0.0.1 Q1.1 Find $P(A|B)$

Points: 0.25

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.29}{0.48} = 0.604$$

0.0.2 Q1.2 Find $P(B|A)$

Points: 0.25

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.29}{0.73} = 0.397$$

0.0.3 Q1.3 Determine whether or not A and B are independent.

Points: 0.25

If events A and B are independent than this equation should be satisfied:

$$P(A \cap B) = P(A) \cdot P(B)$$

Let's try it:

$$P(A \cap B) = 0.29$$
$$P(A) \cdot P(B) = 0.73 \cdot 0.48 = 0.3504 \neq P(A \cap B) = 0.29$$

Because

$$P(A \cap B) \neq P(A) \cdot P(B)$$

, it shows that events A and B are not independent of each other.

0.0.4 Q2.1 What is the probability of picking a 100-dollar bill?

Points: 0.2

$$
\begin{aligned}
P(\$100 \cap red) + P(\$100 \cap green) + P(\$100 \cap blue) &= P(\$100) \\
P(\$100 \cap red) &= P(\$100 \mid red) \cdot P(red) = 0.1 \cdot 0.6 = \mathbf{0.06} \\
P(\$100 \cap green) &= P(\$100 \mid green) \cdot P(green) = 0.5 \cdot 0.3 = \mathbf{0.15} \\
P(\$100 \cap blue) &= P(\$100 \mid blue) \cdot P(blue) = 0.9 \cdot 0.1 = \mathbf{0.09} \\
P(\$100 \cap red) + P(\$100 \cap green) + P(\$100 \cap blue) &= 0.06 + 0.15 + 0.09 = \mathbf{0.3}
\end{aligned}
$$

$$\mathbf{P(\$100) = 0.3}$$

0.0.5 Q2.2 What is the probability of picking a 1-dollar bill?

Points: 0.2

$$
\begin{aligned}
P(\$1 \cap red) + P(\$1 \cap green) + P(\$1 \cap blue) &= P(\$100) \\
P(\$1 \cap red) &= P(\$1 \mid red) \cdot P(red) = 0.9 \cdot 0.6 = \mathbf{0.54} \\
P(\$1 \cap green) &= P(\$1 \mid green) \cdot P(green) = 0.5 \cdot 0.3 = \mathbf{0.15} \\
P(\$1 \cap blue) &= P(\$1 \mid blue) \cdot P(blue) = 0.1 \cdot 0.1 = \mathbf{0.01} \\
P(\$1 \cap red) + P(\$1 \cap green) + P(\$1 \cap blue) &= 0.54 + 0.15 + 0.01 = \mathbf{0.7}
\end{aligned}
$$

$$\mathbf{P(\$1) = 0.7}$$

0.0.6 Q2.3 Given that the picked bill is a 100-dollar bill, what is the probability that it came from the Green box?

Points: 0.2

$$P(green \mid \$100) = \frac{P(green \cap \$100)}{P(\$100)} = 0.15$$

$$\mathbf{P(green \mid \$100) = 0.15}$$

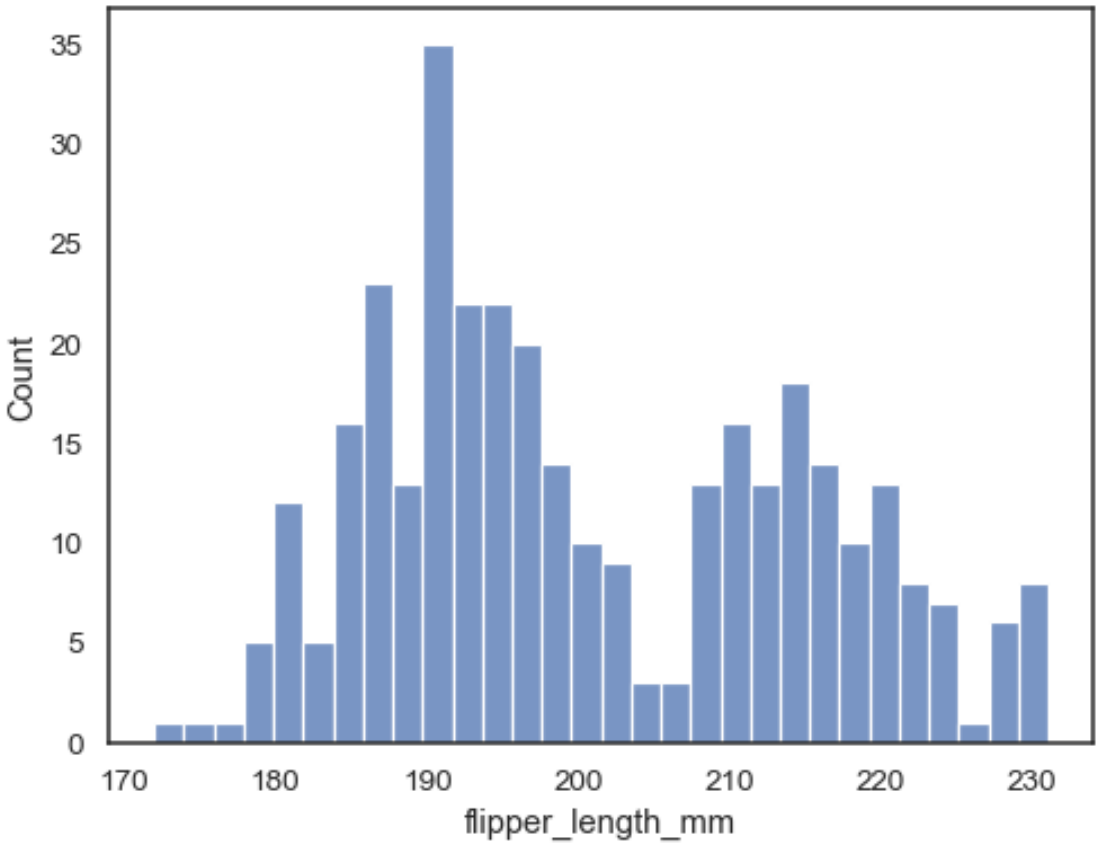
0.0.7 Q2.4 Let's draw a random bill out of the box. What is the expected value of the dollar worth of the bill? What does that mean?

Points: 0.2

$$E[X] = \sum_{i}^{n} x_i \cdot P(X = x_i)$$
$$E[X] = 1 \cdot P(\$1) + 100 \cdot P(\$100) = 1 \cdot 0.7 + 100 \cdot 0.3 = 0.7 + 30 = 30.7$$
$$\mathbf{E[X] = \$30.7}$$

The expected value of the dollar worth of the bill is $30.70.What this means is that when a bill is randomly pulled out of th

0.0.8 Q3.1



Here is some data. Do you think this would be well-fit by a single normal distribution? If not, is there another distribution that you would suggest to fit this data? Do you think this data would be well fit by multiple normal distributions? If so, how many would you suggest, and why that? Don't freak out here… while there are better and worse answers to this question there is not just a single right answer.

Points: 0.2

My Answer I don't think this data would be well-fit by a single normal distribution. I do think this is a mixture of normal distributions. I would think that this data would be fit by adding 2 normal distributions together. I think this because observing the data, we can compare the data to what a normal distribution looks like and can infer that this data could possibly be represented with 2 normal distributions that have a mean of 190 and 215 and some reasonable standard deviation.

0.0.9 Q3.2

Picking a distribution by eye like we did above is NOT good. It's best if you have a theoretical reason, based on the math of the thing you are modeling. With that in mind, what kind of distribution (or distributions!) would you expect to use to model _____ and why: 1. Whether a coin is fair or not? 2. How frequently we expect customers to enter a store? 3. Height of male college basketball players 4. Height among all college freshmen

Points: 0.4

1) Binomial A Binomial distribution focuses on representing the probability of 2 outcomes inside a sample space. This can be catered to how fair a coin is because a coin flip has only 2 outcomes.

2) Poisson A Poisson distribution focuses on representing the probability of how frequent an event is to be inside a certain time frame. This can be catered to how frequently customers enter a store given a certain time frame such as a week.

3) Normal A Normal distribution focuses on representing a set of data that varies around a mean or expected value seen inside the data. This can be catered to the height of male college Basketball players because there is a mean height that can be said to be the expected or average height of male college Basketball players and any other height deviates from that mean.

4) Normal The same explanation can be used from part 3 for part 4. There's an average height among all college freshman and a deviation from that average. This means that the data could be represented too as a normal distribution.

0.0.10 Q4.1

$$\sum_{x=0}^{1} p(x|\mu) = 1$$

Points: 0.25

$$\sum_{x=0}^{1} p(x|\mu) = 1$$

$$\begin{aligned}
\sum_{x=0}^{1} p(x|\mu) &= p(x=0|\mu) + p(x=1|\mu) = 1 \\
&= \mu^{x=0} \cdot (1-\mu)^{1-x=0} + \mu^{x=1} \cdot (1-\mu)^{1-x=1} = 1 \\
&= 1 \cdot (1-\mu)^1 + \mu \cdot (1-\mu)^0 = 1 \\
&= 1 \cdot (1-\mu) + \mu \cdot 1 = 1 \\
&= (1-\mu) + \mu = 1 \\
&= 1 - \mu + \mu = 1 \\
&= 1 = 1
\end{aligned}$$

0.0.11 Q4.2

$$\mathbb{E}[x] = \mu$$

Points: 0.25

$$\mathbb{E}[x] = \mu$$

$$p(x = 1) = \mu$$

$$\mathbb{E}[x] = \sum_x p(x) \cdot x$$

$$\mathbb{E}[x] = p(x = 0) \cdot x = 0 + p(x = 1) \cdot x = 1$$

$$\mathbb{E}[x] = (1 - \mu) \cdot 0 + \mu \cdot 1$$

$$\mathbb{E}[x] = \mu$$

0.0.12 Q4.3

$$Var[x] = \mu(1-\mu)$$

Points: 0.35

$$Var[x] = \mu(1-\mu)$$

$$\begin{aligned}
\mathbb{E}[x] &= \mu \\
Var[x] &= \mathbb{E}[x^2] - \mathbb{E}^2[x] \\
\mathbb{E}^2[x] &= (\mathbb{E}[x])^2 = \mu^2 \\
\mathbb{E}[x^2] &= \sum_x p(x) \cdot x^2 \\
&= p(x=0) \cdot 0^2 + p(x=1) \cdot 1^2 \\
&= 0 + \mu \cdot 1 \\
&= \mu \\
Var[x] &= \mathbb{E}[x^2] - \mathbb{E}^2[x] = \mu - \mu^2 \\
&= \mu \cdot (1-\mu)
\end{aligned}$$

0.0.13 Q5.1

What is the probability of the car being behind Door 1 given that you chose Door 1 and Monty opened Door 2? i.e.

$$P(\text{car} = 1 | \text{choose} = 1, \text{open} = 2)$$

Points: 0.3

$$\begin{aligned}
P(choose = 1, open = 2) &= a \\
P(car = 1 \mid choose = 1, open = 2) &= \frac{P(car = 1, choose = 1, open = 2)}{P(choose = 1, open = 2)} \\
P(car = 1, choose = 1, open = 2) &= P(car = 1) \cdot P(choose = 1 \mid car = 1) \cdot P(open = 2 \mid choose = 1, car = 1) \\
&= \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{18} = 0.056 \\
P(car = 1 \mid choose = 1, open = 2) &= \frac{0.056}{a} = \frac{1}{a} \cdot 0.056
\end{aligned}$$

0.0.14 Q5.2

What is the probability of the car being behind Door 3 given that you chose Door 1 and Monty opened Door 2?

$$P(\text{car} = 3 | \text{choose} = 1, \text{open} = 2)$$

Points: 0.3

$$\begin{aligned}
P(choose = 1, open = 2) &= a \\
P(car = 3 \mid choose = 1, open = 2) &= \frac{P(car = 3, choose = 1, open = 2)}{P(choose = 1, open = 2)} \\
P(car = 3, choose = 1, open = 2) &= P(car = 3) \cdot P(choose = 1 \mid car = 3) \cdot P(open = 2 \mid choose = 1, car = 3) \\
&= \frac{1}{3} \cdot \frac{1}{3} \cdot 1 = \frac{1}{9} = 0.111 \\
P(car = 3 \mid choose = 1, open = 2) &= \frac{0.111}{a} = \frac{1}{a} \cdot 0.111
\end{aligned}$$

0.0.15 Q5.3

Compare your answers from part(1) and part(2), which has a higher probability and should you switch the door?

Points: 0.4

There's a higher probability getting the car after switching the door. If we don't switch our choice, we have a 1 in 18 chances to get the car; however if we switch our choice then we have 1 in 9 chances of picking the car which is double the likelihood than keeping our original choice. We should always switch the door.

0.0.16 Q6.1 What is the correct distribution to describe each dataset AND WHY?

Points: 0.4

For D1, I would say that a binomial distribution would be the correct way to describe that dataset because the data consists of only 2 outcomes and has multiple trials.
For D2, I would say that a normal distribution would be the correct way to describe the dataset because it shows some values cluster around the value 188 and start to deviate from it. This is a lot like a normal distribution with a mean and standard deviation.

0.0.17 Q6.2

Write down the likelihood function $P(D|\theta)$ for each distribution.

Note this will be in term of a product of PDFs for each datapoint in the dataset D because these processes are assumed to be statistically independent.

Points: 0.6

$$x_2 \in \mathcal{D}_2, x_1 \in \mathcal{D}_1$$

$$P(D_2 \mid \theta) = \prod_{k=1}^{n} \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_2-\mu}{\sigma})^2}$$

$$P(D_1 \mid \theta) = \prod_{k=1}^{n} Binomial(n, \theta)$$

$$= \prod_{k=1}^{n} \binom{n}{x_k} \theta^{x_k} (1-\theta)^{n-x_k}$$

0.0.18 Q6.3 Write down the log likelihood function only for the coin datasets, and simplify it enough to get rid of products and exponents.

Note: You don't have to calculate the exact likelihood values, you can use x_k to represent the k^{th} value in the

Points: 0.4

$$
\begin{aligned}
\binom{n}{x} &= \frac{n!}{x!(n-x)!} \\
P(D_1 \mid \theta) &= \prod_{k=1}^{n} \binom{n}{x_k} \theta^{x_k}(1-\theta)^{n-x_k} \\
&= \prod_{k=1}^{n} \frac{n!}{x_k!(n-x_k)!} \theta^{x_k}(1-\theta)^{n-x_k} \\
ln(P(D_1 \mid \theta)) &= \sum_{k=1}^{n} ln(\prod_{k=1}^{n} \frac{n!}{x_k!(n-x_k)!} \theta^{x_k}(1-\theta)^{n-x_k}) \\
ln(P(D_1 \mid \theta)) &= \sum_{k=1}^{n} ln(\frac{n!}{x_k!(n-x_k)!} \theta^{x_k}(1-\theta)^{n-x_k}) \\
ln(P(D_1 \mid \theta)) &= \sum_{k=1}^{n} ln(n!) + x_k ln(\theta) + (n-x_k)(1-\theta) - ln(x_k!) - ln(n-x_k)
\end{aligned}
$$

0.0.19 Q6.4 Only for the coin dataset: how would you use the log likelihood function to analytically solve for the MLE of θ?

Note you don't have to do all the derivation and simplification, but kudos to you if you can. At the very least describe the procedure in words and/or sketch out the beginning of the derivation. The idea is we want you to demonstrate that you understand the concept of MLE, so whatever you think is sufficient to do that.

Points: 0.4

I would take the partial derivative of the log likelihood function with respect to theta and find what value of theta makes that partial derivative equal to 0. If there are more than one parameter in theta, then we would find the gradient of the log likelihood function with respect to each parameter in theta set it to 0. After solving it, we would get the MLE of theta.

0.0.20 Q6.5 For both datasets, what is the equation for the MLE of θ? What are the values for each dataset?

Even if you didn't derive it from first principles, you should still know what the equation is for the MLE. Write down the proper equation, and calculate the value (actual number) of that MLE.

Points: 0.2

This is for the Binomial distribution: $ln(P(D_1 \mid \theta)) = \sum_{k=1}^{n} ln(n!) + x_k ln(\theta) + (n - x_k)(1 - \theta) - ln(x_k!) - ln(n - x_k)$

$$\nabla ln(P(D_1 \mid \theta)) = \sum_{k=1}^{n} \frac{x_k}{\theta} - \sum_{k=1}^{n} \frac{1 - x_k}{1 - \theta}$$

$$\nabla ln(P(D_1 \mid \theta)) = 0$$

$$\text{This simplifies to: } \theta = \frac{\sum_{k=1}^{n} x_k}{n}$$

$$\theta = \frac{8}{10} = \mathbf{0.8}$$

This is for the Normal distribution: $P(D_2 \mid \theta) = \prod_{k=1}^{n} \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_2 - \mu}{\sigma})^2}$

$$\mu_{ML} = \frac{1}{N} \sum_{k=1}^{N} x_k = \mathbf{186.6}$$

$$\sigma_{ML} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (x_k - \mu_{ML})^2} = \mathbf{14.83}$$

0.0.21 Q7.1 What is the correct conjugate prior?

Given the likelihood function you already wrote down in Q6.2, what's the proper conjugate prior distribution for that likelihood function? Why that one?

Points: 0.3

The correct conjugate prior distribution to Q6.2's likelihood function would be a Beta distribution.The reas

0.0.22 Q7.2 Write down that Bayes equation for calculating the posterior distribution $P(\theta|D)$

Do this in terms of the likelihood fuction and prior you have selected. Use n as the number of flips and x as the number of heads in those flips.

Points: 0.3

$$
\begin{aligned}
n = 10, x = 8 \\
P(\theta \mid D) &= \frac{P(D \mid \theta)P(\theta)}{\int P(\theta)P(D \mid \theta)d\theta} = \frac{P(D \mid \theta)P(\theta)}{P(D)} \\
P(\theta \mid D) &\propto P(D \mid \theta)P(\theta) \\
&= \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\
&= \theta^{x+a-1}(1-\theta)^{n+b-1-x}
\end{aligned}
$$

0.0.23 Q7.3 Special interprestation of prior parameters

There is a special interpretation of the parameters for the prior distribution, where we can talk about them as a virtual coin flipping experiment. Describe the relationship in words. Tell us WHY this works in either words or math, your choice.

Hint: see the slides from lecture, or if you simplified your answer to 7.2 above you can see it there

Points: 0.2

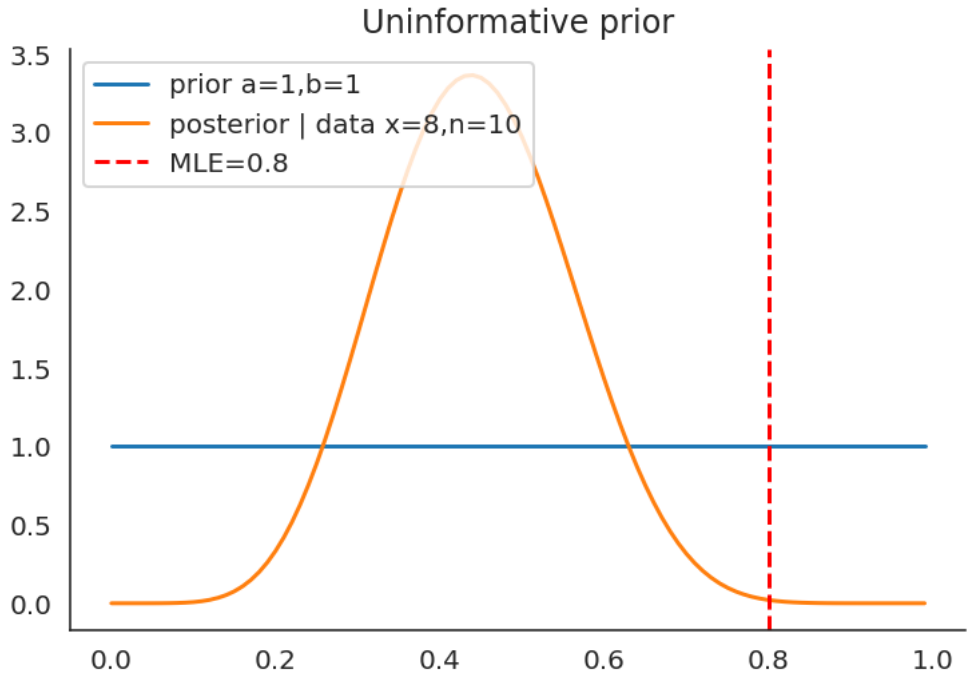
The parameters in the prior distribution have an effect on the fairness of a coinif we were to consider a virt

0.0.24 Q7.5 Graph Interpretation

Describe what you see in the previous graph. Which distribution the uninformative prior looks like? How does prior influence the posterior? How does the MLE similar or different from the posterior distribution with this particular prior?

Points: 0.2

In [4]: `plot_74()`



I see a line for the mean for of the likelihood, a line representing the PDF of the prior and a PDF representing the posterior. The uninformative prior looks like a uniform distribution. The prior has no influence over the posterior. The MLE is different from the posterior's mean. The posterior's mean looks to be around 0.45 and the MLE is 0.8.

0.0.25 Q8.1 Intuition of MAP

What are the similarities between MLE and MAP? What does MAP take into account that MLE does not?

What are the similarities between MAP and Bayesian estimation? What does Bayesian take into account that MAP does not?

Points: 0.4

1) MLE and MAP are both methods used for parameter estimation and do this through maximizing the likelihood. MAP offers more insight into the belief of the parameters while MLE does not.

2) MAP and BE both use prior and to find the posterior. BE takes into account evidence to update our beliefs while MAP does not.

0.0.26 Q8.2 MAP estimate for the coin flipping dataset

Recall that the posterior of the coin flipping task is the same kind of distribution as the prior.

What kind of distribution is the posterior? What is the generic equation for the peak (i.e., mode) of that distribution? In words, describe why that is the peak.

Write down the equations for the MAP estimate of the coin flipping task. Parameterize it in terms of the prior parameters `a,b` and the likelihood parameters `x` (the number of heads) and `n` (the number of total flips). It may help you to look at the answer to Q7.2

Points: 0.4

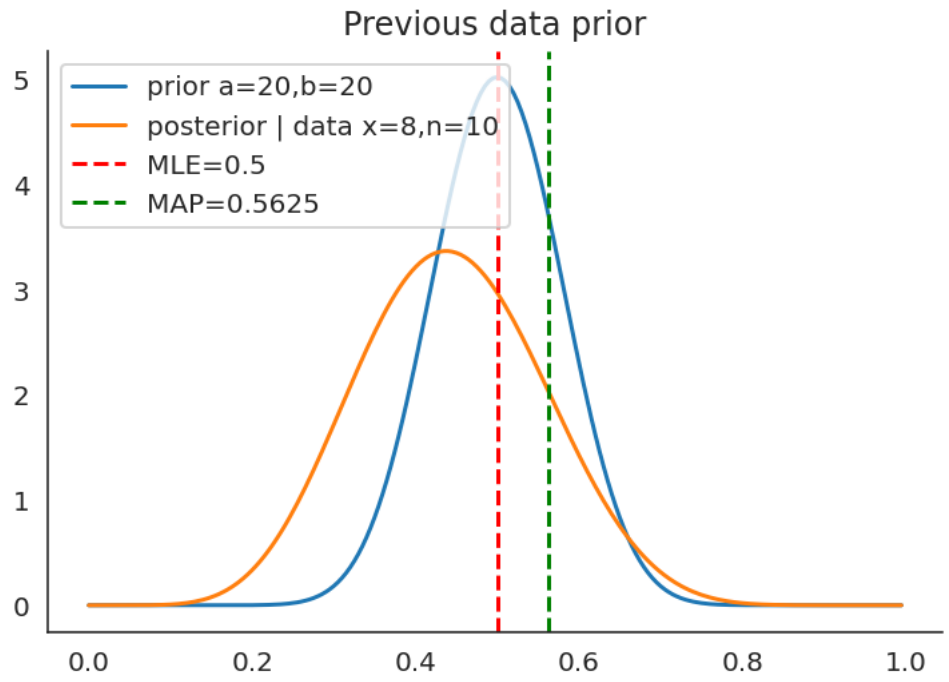
The posterior is a Beta distribution. This is the generic equation:

0.0.27 Q8.4 Graph Interpretation

Describe what you see in the previous graph. How does prior influence the posterior? Compare the MLE and MAP estimate. Explain why the two estimate value are different, or the same.

Points: 0.4

In [7]:
```
plot_83()
```



The prior moved the posterior's mean more to the left. The MLE and MAP estimates are different because MAP were calculated with the posterior in mind.