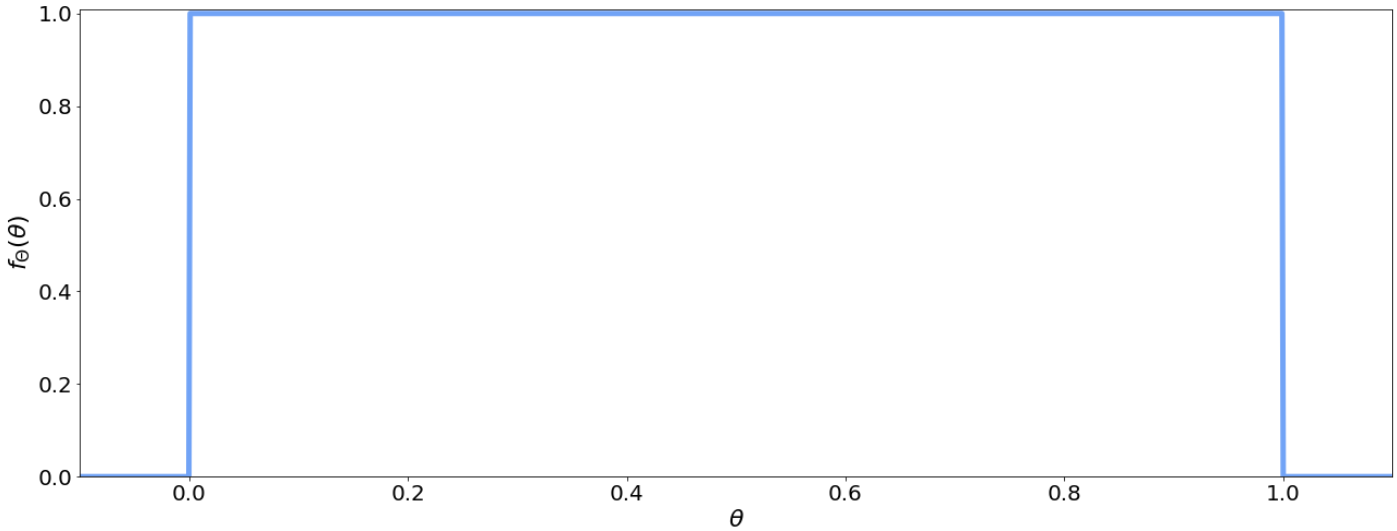
Prior distribution: here is where your initial beliefs come in. If you know nothing about the coin you could start assuming all possible values have the same chance, so you assign an uniform prior:

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

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:

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

Posterior distribution:

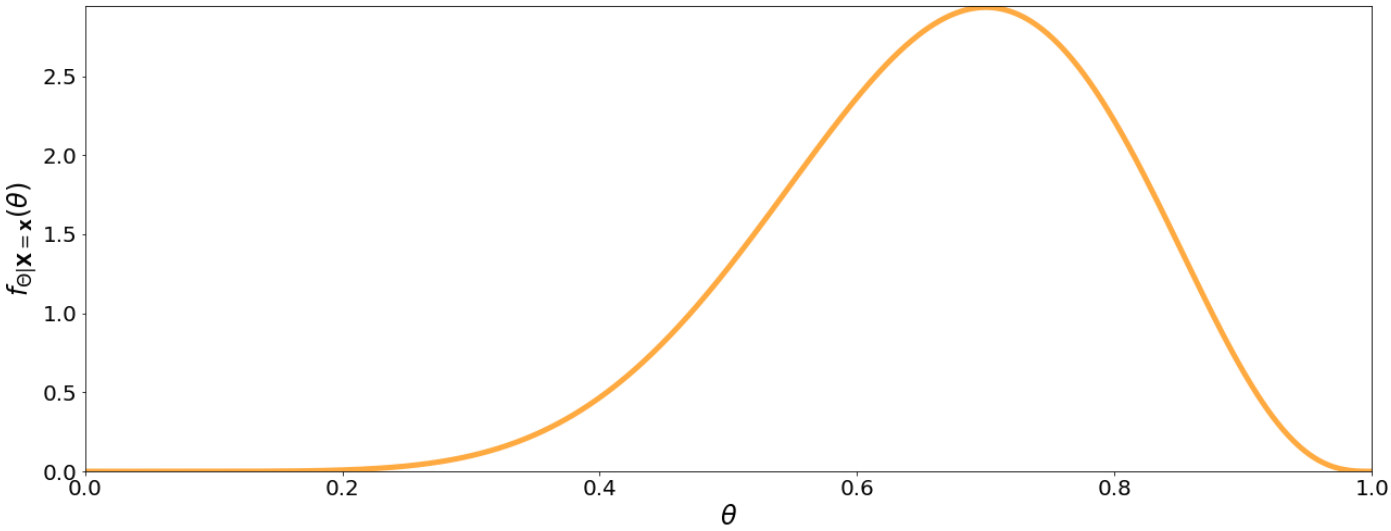
Remember that $p_X(\boldsymbol{x})$ is simply a normalizing constant.

This is a good time to use the information that $\sum_{i=1}^{10} x_i = 7$, so that

$$f_{\Theta|X=\boldsymbol{x}}(\theta) = \frac{\theta^7(1-\theta)^3\mathbf{1}\{\theta \in (0,1)\}}{p_X((1,1,1,1,1,1,1,0,0,0))}$$

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:

$$f_{\Theta|X=\boldsymbol{x}}(\theta) = \frac{11!}{7!3!}\theta^7(1-\theta)^3\mathbf{1}\{\theta \in (0,1)\}$$



This is actually a probability distribution commonly encountered in Bayesian statistics, called the Beta distribution. In this example, the parameters of this distribution are 8 and 4. It is denoted as

$$\Theta|\boldsymbol{X} = (1,1,1,1,1,1,1,0,0,0) \sim \beta(8,4)$$

You can learn more about the Beta distribution in this video by Luis Serrano.

Notice that this posterior distribution has a peak at 0.7, which is exactly the estimate you would get from MLE (Frequentist approach), and it clearly favors values of θ that are close to 0.7. For small values of θ the posterior density is not zero, but it is very close. This implies that the belief that θ can take small values is tiny, which makes sense based on the data. You started with a belief that every value of θ was equally likely, and based on the 7 heads you got you updated these beliefs to favor values of θ close to 0.7.

Discrete parameter, continuous data

Now, let's move on to the third case: discrete parameter with continuous data. In this case, you will have:.

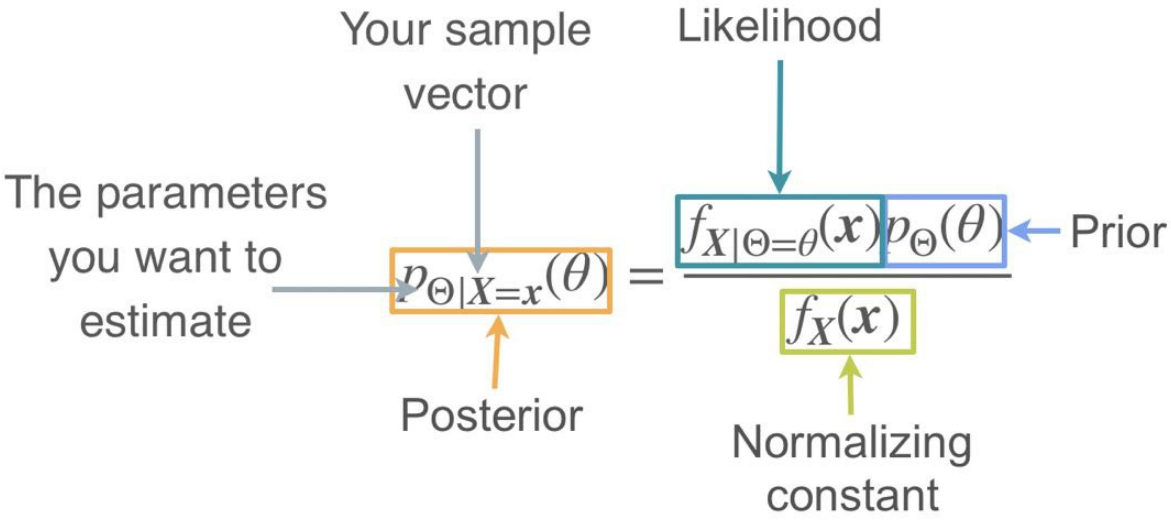
Prior distribution: since Θ is discrete, you will describe your prior beliefs with a PMF ($p_\Theta(\theta)$)

Conditional distribution of the samples: For each possible $\Theta = \theta$ joint distribution of the samples will be represented with conditional probability density function ($f_{X|\Theta=\theta}(x)$). Once again, since the observed values of the sample are fixed, and this is actually a function of θ, this must be interpreted as a **likelihood** and not a density!

Posterior distribution: this will be again a PMF.

Following Bayes rule for updating the beliefs you have that the posterior for the discrete-continuous scenario can be obtained as:

$$p_{\Theta|X=\boldsymbol{x}}(\theta) = \frac{f_{\boldsymbol{X}|\Theta=\theta}(\boldsymbol{x})p_\Theta(\theta)}{f_X(\boldsymbol{x})}$$



Continuous parameter, continuous data

This brings us to the last case. Let's use an example from the lectures. You want to know the mean height of 18 year olds in the US. You already know, that height is a continuos magnitude that nicely fits a Gaussian distribution.. The mean of the population can take any positive value, so it makes sense to also model it as a continuous random variable.

In general, when you have a parameter that is a continuous random variable, and data which is also continuous, you get:

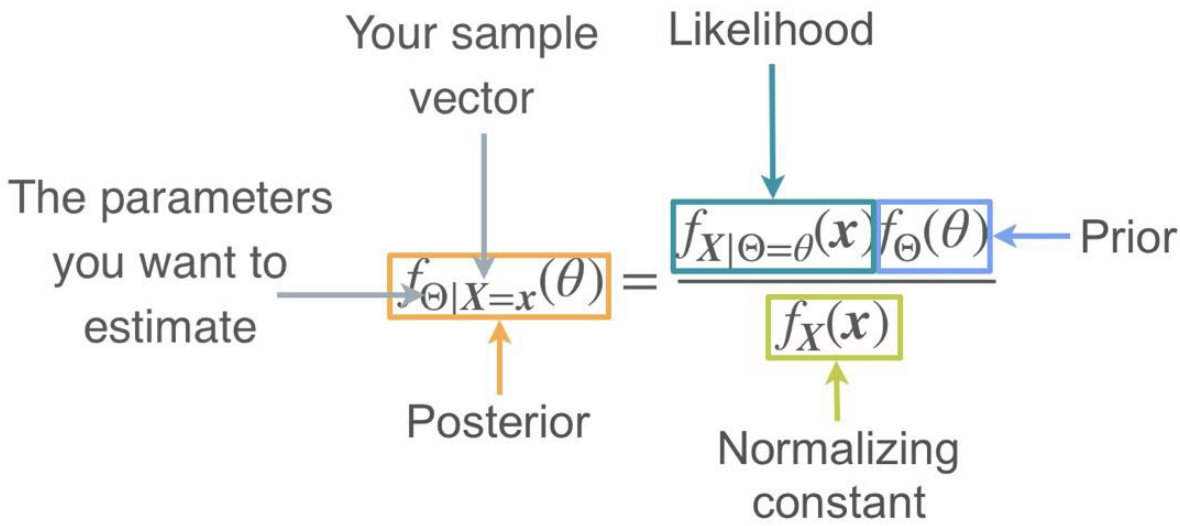
Prior distribution: since Θ is continuous, this will be a PDF $(f_\Theta(\theta))$

Conditional distribution of the samples: For each possible $\Theta = \theta$ the joint distribution of the samples will be described by a conditional PDF $f_{X|\Theta=\theta}(x)$. Remember this is still a **likelihood**.

Posterior distribution: the parameter Θ after observing the data is still continuous, so the posterior distribution will be described also by a PDF.

Following Bayes rule you have that the posterior for the continuous-discrete case can be obtained as

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x)f_\Theta(\theta)}{f_X(x)}$$



For our example, suppose you have 10 measurements
$\boldsymbol{x} = (66.75,\ 70.24,\ 67.19,\ 67.09,\ 63.65,\ 64.64,\ 69.81,\ 69.79,\ 73.52,\ 71.74)$, and that the population standard deviation is known and has value 3.

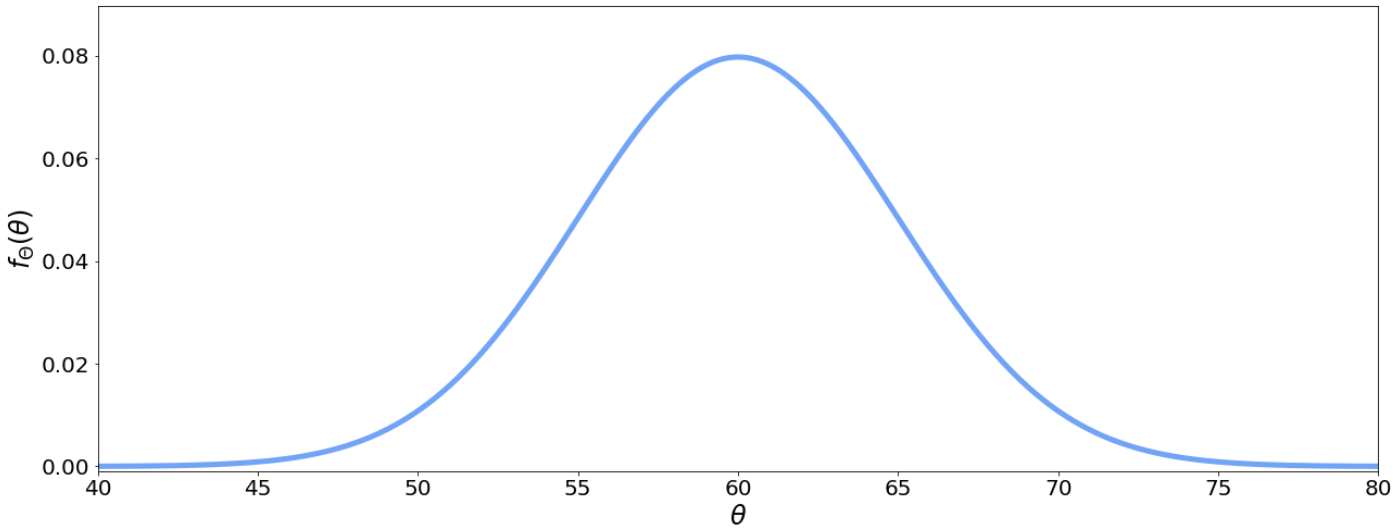
Let's interpret who each of the elements in the Bayesian approach are:

Parameters: $\Theta = \mu = \mathrm{E}(X)$

Sample vector: $\boldsymbol{x} = (66.75,\ 70.24,\ 67.19,\ 67.09,\ 63.65,\ 64.64,\ 69.81,\ 69.79,\ 73.52,\ 71.74)$

Prior distribution: here is where your initial beliefs come in. Suppose you believe that the mean of the population is around 60 inches, with a standard deviation of 5 inches. In fact, you choose a Gaussian distribution with these parameters, so that

$$\Theta \sim \mathrm{N}(60, 5^2) \Rightarrow f_\Theta(\theta) = \frac{1}{\sqrt{2\pi}\,5} e^{-\frac{1}{2}\frac{(\theta-60)^2}{5^2}}$$



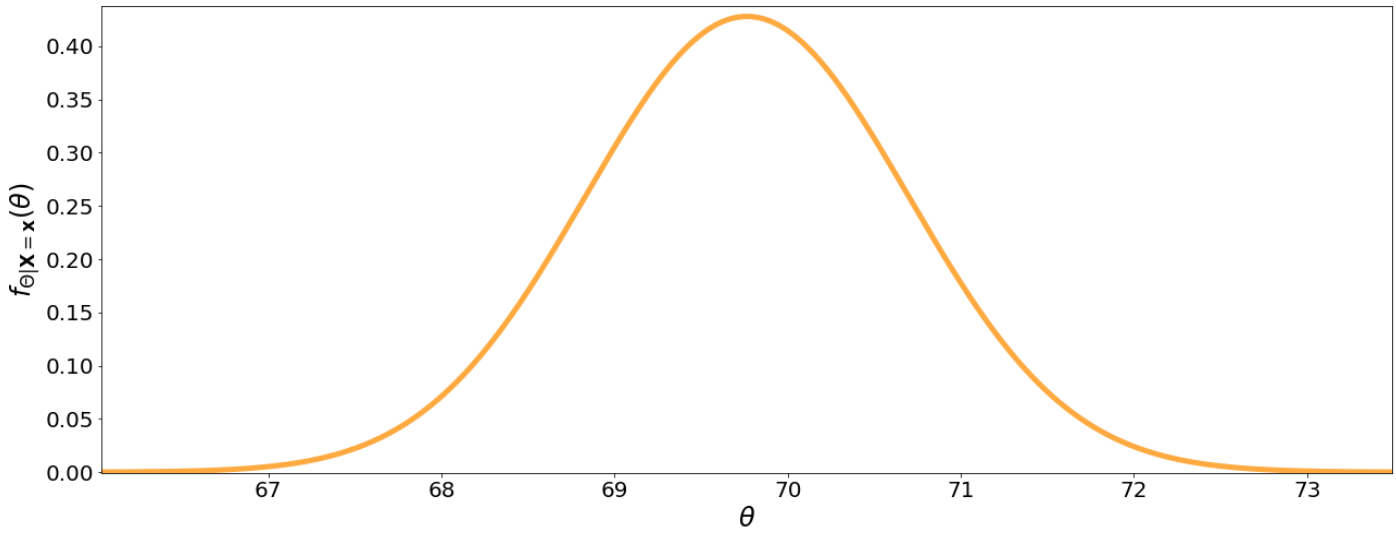
Conditional distribution of the samples: the likelihood will be based on the conditional PDF of the samples: $f_{X|\Theta=\theta}(x) = \frac{1}{\sqrt{2\pi}\,3} e^{-\frac{1}{2}\frac{(x-\theta)^2}{3^2}}$, so that the joint conditional distribution can be written as:

Posterior distribution: This is where math gets a little bit tricky. Let's see how the posterior looks like. For now, remember the $f_X(x)$is simply a normalizing constant.

$$f_{\Theta|X=x}(\theta) = \frac{1}{\text{constant}} \frac{1}{(\sqrt{2\pi}\,3)^{10}} e^{-\frac{1}{2}\frac{\sum_{i=1}^{10}(x_i-\theta)^2}{3^2}} \frac{1}{\sqrt{2\pi}\,5} e^{-\frac{1}{2}\frac{(\theta-60)^2}{5^2}}$$

After some algebraic manipulation and evaluating at the observed data, you get that

$$\Theta|\boldsymbol{X} = \boldsymbol{x} \sim \mathrm{N}\,(69.76, 0.869)$$



Please see the appendix at the end of this reading item for the derivation of the posterior in the above example.

It is interesting to note that the posterior distribution of Θ is still Gaussian, you just updated its parameters based on the observed data.

Maximum a Posteriori (MAP)

Sometimes you still want a point estimate for the parameter Θ . You can use the posterior distribution to define different point estimates. One of the most commonly used point estimates in Bayesian statistics is known as Maximum a Posteriori (MAP). As the name suggests, it the the value θ that maximizes the posterior distribution of the parameter. If Θ is a continuous random variable, this can be expressed as

$$\hat{\theta} = \arg\max_{\theta} f_{\Theta|X=x}.$$

Note that the MAP is nothing more than the **posterior mode.**

Some other frequently used point estimators are

Posterior mean:

$$\hat{\theta} = \mathrm{E}[\Theta|\boldsymbol{X} = \boldsymbol{x}].$$

It can be shown that this estimate has the property of minimizing the L_2, or quadratic, error.

Posterior median: median of the posterior distribution

$$\hat{\theta} = median(\Theta|\boldsymbol{X} = \boldsymbol{x}).$$

This estimate actually minimizes the L_1 error.

The demonstration of the the posterior mean and median minimize L_2 and L_1 respectively is far too advanced for the intended scope of this course. However, if you are interested, you can find the derivations in the following references:
Proof that the posterior mean minimizes the L2 error (page 198, Theorem 12.8)

Proof that the posterior median minimizes the L1 error

Typically, these estimators will give a different value for the point estimate.
Let's take a look at the coin example from earlier (consider the continuous parameter case). Remember that the posterior distribution for Θ was $\Theta|\boldsymbol{X} = (1,1,1,1,1,1,1,0,0,0) \sim \mathrm{B}(8,4)$. You can readily find the mean, median and mode for a Beta distribution on the Wikipedia article:
Posterior mean: $\hat{\theta} = \mathrm{E}[\Theta|\boldsymbol{X} = \boldsymbol{x}] = \frac{8}{8+4} = \frac{2}{3}$
Posterior median: $\hat{\theta} = median(\Theta|\boldsymbol{X} = \boldsymbol{x}) = 0.6762$
MAP: $\hat{\theta} = \arg\max_\theta f_{\Theta|\boldsymbol{X}=\boldsymbol{x}} = \frac{8-1}{8+4-2} = \frac{7}{10}$.
This is the same result you got with the Frequentist approach! Is this a coincidence?

Relationship between MAP and MLE
It turns out the MAP estimation and the MLE have a lot in common. Begin by noting that the conditional distribution of the data has the same expression as the likelihood:

$$p_{\boldsymbol{X}|\Theta=\theta}(\boldsymbol{x}) = \prod_{i=1}^n p_{X_i|\Theta=\theta(x_i)} \quad L(\theta;\boldsymbol{x}) = \prod_{i=1}^n p_{X_i}(x_i)$$

The interpretations are slightly different, because one comes from a conditional distribution and the other does not.. However, the expressions are exactly the same.
If you use an uninformative prior, just like the coin example, then you are assigning equal weight to every possible value of the parameter. You are essentially multiplying the conditional distribution of your data by a constant, so that the posterior is proportional to this function:

$$f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \frac{p_{\boldsymbol{X}|\Theta=theta}(\boldsymbol{x})\text{constant}_1}{\text{constant}_2} \propto p_{\boldsymbol{X}|\Theta=theta}(\boldsymbol{x})$$

Since the posterior is proportional to the conditional distribution of the data, the value θ that maximizes the posterior distribution has to be the same as the θ that maximized the conditional distribution of data. But this conditional distribution coincided with the likelihood. **This means that the MAP estimation with an uninformative prior yields the same result as the MLE**.
Continuing the example
Let's go back to the coin example from earlier. After seeing 7 heads in 10 coin flips, you updated on the parameter $\Theta = \mathbf{P}(H)$, from a Uniform(0,1) distribution to the Beta(8,4) distribution. Imagine you want to improve your beliefs even more, so you decide to flip the coin 5 more times, observing 3 heads and 2 tails.
In this new scenario, you can consider the Beta(8,4) distribution as your prior for Θ, and update the posterior using your new observations. Let's see what this looks like.
Parameters: $\Theta = \mathbf{P}(H)$, this does not change
Sample vector: This time you have 3 heads in 5 coin flips. If heads are represented by 1 and tails by 0, then $\boldsymbol{x} = (1,1,1,0,0)$
Prior distribution: this time you will consider the Beta(8,4) distribution you found earlier as the posterior:

$$f_\Theta(\theta) = \frac{11!}{7!3!}\theta^7(1-\theta)^3\mathbf{1}\{\theta \in (0,1)\}$$

Conditional distribution of the samples: Each sample still comes from a Bernoulli distribution: $p_{X|\Theta=\theta}(x) = \theta^x(1-\theta)^{1-x}$. With a sample size of 5, the joint conditional distribution can be written as:
Posterior distribution: now, all you have to do is use the posterior distribution formula to get the update on your beliefs:

$$f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \frac{p_{\boldsymbol{X}|\Theta=\theta}(\boldsymbol{x})f_\Theta(\theta)}{p_{\boldsymbol{X}}(\boldsymbol{x})}$$

$$= \frac{\theta^{\sum_{i=1}^5 x_i}(1-\theta)^{5-\sum_{i=1}^5 x_i}\frac{11!}{7!3!}\theta^7(1-\theta)^3\mathbf{1}\{\theta\in(0,1)\}}{p_{\boldsymbol{X}}(\boldsymbol{x})}$$

Remember that $p_{\boldsymbol{X}}(\boldsymbol{x})$ is simply a normalizing constant, and for that matter, $\frac{11!}{7!3!}$ is a constant too.
This is a good time to use the information that $\sum_{i=1}^5 x_i = 3$:

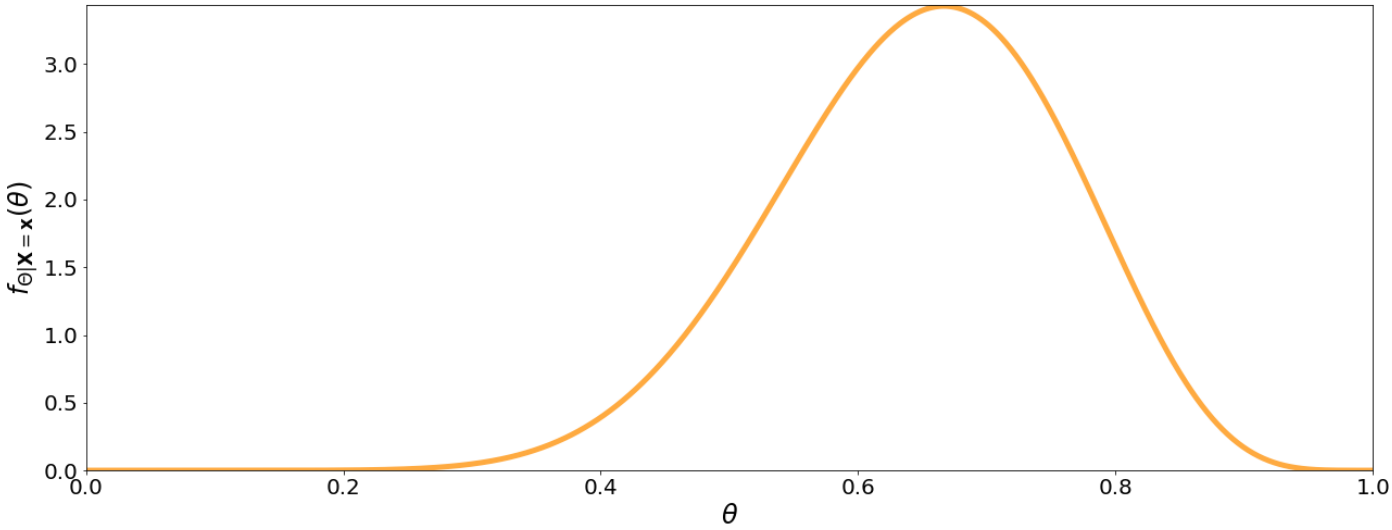
$$f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \frac{\theta^3(1-\theta)^2\frac{11!}{7!3!}\theta^7(1-\theta)^3\mathbf{1}\{\theta\in(0,1)\}}{p_{\boldsymbol{X}}((1,1,1,0,0))}$$

Grouping exponents together, you get:

$$f_{\Theta|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \frac{\theta^{7+3}(1-\theta)^{3+2}\mathbf{1}\{\theta\in(0,1)\}}{\text{constant}}$$

Once again, the constant is there so that the area under the density is exactly 1. If you were to do the math, you would see that the constant is $\frac{16!}{10!5!}$, so that the new posterior results

$$f_{\Theta|X=x}(\theta) = \frac{16!}{10!5!}\theta^{10}(1-\theta)^5 \mathbf{1}\{\theta \in (0,1)\}$$



This is in fact once again a Beta distributions, this time with parameters 11 and 6, and is noted

$$\Theta|X = x \sim \beta(11,6)$$

Note the peak of the distribution has shifted towards smaller values, and it is something between $\frac{3}{5} = 0.6$, the observed sample mean, and $\frac{7}{10} = 0.7$, the peak of the prior. The end result is a mixture of the the previous beliefs and the new observations.
This last example shows how an informative prior models the posterior distribution of the parameter.

Appendix: Derivation of the continuous parameter, continuous data case

Here we show the details of the derivation of the result we obtained for the example of the continuous parameter, continuous data case you saw earlier. The goal is to find the posterior distribution for the mean of a Gaussian population with known variance. We had the following from the example:
Parameters: $\Theta = \mu = \mathrm{E}(X)$

Sample vector: $\boldsymbol{x} = (66.75,\ 70.24,\ 67.19,\ 67.09,\ 63.65,\ 64.64,\ 69.81,\ 69.79,\ 73.52,\ 71.74)$

Prior distribution:

$$\Theta \sim \mathrm{N}(60, 5^2) \Rightarrow f_\Theta(\theta) = \frac{1}{\sqrt{2\pi}\,5} e^{-\frac{1}{2}\frac{(\theta-60)^2}{5^2}}$$

Conditional distribution of the samples:

$$f_{\boldsymbol{X}|\Theta=\theta}(\boldsymbol{x}) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\,3} e^{-\frac{1}{2}\frac{(x_i-\theta)^2}{3^2}} = \frac{1}{(\sqrt{2\pi}\,3)^{10}} e^{-\frac{1}{2}\frac{\sum_{i=1}^{10}(x_i-\theta)^2}{3^2}}$$

Posterior distribution:

$$f_{\Theta|X=x}(\theta) = \frac{1}{\text{constant}} \frac{1}{(\sqrt{2\pi}\,3)^{10}} e^{-\frac{1}{2}\frac{\sum_{i=1}^{10}(x_i-\theta)^2}{3^2}} \frac{1}{\sqrt{2\pi}\,5} e^{-\frac{1}{2}\frac{(\theta-60)^2}{5^2}}$$

There are a lot of constants there, let's absorb them together. Also, note a little trick that $(x_i-\theta)^2 = (\theta - x_i)^2$, so you can rewrite

$$f_{\Theta|X=x}(\theta) = \frac{1}{\text{constant}_2} e^{-\frac{1}{2}\frac{\sum_{i=1}^{10}(\theta-x_i)^2}{3^2}} e^{-\frac{1}{2}\frac{(\theta-60)^2}{5^2}} = \frac{1}{\text{constant}_2} e^{-\frac{1}{2}\left(\frac{\sum_{i=1}^{10}(\theta-x_i)^2}{3^2} + \frac{(\theta-60)^2}{5^2}\right)}$$

The exponent is clearly quadratic function of θ. Could it be rewritten in as something with the structure $k(\theta-\mu)^2 + \text{constant} = k(\theta^2 - 2\mu\theta + \mu^2) + \text{constant}$? Start with expanding the quadratic terms and gathering common factors.
Remember that the x_i are fixed, so they are constants. Let's group every term that is inside the sum into a single additive constant

$$\frac{\sum_{i=1}^{10}(\theta-x_i)^2}{3^2}+\frac{(\theta-60)^2}{5^2} = \frac{10}{3^2}\theta^2-\frac{2}{3^2}\theta\sum_{i=1}^{10}x_i+\frac{1}{5^2}\theta^2-\frac{2}{5^2}\cdot 60\theta+\text{constant}$$

$$= \theta^2\left(\frac{10}{3^2}+\frac{1}{5^2}\right)-2\theta\left(\frac{1}{3^2}\sum_{i=1}^{10}x_i+\frac{1}{5^2}60\right)+\text{constant}$$

We're really close! First, since you want to have something like $\text{constant}+k(\theta^2-2\mu\theta+\mu^2)$, you need to factor out the constant affecting θ^2.

$$\frac{\sum_{i=1}^{10}(\theta-x_i)^2}{3^2}+\frac{(\theta-60)^2}{5^2} = \left(\frac{10}{3^2}+\frac{1}{5^2}\right)\left(\theta^2-2\theta\frac{\left(\frac{1}{3^2}\sum_{i=1}^{10}x_i+\frac{1}{5^2}60\right)}{\left(\frac{10}{3^2}+\frac{1}{5^2}\right)}\right)+\text{constant}$$

Now, note that $\left(\frac{10}{3^2}+\frac{1}{5^2}\right)=\frac{5^2 10+3^2}{3^2 5^2}$ so that $\frac{\left(\frac{1}{3^2}\sum_{i=1}^{10}x_i\right)}{\left(\frac{10}{3^2}+\frac{1}{5^2}\right)}=\frac{5^2}{5^2 10+3^2}\sum_{i=1}^{10}x_i$ and $\frac{\left(\frac{1}{5^2}60\right)}{\left(\frac{10}{3^2}+\frac{1}{5^2}\right)}=\frac{3^2}{5^2 10+3^2}60$, and the equation above becomes

$$\frac{\sum_{i=1}^{10}(\theta-x_i)^2}{3^2}+\frac{(\theta-60)^2}{5^2} = \left(\frac{10}{3^2}+\frac{1}{5^2}\right)\left(\theta^2-2\theta\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3^2}\right)+\text{constant}$$

Now you're truly almost there!! To be able to get the desired form you just need to add the constant $\left(\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3^2}\right)^2$. That's no problem, you can just add it and subtract it from the equation. Note the the subtracted term will simply be absorbed into the additive constant

$$\frac{\sum_{i=1}^{10}(\theta-x_i)^2}{3^2}+\frac{(\theta-60)^2}{5^2} = \left(\frac{10}{3^2}+\frac{1}{5^2}\right)\left(\theta^2-2\theta\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3^2}+\left(\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3^2}\right)^2\right)+\text{constant}_2$$

$$= \left(\frac{10}{3^2}+\frac{1}{5^2}\right)\left(\theta-\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3}\right)^2+\text{constant}_2$$

Replacing this rearranged expression into the posterior PDF for Θ you will find that it should have a very familiar:

$$f_{\Theta|X=x}(\theta) = \frac{1}{\text{constant}}e^{-\frac{1}{2}\left(\frac{\sum_{i=1}^{10}(\theta-x_i)^2}{3^2}+\frac{(\theta-60)^2}{5^2}\right)}$$

$$= \frac{1}{\text{constant}}e^{-\frac{1}{2}\left(\frac{10}{3^2}+\frac{1}{5^2}\right)\left(\theta-\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3}\right)^2+\text{constant}_2}$$

$$= \frac{1}{\text{constant}_3}e^{-\frac{1}{2}\left(\frac{10}{3^2}+\frac{1}{5^2}\right)\left(\theta-\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3}\right)^2}$$

Note that this looks a lot like a Gaussian density!

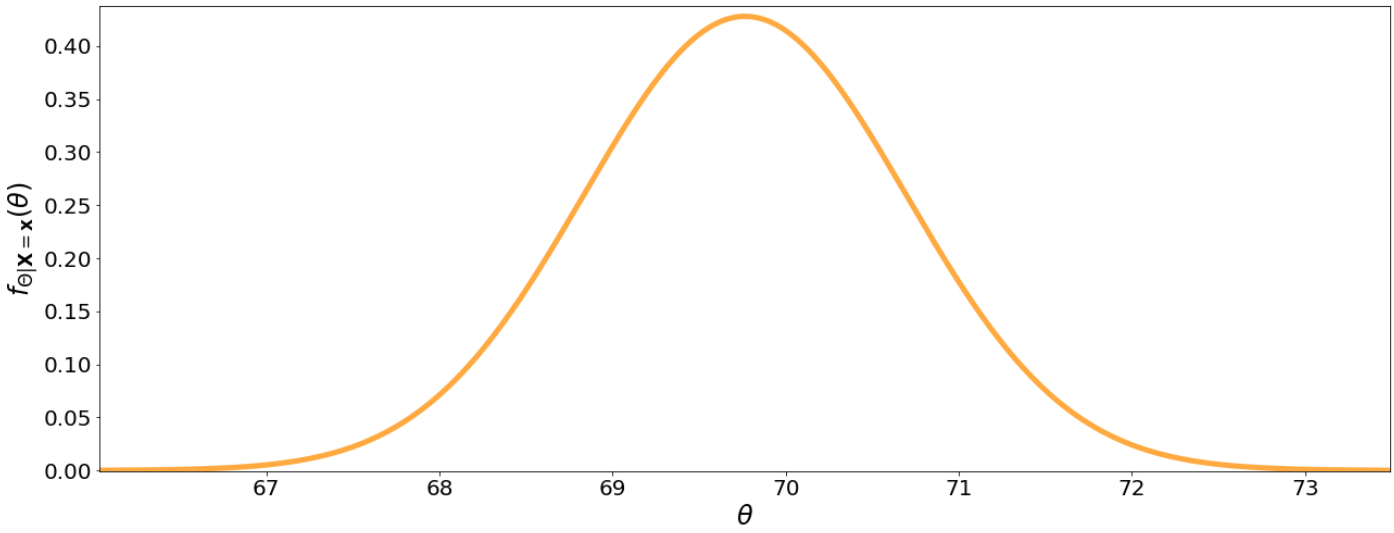
$$f_{\Theta|X=x}(\theta) = \frac{1}{\text{constant}_3}e^{-\frac{1}{2}\underbrace{\left(\frac{10}{3^2}+\frac{1}{5^2}\right)}_{1/(\sigma^2)}\left(\theta-\underbrace{\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3}}_{\mu}\right)^2}$$

Then, you can say that

$$\Theta|X=x \sim \text{N}\left(\frac{5^2\sum_{i=1}^{10}x_i+3^2 60}{5^2 10+3},\ \frac{1}{\frac{10}{3^2}+\frac{1}{5^2}}\right)$$

Replacing with the observed values of x, where $\sum_{i=1}^{10}x_i=684.42$ you have that

$$\Theta|X=x \sim \text{N}(69.76, 0.869)$$


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta)$
67
68
69
70
71
72
73
θ