

# Data 102, Fall 2023

## Homework 2

Due: **5:00 PM** Friday, September 29, 2023

### Submission Instructions

Homework assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

#### Written Portion:

- Every answer should contain a calculation or reasoning.
- You may write the written portions on paper or in  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ .
- If you type your written responses, please make sure to put it in a markdown cell instead of writing it as a comment in a code cell.
- Please start each question on a new page.
- It is your responsibility to check that work on all the scanned pages is legible.

#### Code Portion:

- You should append any code you wrote in the PDF you submit. You can either do so by copy and paste the code into a text file or convert your Jupyter Notebook to PDF.
- Run your notebook and make sure you print out your outputs from running the code.
- It is your responsibility to check that your code and answers show up in the PDF file.

#### Submitting:

You will submit a PDF file to Gradescope containing all the work you want graded (including your math and code).

- When downloading your Jupyter Notebook, make sure you go to File → Save and Export Notebook As → PDF; do not just print page from your web browser because your code and written responses will be cut off.
- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so. As a Berkeley student, you get free access to Adobe Acrobat, which you can use to merge as many PDFs as you want.
- Please see this guide for how to submit your PDF on Gradescope. In particular, for each question on the assignment, please make sure you understand how to select the corresponding page(s) that contain your solution (see item 2 on the last page).

Late assignments will count towards your slip days; it is your responsibility to ensure you have enough time to submit your work.

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

## The One with all the Beetles

- (9 points) Cindy has an inordinate fondness for beetles and for statistical modeling. She observes one beetle everyday and keeps track of their lengths. From her studies she feels that the beetle lengths she sees are uniformly distributed. So she chooses a model that the lengths of the beetles come from a uniform distribution on  $[0, w]$ : here  $w$  is an unknown parameter corresponding to the size of the largest possible beetle. Since the maximum size  $w$  is unknown to her, she would like to estimate it from the data. She observes lengths of  $n$  beetles, and calls them  $x_1, \dots, x_n$ .

- (1 point) What is the likelihood function of the observations  $x_1, \dots, x_n$ ? Express your answer as a function of the largest size parameter  $w$ .

**Hint:** Your answer should include the indicator function  $\mathbb{1}(\max_i x_i \leq w)$ . To see why, consider what happens if  $w = 3$  cm and  $x_1 = 5$  cm.

- (2 points) Use your answer from Part (a) to explain why the maximum likelihood estimate (MLE) for  $w$  is the maximum of the observed lengths, that is,

$$\hat{w}_{\text{MLE}} = \max\{x_1, x_2, \dots, x_n\}$$

**Hint:** You don't need to use calculus; a qualitative argument would suffice.

- (2 points) Cindy decides to instead use a Bayesian approach. She has a prior belief that  $w$  follows a *Pareto distribution* with parameters  $\alpha, \beta > 0$ . We can write:

$$w \sim \text{Pareto}(\alpha, \beta)$$

Then the density function of  $w$  is

$$p(w) = \frac{\alpha\beta^\alpha}{w^{\alpha+1}} \mathbb{1}(w \geq \beta)$$

Show that the posterior distribution for  $w$  is also a Pareto distribution, and compute the parameters as a function of  $\alpha, \beta$ , and the observations  $x_1, \dots, x_n$ .

- (2 points) Provide a short description in plain English that explains what the parameters of the Pareto distribution mean, in the context of the Pareto-uniform conjugate pair.

**Hint:** For the Beta-Binomial conjugate pair that we explored in class, the answer would be that the Beta parameters acted as pseudo-counts of observed positive and negative examples.

- (e) (2 points) Cindy started with the initial prior with parameters  $\alpha = 1$  and  $\beta = 10$  on day 0. Use the code in `beetledata.py` to generate the data for the lengths of the beetles she sees, starting from Day 1 to Day 100. Use the data to make a graph of one curve for each of the days 1, 10, 50 and 100 (so four curves total), where each curve is the probability density function of Cindy's posterior for the respective day. **Note:** For the Pareto distribution, code the density function by hand rather than relying on the distribution provided by `scipy`.
- (f) (0 points) (Optional) Use `pymc3` to sample from the posterior for days 1, 10, 50 and 100 and plot a density function for each of the cases. Compare the results from the analytic and simulation based computation of the densities.

## Bayesian Fidget Spinners

2. (8 points) Nat's company manufactures fidget spinners. The company uses two factories, which we'll call factory 0 and factory 1. Each fidget spinner from factory  $k$  is defective with probability  $q_k$  ( $k \in \{0, 1\}$ ). Nat knows that factory 0 produces fewer defective fidget spinners than factory 1 (in other words,  $q_0 < q_1$ ).

She receives  $n$  boxes full of fidget spinners, but the boxes aren't labeled (in other words, she doesn't know which box is from which factory). For each box, she starts randomly pulling out fidget spinners until she finds a defective one, and records how many fidget spinners she pulled out (including the defective one). She calls this number  $x_i$  for box  $i$ , for  $i = 1, \dots, n$ .

She wants to estimate the following pieces of information:

- Which boxes came from factory 0, and which came from factory 1? She defines a binary random variable for each box  $z_i$  with the factory label (i.e.,  $z_i = 0$  if box  $i$  is from factory 0, and  $z_i = 1$  if box  $i$  is from factory 1).
- How reliable is each factory? In other words, what are  $q_0$  and  $q_1$ ?

Inspired by what she learned about Gaussian mixture models, she sets up the following probability model:

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\pi) \\ q_k &\sim \text{Beta}(\alpha_k, \beta_k) \\ x_i | z_i, q_0, q_1 &\sim \text{Geometric}(q_{z_i}) \end{aligned}$$

- (a) (1 point) Draw a graphical model for the probability model described above if  $n = 2$  (i.e., there are only two boxes of fidget spinners).

Nat decides to implement the model above setting the following hyperparameters:

$$\pi = 0.45, \quad q_0 \sim \text{Beta}(1, 5), \quad q_1 \sim \text{Beta}(5, 1)$$

- (b) (2 points) The choices of the parameters in Nat's model represents her beliefs about the factories.

- i. (1 point) From her choice for  $\pi$ , what can we infer about her beliefs about the two factories?
  - ii. (1 point) Similarly, from her choices for  $\alpha$  and  $\beta$ , what can we infer about her beliefs about the factories?
- (c) (5 points) Use `fidget_data.py` to generate the data that Nat observes, then, using PyMC3, fit the model outlined above, setting the hyperparameters to the values that Nat chose. Obtain 1000 samples from the posterior distribution  $p(q_0, q_1 | x_1, \dots, x_n)$ , and generate a scatterplot (one point per sample).

You can use the code below (also provided to you in `fidget_model.py`) to help you get started.

```

1 import pymc3 as pm
2
3
4 alphas = ...
5 betas = ...
6 pi = ...
7
8 with pm.Model() as model:
9     z = pm.Bernoulli(
10         # Define the Bernoulli Model Here
11     )
12
13     # Hint: you should use the shape= parameter here so that
14     # q is a PyMC3 array with both q0 and q1.
15     q = ...
16
17     # Hint: it may be useful to use "fancy indexing" like we did in
18     # class.
19     # See below for an example
20     X = pm.Geometric(
21         # DEFINE THE GEOMETRIC MODEL HERE
22     )
23
24     trace = ....
25
26 # FANCY INDEXING
27 my_binary_array = np.array([0, 0, 1, 1, 0, 1])
28 my_real_array = np.array([0.27, 0.34])
29 print(my_real_array[my_binary_array])

```

- i. Under the posterior, what is the probability that factory 0 produces more boxes than factory 1?
- ii. What is your median estimate of factory 0's defect rate, based on the samples from the posterior?

## Baseball Average Prediction

3. (8 points) The following historical dataset is famous in the field of statistics ever since it was used by Brad Efron and Carl Morris to illustrate the James-Stein estimator and the Stein shrinkage phenomenon (see, for example, the Scientific American paper titled “Stein’s Paradox in Statistics” by Efron and Morris, or Section 1.2 of Efron’s book on Large Scale Inference).

In baseball, an **at-bat** (AB) is a hitter’s turn batting against a pitcher. In each at-bat, the hitter can reach (or pass) first base on a **hit** (H). **Batting average** is used to measure a hitter’s success and is calculated as the fraction  $AVG = \frac{H}{AB}$ .

The `baseball.csv` dataset (shown in Table 1) contains 18 rows and 3 columns. Each row represents a baseball player and contains the following information:

- The player’s name
- The player’s number of *hits* (H) in the first 45 *at-bats* (AB)
- The player’s *End of Season Batting Average* (EoSAverage), calculated as the proportion of hits over the total number of at-bats over the entire season

For example, the first row shows that Clemente had 18 hits in his first 45 at-bats and a .346 EoSAverage.

The goal is to use the players’ early season performance (as indicated by the second column) to predict their end of season performance (as indicated by the third column).

| Player Name | Number of Hits in the first 45 At-Bats | EoSAverage |
|-------------|----------------------------------------|------------|
| Clemente    | 18                                     | .346       |
| F Robinson  | 17                                     | .298       |
| F Howard    | 16                                     | .276       |
| Johnstone   | 15                                     | .222       |
| Berry       | 14                                     | .273       |
| Spencer     | 14                                     | .270       |
| Kessinger   | 13                                     | .263       |
| L Alvarado  | 12                                     | .210       |
| Santo       | 11                                     | .269       |
| Swoboda     | 11                                     | .230       |
| Unser       | 10                                     | .264       |
| Williams    | 10                                     | .256       |
| Scott       | 10                                     | .303       |
| Petrocelli  | 10                                     | .264       |
| E Rodriguez | 10                                     | .226       |
| Campaneris  | 9                                      | .286       |
| Munson      | 8                                      | .316       |
| Alvis       | 7                                      | .200       |

Table 1: Some Statistics of 18 Baseball Players from the 1970 Season

- (a) (1 point) For the  $i^{\text{th}}$  player, we model their number of hits in the first 45 at-bats  $H_i$  as

$$H_i \sim \text{Bin}(45, \theta_i),$$

where  $\theta_i$  is their EoSAverage. This model places the problem of predicting EoSverages (based on hits in the first 45 at-bats) inside the framework of estimating probabilities in a Bernoulli/Binomial model. Is this a sensible model? Why or why not?

- (b) (1 point) Calculate the mean squared error (MSE) of the naive proportion estimates of  $\theta_i$  given by  $\hat{\theta}_i = \frac{H_i}{45}$ . Note that you are given values of  $\theta_i$  in the last column of Table 1.
- (c) (2 points) The goal now is to compare the naive estimate with Bayes estimates. To calculate Bayes estimates, we shall use a suitable Beta( $a, b$ ) prior.

To find the appropriate  $a$  and  $b$ , use the following procedure:

- Ignore the top four players as well as the bottom four players in Table 1, as these players have either performed exceptionally well or exceptionally poorly in the first 45 at-bats so their current averages may not be reflective of their EoSverages.
- Calculate the mean  $m$  and variance  $v$  of the remaining 10 players.
- Find  $a$  and  $b$  such that the mean and variance corresponding to the Beta( $a, b$ ) distribution matches with  $m$  and  $v$ .

Report the values of  $a$  and  $b$ , and plot the Beta( $a, b$ ) density function.

- (d) (1 point) Calculate the Bayes estimates using the posterior mean for each  $\theta_i$  using the Beta( $a, b$ ) prior from the previous part.
- (e) (1 point) Calculate the MSE of the Bayes estimates you calculated in part (d). The MSE of these estimates should be much smaller than the MSE of the naive proportions from part (b).
- (f) (2 points) The naive estimates and the Bayes estimates differ in one crucial aspect. The naive estimate of the EoSAverage for a player only uses data on this player's current record. On the other hand, the Bayes estimate uses also data from other players' current records (because this data was used to calculate  $a$  and  $b$ ). Some people find this paradoxical that the EoSAverage prediction for a particular player should use data from other players, and find it hard to reconcile that these paradoxical estimates often significantly outperform the naive estimates in terms of accuracy. Provide a brief explanation of this paradox which sometimes goes by the name "Stein's Paradox".