# Data 102, Fall 2023
# Homework 2

Due: **5:00 PM** Friday, February 16, 2024

## 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 LaTeX.
- 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

1. (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 $x_1, \ldots, x_n$.

   (a) (1 point) What is the likelihood function of the observations $x_1, \ldots, x_n$? Express your answer as a function of the 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.

   (b) (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,

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

   **Hint**: You don't need to use calculus.

   (c) (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 $x_1, \ldots, x_n$.

   (d) (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 act 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. Using the starter code in beetledata.py, 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 `pymc` 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.

## Baseball Average Prediction

2. (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 Scientic 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 $= \frac{\text{H}}{\text{AB}}$.

   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 $i^{th}$ 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 EoSAverages (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 EoSAverages.
- 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".

## School District Funding Gaps

3. (15 points) In this question, you'll work with data on school funding provided by the School Finance Indicators Database. The dataset contains information on each school district in the US, including student demographics, district spending per student, test score outcomes, and more. You'll work with the following three columns:

   - `state_name`
   - `fundinggap`, the difference in how much the district should spend on each student and the amount it actually spends per student. Negative values indicate insufficient spending.

   You can find more information about the data at the SFID website.

   For ease of visualization, we'll limit our analysis to the following five states: California, the District of Columbia, Nevada, Oregon, and Texas.

   The file `dcd.csv` contains the dataset we will be using. You should use the provided `q3.ipynb` file to get started, which has cells with some useful variables already defined, and a hint about how to use fancy indexing. This notebook is **not** comprehensive: it only has some starter code and useful functions for this question.

   (a) (1 point) Visualize the funding gap for all districts in the five states above. In two sentences or less, describe any differences you see between the data from larger states (California and Texas) and smaller ones (Nevada and DC).

   (b) (3 points) We'll use a hierarchical model to help us understand state-level averages in the funding gap: each state will have a state-level mean $\mu_i$ with common mean $\alpha$, and for each district $j$ in state $i$, the funding gap $y_{ij}$ will be normally distributed with mean $\mu_i$. We'll assume the variances are known, so the model can be written as:

   $$\mu_i \sim \text{Normal}(\alpha, \sigma_0^2)$$
   $$y_{ij} \sim \text{Normal}(\mu_i, \sigma^2)$$

   Draw a graphical model for this setup.

   (c) (0 points) (Optional) Use empirical Bayes and the data from all 50 states to determine the values of $\alpha$, $\sigma$, and $\sigma_0$ to use. Explain in three sentences or less how and why you chose the data to use when computing this value.

   (d) (4 points) Implement the model from part (b) in PyMC, using $\alpha = \$700$, $\sigma_0 = \sigma = \$4000$. Using the `plot_state_posterior_means` function provided for you in the notebook, visualize the posterior distributions for the means of each of the five states. For which state(s)/district(s) is the posterior mean the most certain? For which state/district is it least certain? Explain why.

   (e) (2 points) Re-run your model from the previous part, changing only one variable at a time as follows:

   (i) $\alpha = \$700, \sigma_0 = \$4000, \sigma = \$\mathbf{400}$

(ii) $\alpha = \$700, \sigma_{\mathbf{0}} = \$\mathbf{400}, \sigma = \$4000$

(iii) $\alpha = -\$\mathbf{700}, \sigma_0 = \$4000, \sigma = \$4000$

What changes in each of the three cases, and why?

*Hint: you can answer this question by focusing on the changes in the mean for the District of Columbia.*

(f) (2 points) Suppose we had treated $\alpha$ as a normal random variable with mean $\gamma$ and standard deviation $\lambda$. Draw a graphical model for this new model.

*Hint: the answer should only require a small change from your answer to part (b).*

(g) (3 points) Implement the model from the previous part in PyMC, using $\gamma = 0$, $\sigma = 4000$, and $\lambda = 10000$. Using your samples, compute the posterior variance of the mean for the District of Columbia (DC), $\text{var}(\mu_{DC}|y)$., and the posterior variance of the mean for California, $\text{var}(\mu_{CA}|y)$.

(h) (0 points) (Optional) Re-run your model from the previous part, changing only one variable at a time as follows:

(i) $\gamma = \$0, \lambda = \$10000, \sigma = \$\mathbf{400}$

(ii) $\gamma = \$0, \lambda = \$\mathbf{100}, \sigma = \$4000$

(iii) $\gamma = -\$\mathbf{0}, \lambda = \$10000, \sigma = \$4000$

What changes in each of the three cases, and why? Explain any differences between your findings here and your findings from part (e)

(i) (0 points) (Optional) The histograms from parts (a) and (d) both have one color per state, but the quantity being visualized in each graph is fundamentally different. Explain this difference.

(j) (0 points) (Optional) Re-run your model from part (g) on all 50 states. How do the results change?