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

Points: 0.25

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.29}{0.48} \approx 0.604167$

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

Points: 0.25

$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.604167*0.48}{0.73} \approx 0.397260$

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

Points: 0.25

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

A and B are not independent

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

Points: 0.2

R = chose red box
G = chose green box
B = chose blue box
H = pick a 100-dollar bill
O = pick a 1-dollar bill

$P(H) = P(H|R)P(R) + P(H|G)P(G) + P(H|B)P(B) = \frac{1}{10} * \frac{6}{10} + \frac{1}{2} * \frac{3}{10} + \frac{9}{10} * \frac{1}{10} = 0.3$

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

Points: 0.2

$P(O) = 1 - P(H) = 1 - 0.3 = 0.7$

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

$P(G|H) = \frac{P(H|G)P(G)}{P(H)} = \frac{1}{2} * \frac{3}{10} * \frac{10}{3} = \frac{1}{2}$

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?

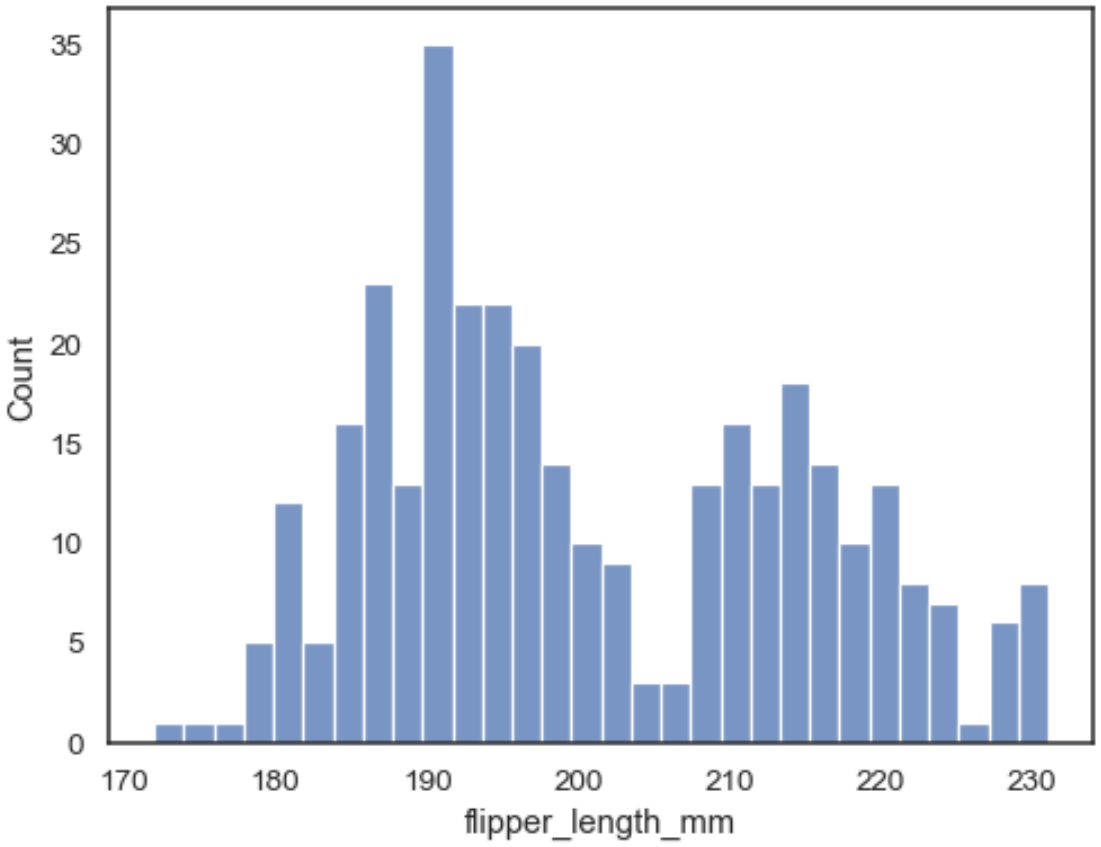
Points: 0.2

W = random variable; worth of dollar bill

$E[W] = \sum_{x=0}^{\infty} xP(W = x) = 100 * 0.3 + 1 * 0.7 = \$30.7$

The expected value of the selected bill is \$30.7, which means that if the outcome of the experiment (value of the dollar bill selected) is averaged across many trials, it will converge to the value of \$30.7.

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

In my opinion, this data would not be well-fit by a single gaussian distribution, as there is not a clear mean around which most of the values are distributed, and the overall shape of the graph doesn't resemble a gaussian. Instead, I would model the data by two normal distributions, because the graph seems to show 2 peaks/means with a high concentration of data ('count'), that diminish with further distance from the relative mean.

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 distribution. A fair coin lands heads 50% of the time on average, so to model whether a coin is fair or not, one must consider the number of heads relative to the total number of flips (n). The distribution that counts the number of wins in n independent trials is the binomial.
2. Poisson distribution. The poisson distribution is suitable to estimate the frequency of costumers entering a store, because it models the probability of an event happening n times during a fixed time interval.
3. Normal distribution. Since the number of college basketball players is large, the central limit theory holds. According to the theorem, any distribution tends to the Gaussian when the sample size tends to infinity, hence a normal distribution is appropriate. The mean of the distribution, however, will be higher than the mean height of the general population (the distribution will likely be left skewed)
4. Normal distribution. Similarly to the previous point, the central limit theorem suggest a normal distribution is appropriate due to the large sample size.

0.0.10 Q4.1

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

Points: 0.25

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

0.0.11 Q4.2

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

Points: 0.25

$$E[x] = 0 * \mu^0(1-\mu)^1 + 1 * \mu^1(1-\mu)^0 = \mu$$

0.0.12 Q4.3

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

Points: 0.35

$$Var[x] = E[x^2] - E[x]^2 = 0 * \mu^0(1-\mu)^1 + 1 * \mu^1(1-\mu)^0 - \mu^2 = \mu - \mu^2 = \mu(1-\mu)$$

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

$$P(\text{car} = 1 | \text{choose=1,open=2}) = \frac{P(\text{choose=1,open=2} | \text{car} = 1) P(\text{car} = 1)}{P(\text{choose=1,open=2})} = \frac{\frac{1}{6} * \frac{1}{3}}{a} = \frac{1}{18\text{a}}$$

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

$$P(\text{car} = 3 | \text{choose} = 1, \text{open} = 2) = \frac{P(\text{choose=1,open=2} | \text{car} = 3) P(\text{car} = 3)}{P(\text{choose=1,open=2})} = \frac{\frac{1}{3} * \frac{1}{3}}{a} = \frac{1}{9\text{a}}$$

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

If the player chooses door 1 and Monty opens door 2, there is a higher probability that the car is behind door 3 compared to door one (twice as much!), hence the player is more likely to win if he switches the door.

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

Points: 0.4

- Coin flips dataset: Bernoulli distribution. The bernoulli distribution models the probability of a single coin flipping heads (x = 1). In this dataset there are 10 statistically independent coin flips, hence the 10 Bernoulli random variables can be multiplied together to describe the dataset.
- Male heights: Normal distribution. Since the number of male human beings tends to infinity, the central limit theorem applies, suggesting that male heights is best model by a gaussian distribution.

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

$$\text{Bernoulli distribution: } B = Ber \sim (10, \theta)$$

$$P(D_1|\theta) = \prod_{k=1}^{10} \theta^{x_k}(1-\theta)^{1-x_k}$$

$$\text{Normal distribution: } H = N \sim (\theta, \sigma^2)$$

$$P(D_2|\theta) = \prod_{k=1}^{10} (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}\frac{(x_k-\theta)^2}{\sigma^2})$$

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

$$ln(P(D_1|\theta)) = ln(\prod_{k=1}^{10} \theta^{x_k}(1-\theta)^{1-x_k}) = \sum_{k=1}^{10} ln(\theta^{x_k}(1-\theta)^{1-x_k}) = \sum_{k=1}^{10} ln(\theta^{x_k}) + ln((1-\theta)^{1-x_k}) = \sum_{k=1}^{10} x_k ln(\theta) + (1-x_k) ln(1-\theta)$$

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

$\theta^* = \arg\max_x P(D_1|\theta)$
The equation above means that the maximum likelihood estimate of θ is the parameter θ that maximizes the likelihood function $P(D_1|\theta)$. This is done in the following way: 1. Replace the likelihood function with the log likelihood function $ln(P(D_1|\theta))$. The log likelihood function allows for easier manipulation as it involves a sum and not a product over each data point. Most importantly, the log function is convex by definition, hence it is easy to find the optimum. 2. Find the gradient of the log likelihood function with respect to θ, because the optimum occurs where the gradient equals zero. $\frac{\partial(D_1|\theta)}{\partial\theta} = \sum_{k=1}^{10} \frac{x_k}{\theta} + \frac{1-x_k}{1-\theta}$ 3. Set the gradient equal to 0 and solve for θ. This simplifies to $\theta = \frac{1}{n}\sum_{k=1}^{n} x_k$, the expectation of the distribution.

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

Coin flip dataset: $\theta^* = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{10}\sum_{k=1}^{10} x_k = \frac{8}{10} = 0.8$

Male height dataset: $\theta^* = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{10}\sum_{k=1}^{10} x_k = \frac{202.+193.+190.+189.+210+188.+175.+185.+152.+182.}{10} = 186.6$

The maximum likelihood estimate is the empirical mean/expectation of the data points observed.

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

Conjugate prior of bernoulli distribution: beta distribution $B \sim Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

A beta prior is the correct pick because, when a beta distribution is multiplied by a bernoulli one, the result is another beta distribution (definition of conjugate distributions). This is important because a beta prior multiplied by a bernoulli likelihood results in a beta posterior, which means that the observed data does not change the type of distribution, making the math and computation easier.

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

$p(\theta|x) = p(x|\theta)p(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+b-1}$

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

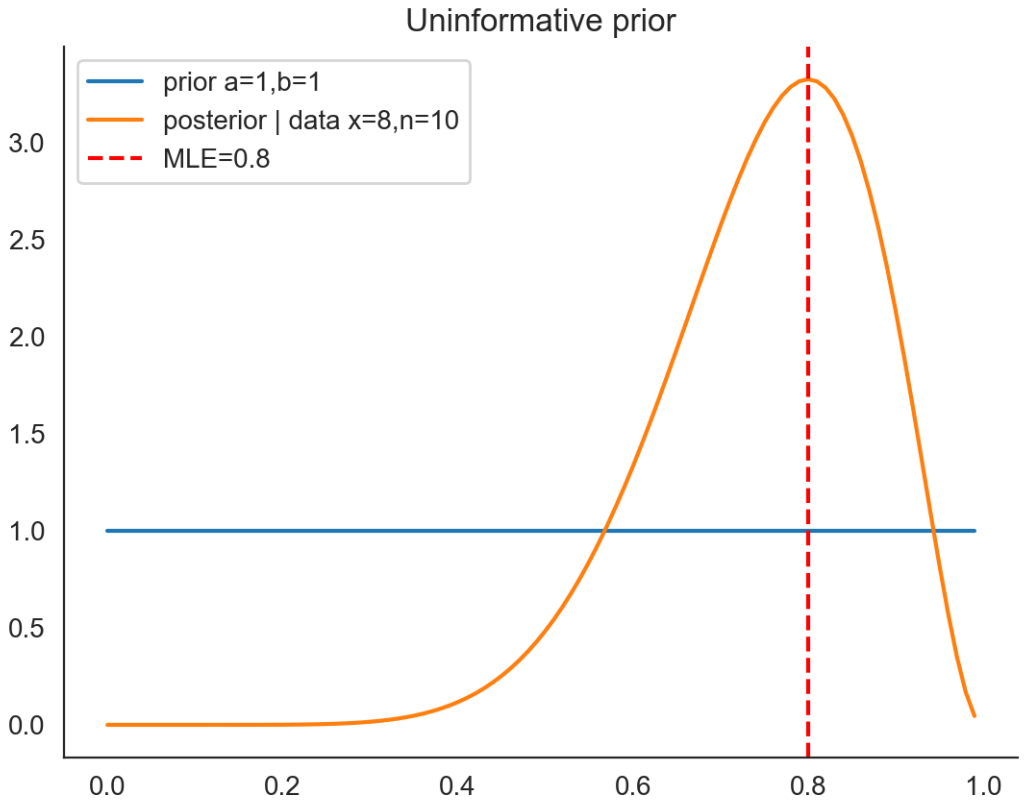
As the simplified answer above shows, the posterior obtained by multiplying the bernoulli likelihood with the beta prior, is itself a beta distribution. Besides a multiplicative constant, the only difference between the bernoulli likelihood and the beta posterior is the addition of $(\alpha - 1)$ and $(\beta - 1)$ to the exponents of θ and $(\theta - 1)$ respectively. Since θ is the probability of a coin landing heads, this can be interpreted as adding $(\alpha - 1)$ heads and $(\beta - 1)$ tails to the original dataset.

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()`



In the plot above, there is a prior distribution (blue line); a posterior, beta-shaped distribution with mean around 0.85 (orange line); and the maximum likelihood estimate of 0.8 (red line). The uninformative prior looks like a uniform distribution; The prior does not influence the posterior, as a uniform distribution assigns the same probability to each data point. In the case of an uninformative prior (Beta$(1, 1)$), the MLE and the x-value corresponding to the peak of the posterior (given by Bayesian estimation) are the same. This is because the uninformative prior contributes no information to the Bayesian estimates, which is therefore only given by the empirical data.

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

MLE and MAP are similar in that they are both ways to find the best parameter to characterize a distribution. They both compute a single estimate instead of the full distribution, but whereas MLE is an estimate entirely based on empirical data, MAP integrate the empirical data with prior beliefs about the subject.

MAP and Bayesian estimation are related: Bayesian estimation is an algorithm to compute the entire distribution of parameters that maximize the posterior given the observed data and the prior; and the Maximum A Posteriori estimate is a method for computing the optimal parameter to maximize the posterior given the data observed and the prior. The MAP is the value corresponding to the peak of the posterior distribution. Hence, whereas Bayesian takes into account the full distribution, MAP is concerned with one value only.

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

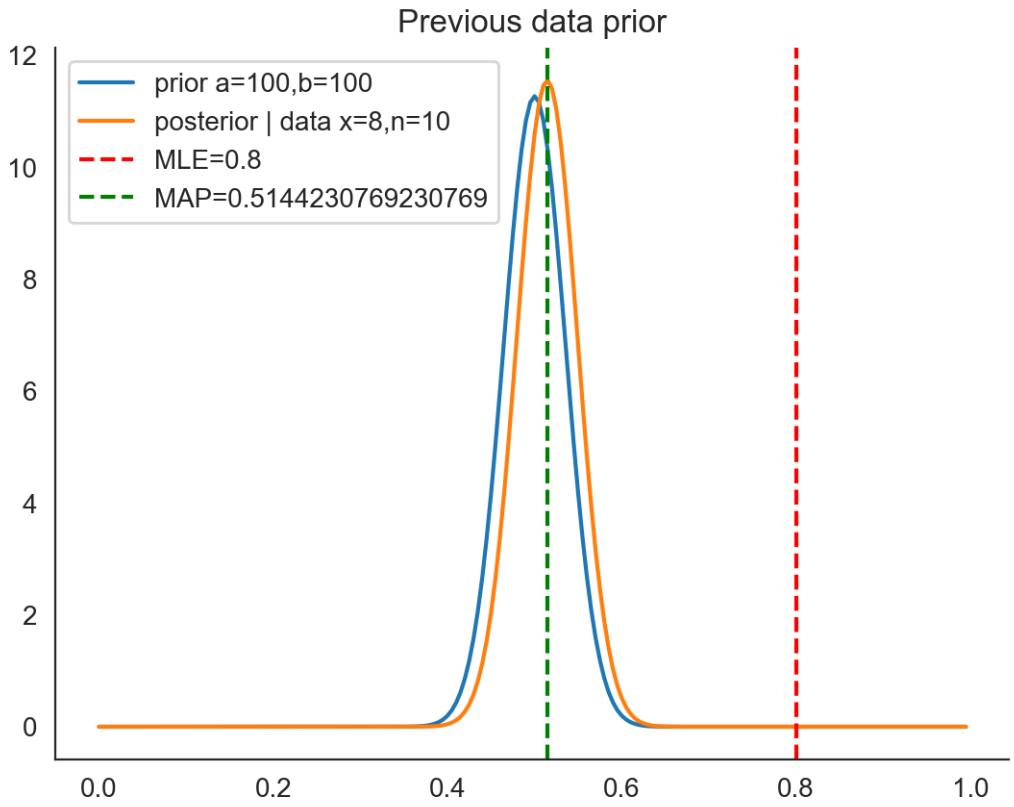
The posterior is a beta distribution with parameters $(a + x, b + n - x)$
The generic equation for the peak of the beta distribution is: $\frac{\alpha-1}{\alpha+\beta-2}$ for B $\sim$ Beta(α, β). This is the result of taking the derivative of the posterior distribution equation with respect to θ, setting it equal to 0, and solving for θ.
MAP estimate for coin flipping task: $\theta^* = \arg\max_x P(\theta|D_1) = \frac{x+a-1}{n+a+b-2}$

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()
```



In the graph above, there is a prior showing the distribution of flip outcomes of a fair coin (peak at 0.5); The MLE, which estimates the best parameter underlying the distribution given the empirical data only (MLE = 0.8); The posterior, which is an updated distribution that integrated the prior and the data observed; And the MAP, the estimation of the best parameter underlying the distribution given the empirical data and prior knowledge (MAP $\sim$ 0.514).

The prior influences the posterior as it prevents the empirical data to overly influence the distribution. This is reflected in the discrepancy between the MLE and MAP estimates. The MLE is greater than the MAP

because it is the empirical mean of the data, which in this dataset are highly skewed towards heads. On the other hand, the MAP takes into account the fact that in a small dataset, the chance of observing a skewed ratio of heads and tails is considerable.