# Math76 (Summer 2020): Introduction to Bayesian Computation
# Final Project

Dartmouth College

Due Monday August 31 at 12:00pm (noon) EST (GMT-4)

This project is designed to give you experience thinking through an entire Bayesian analysis problem: from constructing a statistical model of the data to using the posterior distribution on the model's parameters to compute an expectation and answer questions about the data or model parameters. I've provided two possible topics you might want to consider, but you are also welcome to find another dataset that's interesting to you. Descriptions of each topic and guidance on choosing your own data is provided below. If you do decide to find your own data, I suggest looking through some of the datasets on Kaggle (`https://www.kaggle.com/datasets`).

There are no requirements on what techniques you can and cannot use in this project, as long as they are related to the Bayesian concepts discussed in class. You can do equally well by applying conjugate distributions or using MCMC to explore a more complicated posterior density. Just show me that you can apply concepts we learned in class to your topic.

Unlike the midterm, *I highly encourage you to discuss your project with other students and myself.* Feel free to send me an email, post a discussion topic, or chat with other students over Zoom. You will have to write up your own work, but it is ok if the report includes ideas that stemmed from conversations with other students.

This project will require some level of programming, but the extent of the programming is very much up to you. I encourage you to reuse anything I've posted in class; either as part of the homework, in a demo, or just to create plots in one of the lectures. Remember the goal of this project is for you to apply concepts from class to a new problem so you do not need to reinvent the wheel by rewriting code we have already used or re-deriving identities we've already discussed. If you use identities or code from another source, just make sure to cite where you found it.

## Grading Criteria:

This project is worth a total of 20 points. The categories below will be used for assessing your grade. Note that the majority of the points are devoted to the report itself. Partial point values may be awarded for each section.

- (1 pt) The solution to the project is a typed report of approximately 4-8 pages in length with single-spaced 12pt font.
- (3 pts) Well documented Python code accompanies the final report and can be used to recreate the results.
- (2 pts) The mathematics used in the report is clearly stated in words as well as equations at a level that another Math76 student could reasonably follow.
- (2 pts) Assumptions, such as the particular probability distributions you've chosen or the conditional independence of observations, are clearly explained and justified in the text. Note that the justification for some of your assumptions may simply be that it was necessary to simplify your computations.
- (12 pts) The final report contains all relevant information and is divided into the following sections:
  - **Introduction** This section should define the question you are trying to answer and describe the data you are going to use. Why is the problem important and what features of the data will help you?
  - **Formulation** This section should describe your statistical model for the data (i.e., the likelihood function), what parameters are in the model (i.e., what you are trying to infer with Bayes rule), what is your prior distribution and how does your choice reflect information known a priori about the parameters, and finally, how can you use the posterior density to answer the question you defined in the introduction (e.g., what posterior expectation will you compute?). In essence, this section should define a posterior density (or mass function) $f(x|y)$ and at least one function $h(x)$ such that the posterior expectation $\mathbb{E}_{x|y}[h(x)]$ helps answer the question you define in the introduction.
  - **Solution Strategy** This section should describe the details of how you will compute the posterior density and posterior expectations defined in the formulation section. For example, if you are using a conjugate prior and likelihood, how are the parameters in the posterior distribution related to the parameters in the prior distribution? Depending on your $h(x)$, how will you use these parameters to compute $\mathbb{E}[h(x)]$? If you are using a non-conjugate distribution, how will you generate posterior samples? Also, what software did you use or develop?
  - **Results** This is where you describe what you found after solving your inference problem and computing the posterior expectations. You should include any relevant plots of your posterior density (e.g., plot posterior samples), report Monte Carlo standard errors if appropriate (applying the CLT to both MCMC and Monte Carlo results), and list values of your posterior expectations.
  - **Discussion** This section should provide analysis of your results. What is the answer to your question? How valid do you think your conclusions are? Do you think they would change if you relaxed any of your assumptions? Does the posterior distribution itself (e.g., means, variances, probabilities, etc..) give us any additional insight into the data or problem?

## Suggested Workflow

(a) Look through the data. Adapt existing code to read the data into python (e.g., the `read_csv` function that we've used in pandas).

(b) Choose a question you might want to ask about the data. For example, is American happiness increasing or decreasing?

(c) Develop a statistical "model" of the data that would allow you to answer your question. This will form the likelihood in Bayes' rule. Think back to the examples in class, is your question similar to one we've studied in class? Can your question be related to the coefficients in a regression problem? or the parameters of a Poisson distribution?

(d) Think about what you might know about the model parameters. Do they have to be positive? What would be a reasonable value? Is there any historical data that might tell you about their values? Use the answers to these questions to develop a prior distribution.

(e) Characterize the posterior distribution. Do we have a prior that is conjugate to the likelihood? Do we need to use MCMC to sample the posterior? Can you borrow code from a previous homework or demo?

(f) Using the posterior, compute an expectation that helps answer your question. For example, what is the probability that the slope of a regression line is less than 1?

## Possible Topic 1: World Happiness

The United Nations has released several ”World Happiness Reports” that attempt to quantify levels of happiness around the world. Some of the data for these reports is provided at `https://www.kaggle.com/unsdsn/world-happiness`. The goal of this topic is to use Bayesian techniques to analyze these data and assess trends in world happiness. Questions you might want to answer in your analysis include: How is happiness changing over time? or Are countries with higher GDPs happier? Note that the data is available from Kaggle as CSV files. We have used pandas in many different examples for reading CSV files, so you should adapt previous code to read the data into python and perform your analysis.

# Possible Topic 2: Sea Ice Thickness

Measuring the thickness of sea ice in both the Arctic and Antarctic is important for understanding climate change. The IceSat-2 satellite[1], which was launched was launched in 2018, uses a laser altimeter to measure the surface height of the ice to an accuracy of a few centimeters, but does not measure the ice thickness. The height of the ice above the water, which can be extracted from the IceSat-2 measurements, is called the ice ”freeboard” and is a function of the thickness. The problem is that these measurements include an unknown snow thickness. In particular, the relationship between ice thickness $H$, ice freeboard $F$, and snow depth $S$ is given by

$$F = \frac{\rho_w - \rho_i}{\rho_w} \left( H - S \frac{\rho_s}{\rho_w - \rho_i} \right),$$

where $\rho_w$, $\rho_i$, and $\rho_s$ are the densities of the water, sea ice, and snow, respectively. Note that the density of sea water $\rho_w$ in the Arctic ocean range from 1010 to 1030 $kg/m^3$ depending on location [2]. The density of sea ice $\rho_i$ is typically around 910 $kg/m^3$ but measurements can range from around 750 to 940 $kg/m^3$ [3]. The density of snow can also vary from 100 $kg/m^3$ to 400 $kg/m^3$ depending on time of year and location but has an average density of $300kg/m^3$ [4].

The goal of this topic is to estimate the ice thickness given a freeboard measurement of $0.1m$ from IceSat-2. Note there are many ways to do this, ranging from simple to complex. There is no wrong answer as long as you clearly state any simplifications or assumptions you make and discuss the possible impacts of these simplifications on your results.

[1] See `https://icesat-2.gsfc.nasa.gov/mission` for more details on the IceSat-2 mission

[2] See `https://svs.gsfc.nasa.gov/3652`

[3] Timco and Frederking (1996), “A review of sea ice density,” Cold Regions Science and Technology.

[4] Warren et al., (1999) “Snow Depth on Arctic Sea Ice”, Journal of Climate.

## Possible Topic 3: Your Own Problem

If the previous two topics are not interesting to you, or you simply have a burning desire to study a different problem, you are more than welcome to do so as long as you use techniques from this course in your analysis. Your problem may involve a physical model, like the sea ice thickness example, or it may rooted entirely in observational data, like the World Happiness example. In either case, make sure your problem will allow you to follow the suggested workflow above in order to satisfy all of the grading criteria. If you have a problem in mind but aren't sure if it's a good fit for this project, please discuss it with me as soon as possible.