JSC270 Assignment 3

Simulation (with `numpy`)

February 10, 2021

Submission Deadline: Monday, March 8th at 11:59AM EST

What and where to submit:

- A PDF report containing your responses, including any figures or derivations.
- A Google Colab notebook containing any relevant python code. Please provide a link to this notebook in your PDF report. Optionally, you may place the code on GitHub, and instead provide a link to that repo.

The PDF report is to be submitted through Quercus. This assignment is to be completed individually.

Grading Scheme: This assignment contains 3 questions, for a total of 50 marks, plus 2 optional bonus points.

1. Bayesian Simulation (20 pts)

Recall that the Bayesian approach to modelling assumes we have some prior belief about an unknown parameter of interest (say, θ). Our uncertainty about the true value is captured in a distribution called a **prior**. Instead of treating data as random, and assuming a fixed parameter to explain the sample we see, we assume the parameter θ is random, and we use a given sample of data (X) as evidence to update our belief about θ's true distribution. This process of updating beliefs is done via Bayes' Rule:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int_\theta f(x|\theta')f(\theta')d\theta'}$$

Each of the terms in the rule has its own name:

$$posterior \ = \ \frac{likelihood \cdot prior}{normalizing \ constant}$$

Where the **prior** encodes previous beliefs about θ, and the **likelihood** captures how well the data fits our current beliefs. The **posterior** distribution then represents our new (updated) belief about θ, given the data we have observed. The **normalizing constant** is often ignored, but is needed to ensure that the posterior integrates to 1. Although there are several ways of estimating each component, for this question, we'll focus on how the choice of prior affects the posterior.

A) (4 pts) Figure 1 shows 4 common types of statistical distributions. For each of the four quantities given below, state which of the four distributions is the best fit[1]. For each choice, explain your answer (possibly including why other distributions do not work well).

[1]Note: I do not claim that these are the exact distributions of the given quantities. This exercise is just to show that what constitutes a reasonable choice of prior depends on the context.

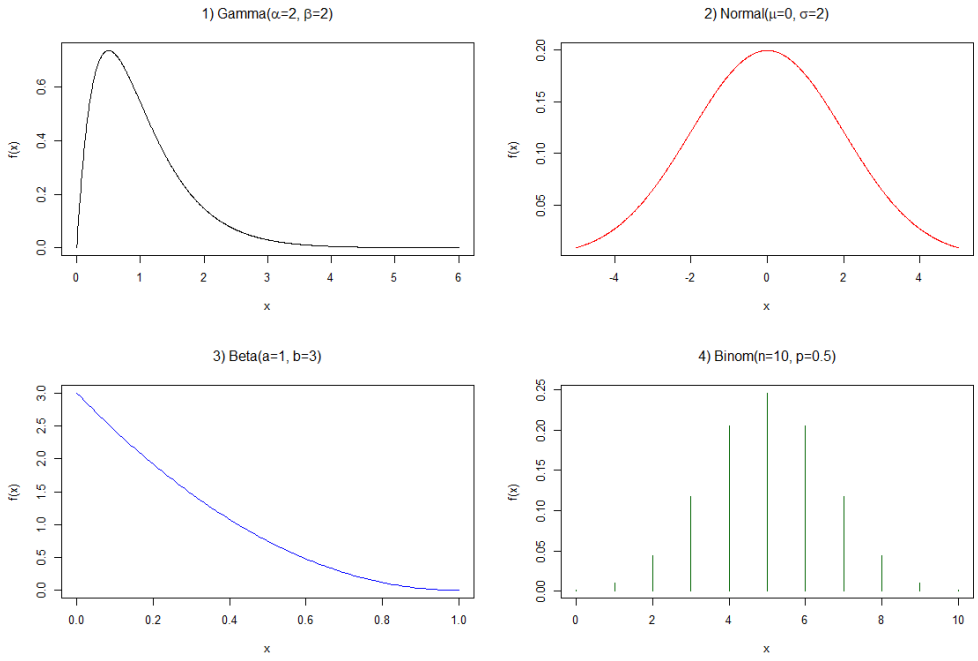

Figure 1: Some possible choices of prior.

i) The average price of a cup of coffee in Toronto
ii) The proportion of individuals who catch a cold in February.
iii) The number of people (among a small group) who own a dog.
iv) The average weekly change in US unemployment (measured in hundreds of individuals).

For parts B-G, assume the following scenario:

Suppose we have a sample of n individuals, and we're interested in measuring the number of persons who contract COVID 19. We know (roughly speaking), that each person will either contract the disease or not (i.e. the outcome is binary). Thus, $X_i = 1$ if individual i contracted the disease, or 0 if not. If we assume that individuals contract the disease independent of one another, the **likelihood** function of our n individuals is given by:

$$f(x_1, \cdots, x_n|\theta) = f(x|\theta) = (\theta)^{\Sigma_{i=1}^n x_i}(1-\theta)^{n-\Sigma_{i=1}^n x_i}$$

Note that this is just the joint density of our n data points, assuming each point follows a Bernoulli(θ) distribution. But how do we reasonably estimate θ? Let's assume that $\theta \sim Beta(a, b)$. In other words, our prior distribution is given by:

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

for $\theta \in (0, 1)$. Note that $\Gamma(c) = (c-1)!$, and that, in general, the Beta distribution has expectation equal to $E(\theta) = \frac{a}{a+b}$.

B) (3 pts) Using Bayes rule, show that the posterior distribution is $Beta(a+\Sigma_{i=1}^{n}x_i, b+n-\Sigma_{i=1}^{n}x_i)$. *Hint: You don't actually need to integrate to obtain the normalizing constant. Just use the definition of the beta distribution, and the fact that the posterior must integrate to one (i.e.* $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta=1$*).*

C) (1 pt) What is the expectation of the posterior distribution (i.e. the posterior mean)? *Note: No need for a full derivation here, you can just state (and possibly simplify) your answer.*

D) (2 pts) Suppose $a=b=1$. This gives a special case of the beta distribution. What is the name of this special case? In using this special prior distribution, what are we saying about our beliefs across possible values of θ?

E) (4 pts) Along with the values of a and b above, suppose that we have a sample of 20 individuals, 4 of whom have contracted the disease (i.e. $n=20$, $\Sigma_{i=1}^{n}x_i=4$. Plot the prior and posterior distributions. You may find the `numpy` function `np.random.beta()` useful here. How does the posterior mean compare to 1) the prior mean, and 2) the traditional likelihood estimate, $\frac{1}{n}\Sigma_{i=1}^{n}x_i$?

F) (3 pts) Suppose scientists are fairly certain that the contraction rate of COVID is close to that of the flu, which is fairly low. To encode this belief, we change the hyperparameters of our prior to $a=1, b=3$. Using the same sample as before, plot the new prior and posterior distributions. What are the new prior and posterior means, and how do they compare to those you found in part E?

G) (3 pts) Now suppose we have the same prior as in F, but we get a different sample: $n=20$, $\Sigma_{i=1}^{n}x_i=12$. Plot the prior and posterior distributions. Compare the prior and posterior means to the likelihood estimate defined earlier. How do these results compare to those you obtained in parts E and F?

H) (2 bonus pts) In this question, the fact that the prior and posterior were from the same family of distributions is not a coincidence. What is the name for the class of priors that has this property? Can you give another example of such a prior (and the necessary likelihood)?

2. Asymptotic Behavior (15 pts)

Suppose we have a random variable x that follows an **Exponential Distribution**. Thus, X has the following density:

$$f(x)=\lambda e^{-\lambda x}$$

for some parameter λ, and is defined over $x\in[0,\infty)$. Recall also the definition for the expectation (mean) of a random variable: $E(X)=\int_{-\infty}^{\infty}x\cdot f(x)dx$.

A) (4 pts) Determine the expectation of X, using the definition given (*Hint: Your answer should be a function of* λ *only*). What is the exact value of the expection of X when $\lambda=4$?

B) (6 pts) Suppose now that we know the exact distribution of X is given by $\lambda = 2$. In python, generate 6 different random samples from this distribution with sizes $\{10, 20, 50, 100, 500, 2000\}$. You may find the function `np.random.exponential()` from the numpy package useful here (note that the parameter argument for this function is inverted, ie $\lambda = \frac{1}{\beta}$). Plot the distribution of each sample (i.e. 6 plots total, or optionally 1 plot with 6 curves).

C) (2 pts) Report the means of each of the samples in B (`numpy` has a handy `np.mean()` function for this). What is the theoretical mean when $\lambda = 2$? What happens to the sample mean as sample size grows?

D) (2 pts) What is the name of the statistical 'law' that explains the sample average behavior you've just seen?

3. Bootstrapping (15 pts)

In this question, you'll apply the statistical process of **bootstrapping** from scratch. In class, we saw bootstrapping applied to decision trees (ie bagging), but its uses expand well beyond classification. In general, bootstrapping is quite valuable in cases where repeated sampling from a population is prohibitively expensive, or noisy.

Suppose we have a sample of iid data, $X_1, \cdots, X_n$ from a $N(\mu, \sigma^2)$ distribution. Let $\bar{X} = \frac{1}{n}\Sigma_{i=1}^{n} X_i$ be the sample mean. Our goal here is to estimate the true mean (μ) using the sample mean.

A) (2 pts) What is the theoretical distribution of the sample mean ($\bar{X}$)? *No need to show full derivations here, but for full marks, the parameters should be specified (i.e. $\bar{X} \sim Beta(a = 2\mu, b = \frac{\mu}{\mu+\sigma})$).*

B) (1 pts) Suppose $\mu = 2$, and $\sigma^2 = 4$. Simulate a single sample of 1000 data points using these parameters. You might find the `numpy` function `np.random.normal()` useful here.

C) (5 pts) Now implement the bootstrapping algorithm described above. The strategy here, using the main sample you just generated, is to:

i) Generate B resamples from the main sample (sampling with replacement)
ii) Compute the mean of each resample

For this question, you won't need the resamples themselves, only their means. To help you get started, we've provided some (rough) pseudocode below. Use a resample size of $M = 100$, and compute $B = 5000$ resample means. **Your result should be a single vector** (or list/array/etc. if you used a different data structure) **of length B**. You might find the function `np.random.choice()` helpful.

Provide a plot showing a histogram of the bootstrapped sample means (i.e. the resample means).

D) (2 pts) What is the distribution of bootstrapped sample means, and how does it compare to the true distribution of $\bar{X}$? In addition to `np.mean()`, you might also use `np.std()` here. *Hint: The normal distribution is fully characterized by its mean and variance.*

Algorithm 1: Bootstrapping the Sample Mean

```
Instantiate array of size B to store resample means
for j = 1 : B do
    Generate resample j of size M from main sample
    Compute mean of resample
    array[j] ← computed resample mean
end for
```

E) (3 pts) Suppose we didn't know the true distribution of X, but were able to obtain a sample like the one you generated in part B. Assume also that you have obtained B = 5000 resample means using your main sample. How might you generate a 95% Confidence Interval for μ, the population mean? *Note: You don't have to compute anything for this question, just describe what you would do.*

F) (2 pts) Using your method from part E, state the 95% Confidence Interval for the population mean based on the bootstrapped means collected in part C.