# CS146 Assignment 3

## Using Stan & other exercises

You should submit your work as a Python notebook, or a Python notebook and PDF both if you want to separate your code and your report. (Do not submit your Python code as a PDF only. This usually results in lines being truncated.) Typeset your PDF using Google Docs, LaTeX, Jupyter notebooks, CoCalc, or any other software that allows you to type text and math. Make sure your code is readable and commented.

**The #InterpretingProbabilities LO** is there for the stretch goal only. You are not expected to interpret results for the main part of the assignment – only to check that they match the results we got in class.

**If you have trouble installing Stan**, you might be better off using it on Google Colab or on CoCalc. Both platforms come with Stan pre-installed.

## 1. Implement models in Stan (required)

Implement each of the models below using Stan and produce the results or plots requested for each model. You have seen each of these models before in class.

The goal of this exercise is to learn how to implement different types of parameters, likelihood functions, and prior distributions using Stan. Stan always generates samples for estimating posterior distributions, while we used conjugate distributions in class. Check that your results from Stan's samples match the results we computed in class.

1. **Call center data set – exponential likelihood with a gamma prior.** Estimate the number of calls per minute for the 13th hour of the call center data set.

   Results to compute:

   - Posterior 98% confidence interval over $\lambda$ (check that it matches results in the solution notebook below)
   - Histogram of posterior $\lambda$ samples

   Resources for you to use:

   - Data set: call_center.csv
   - Solution for class activity (call_center_solution.ipynb)

2. **Normal likelihood with normal-inverse-gamma prior.**

   Results to compute:

   - 95% posterior confidence intervals for the mean $\mu$ and variance $\sigma^2$ of the data.
   - Take 10 samples from your posterior over $\mu$ and $\sigma^2$ and plot the normal distributions corresponding to them. See Task 3 in the solutions below – you should produce a plot similar to the one you find there.

   Resources for you to use:

   - Data and solution for class activity (normal_inverse_gamma_solution.ipynb)

3. **Log-normal HRTEM data.** Normal likelihood log-transformed data and using a normal-inverse-gamma prior.

   Results to compute:

   - 95% posterior confidence intervals for the $\mu$ and variance $\sigma^2$ of the log-transformed data. (Should match results under Task 3 of the solutions.)
   - Take 10 samples from your posterior over $\mu$ and $\sigma^2$ and plot the log-normal distributions corresponding to them. See Task 5 in the solutions below – you should produce a plot similar to the one you find there, but with 10 pdfs rather than one.

   Resources for you to use:

   - Data set: hrtem.csv (remember to log-transform the data)
   - Solution for class activity (hrtem_solution.ipynb)

## 2. Stretch goal (optional)

Fit a mixture of two Gaussians to the HRTEM data set using Stan. Your likelihood function for each datum should be

$$\begin{aligned} P(datum \mid parameters) &= P(\log x_i \mid p, \mu_1, \mu_2, \sigma_1, \sigma_2) \\ &= p\, N(\log x_i \mid \mu_1, \sigma_1^2) + (1-p)\, N(\log x_i \mid \mu_2, \sigma_2^2) \end{aligned}$$

That is, with probability $p$ the log of the datum is generated from the first Gaussian, and with probability $(1-p)$ it is generated from the second Gaussian.

Produce plots and summaries of your posteriors over all parameters. Use samples from the posterior to show whether and how well the likelihood functions corresponding to the samples match the data histogram.

**Note:** You have to be really careful when fitting a mixture model like this one since the posterior is always bimodal. For example if

$$\theta' = (p',\ \mu_1',\ \mu_2',\ \sigma_1',\ \sigma_2')$$

is a mode of the posterior, then so is

$$\theta'' = (1-p',\ \mu_2',\ \mu_1',\ \sigma_2',\ \sigma_1')$$

since the likelihood is identical with these two settings of the parameters. This can really mess up sampling-based tools like Stan.

To confirm that sampling from your model really is working, make sure that the Stan diagnostic outputs are reasonable – `n_eff` should be large (at least a few hundred; ideally 1000 or more) and `Rhat` should be close to 1.

If your sampling does not want to converge, review this video on Stan models and make sure you use appropriate data types and priors for your model parameters.

# More practice exercises (optional)

Below are additional practice exercises for you to attempt. These are optional and you can choose to do as many or as few as you want. These exercises will not be graded.

1. Comparison of two multinomial observations. On the evening of a presidential campaign debate in 1988, ABC News surveyed voters in the United States before and after the debate, to measure whether the debate had any impact on voter preferences. The survey results are given below.

| Survey | Bush | Dukakis | Other |
|---|---|---|---|
| Before debate | 294 | 307 | 38 |
| After debate | 288 | 332 | 19 |

   a. Assume that the before and after surveys are independent samples from the population of voters, and model the survey results as samples from two different multinomial distributions.
   b. Let $\alpha_0$ be the proportion of voters who preferred Bush to Dukakis before the debate, and $\alpha_1$ be the same proportion but after the debate. Plot a histogram over $\alpha_1 - \alpha_0$.
   c. What is the probability that voter preference shifted towards Dukakis as a result of the debate?

2. Derivation of conjugate distribution update equations. Derive the posterior distribution over the parameter $p$ of the binomial likelihood function if it has a conjugate beta prior distribution over $p$.

3. Derivation of the mode of a pdf. Calculate the derivatives of the normal-inverse-gamma pdf with respect to its random variables, $x$ and $\sigma^2$. Set all derivatives equal to 0 and solve the resulting equations to determine the values of $x$ and $\sigma^2$ that maximize the pdf in terms of the parameters $\mu, \lambda, \alpha, \beta$.