

Homework 4

PSTAT 115, Spring 2023

Due on June 9, 2023 at 11:59 pm

Installing cmdstanr

```
install.packages("cmdstanr",  
  repos = c("https://mc-stan.org/r-packages/",  
    getOption("repos")))  
  
library(cmdstanr)  
  
install_cmdstan()
```

Problem 1. Goal Scoring in the Women's World Cup

The Chinese Women's soccer team recently won the AFC Women's Asian Cup. Suppose you are interested in studying the Asian Cup performance of this soccer team. Let λ be the average number of goals scored by the team. We will analyze λ using the Gamma-Poisson model where data Y_i is the observed number of goals scored in the i th Asian Cup game, i.e. we have $Y_i | \lambda \sim \text{Pois}(\lambda)$. The Chinese Women's team scored 4, 7, 3, 2, 3 goals in each one of the matches. We ignore the match against India, which didn't take place due to a Covid outbreak on the Indian team.

A priori, we expect the rate of goal scoring to be $\lambda \sim \text{Gamma}(a, b)$. According to a sports analyst, they believe that λ follows a Gamma distribution with $a = 1$ and $b = 0.25$.

1a. (5pts) Compute the theoretical posterior parameters a , b , and also the posterior mean.

1b. (10pts) Create a new Stan file by selecting "File -> New File -> Stan file" in RStudio and name it `women_cup.stan`, use Rstan to report and estimate the posterior mean of the scoring rate by computing the sample average of all Monte Carlo samples of λ .

1c. (5pts) Create a histogram of the Monte Carlo samples of λ and add a line showing the theoretical posterior of density of λ . Do the Monte Carlo samples coincide with the theoretical density?

1d. (10pts) Use the Monte Carlo samples from Stan to compute the mean of the predictive posterior distribution to estimate the distribution of expected goals scored for next game played by the Chinese women's soccer team.

Problem 2. Bayesian inference for the normal distribution in Stan.

Create a new Stan file and name it `IQ_model.stan`. We will make some basic modifications to the template example in the default Stan file for this problem. Consider the IQ example in the class slides. Scoring on IQ tests is designed to yield a $N(100, 15)$ distribution for the general population. We observe IQ scores for a sample of n individuals from a particular town, $y_1, \dots, y_n \sim N(\mu, \sigma^2)$. Our goal is to estimate the population mean IQ in the town. Assume the $p(\mu, \sigma) = p(\mu | \sigma)p(\sigma)$, where $p(\mu | \sigma)$ is $N(\mu_0, \sigma/\sqrt{\kappa_0})$ and $p(\sigma)$ is $\text{Gamma}(a, b)$. Before you administer the IQ test you believe the town is no different than the rest of the population, so you assume a prior mean for μ of $\mu_0 = 100$, but you aren't too sure about this a priori and so you set $\kappa_0 = 1$ (the effective number of pseudo-observations). Similarly, a priori you assume σ has a mean

of 15 (to match the intended standard deviation of the IQ test) and so you decide on setting $a = 15$ and $b = 1$ (remember, the mean of a Gamma is a/b). Assume the following IQ scores are observed:

```
y <- c(70, 85, 111, 111, 115, 120, 123)
n <- length(y)
```

2a. (10pts) Make a scatter plot of the posterior distribution of the mean, μ , and the precision, $1/\sigma^2$. Put μ on the x-axis and $1/\sigma^2$ on the y-axis. What is the posterior relationship between μ and $1/\sigma^2$? Why does this make sense? *Hint:* review the lecture notes.

2b. (5pts) You are interested in whether the mean IQ in the town is greater than the mean IQ in the overall population. Use Stan to find the posterior probability that μ is greater than 100.

2c. (15pts) You notice that two of the seven scores are significantly lower than the other five. You think that the normal distribution may not be the most appropriate model, in particular because you believe some people in this town are likely have extreme low and extreme high scores. One solution to this is to use a model that is more robust to these kinds of outliers. The [Student's t distribution](#) and the [Laplace distribution](#) are two so called “heavy-tailed distribution” which have higher probabilities of outliers (i.e. observations further from the mean). Heavy-tailed distributions are useful in modeling because they are more robust to outliers. Fit the model assuming now that the IQ scores in the town have a Laplace distribution, that is $y_1, \dots, y_n \sim \text{Laplace}(\mu, \sigma)$. Create a copy of the previous stan file, and name it `IQ_laplace_model_1.stan`. *Hint:* In the Stan file you can replace `normal` with `double_exponential` in the model section, another name for the Laplace distribution. Like the normal distribution it has two arguments, μ and σ . Keep the same prior distribution, $p(\mu, \sigma)$ as used in the normal model. Under the Laplace model, what is the posterior probability that the median IQ in the town is greater than 100? How does this compare to the probability under the normal model? Why does this make sense?

Problem 3. Cows and milk

Farmer John has a huge number of cows. Earlier this year he ran an experiment where he gave 10 cows a special diet that he had heard could make them produce more milk. He recorded the number of liters of milk from these “diet” cows and from 15 “normal” cows during one month. Here is the data:

```
diet_milk = c(651, 679, 374, 601, 401, 609, 767, 709, 704, 679)
normal_milk = c(798, 1139, 529, 609, 553, 743, 151, 544, 488, 555, 257, 692, 678, 675, 538)
```

We want to answer the question: Was the diet any good? Did it result in higher milk production?

Assume the milk production of the cows on the diet is given by a random variable Y_1 and the production of the normal cows is denoted by Y_2 . Assume the statistical model is such that

$$\begin{aligned} Y_{1,i} \mid \mu_1, \sigma_1 &\sim N(\mu_1, \sigma_1^2), \text{ for } i = 1, \dots, 10 \\ Y_{2,j} \mid \mu_2, \sigma_2 &\sim N(\mu_2, \sigma_2^2), \text{ for } j = 1, \dots, 15 \\ \sigma_1 &\sim \text{Unif}[0, 1000] \\ \sigma_2 &\sim \text{Unif}[0, 1000] \\ \mu_1 \mid \sigma_1 &\sim \text{Unif}[0, 2000] \\ \mu_2 \mid \sigma_2 &\sim \text{Unif}[0, 2000] \end{aligned}$$

3a. (10pts) Write the Stan model into object `cows_string` below.

```
cows_string = "
  INCLUDE YOUR STAN CODE HERE.

  YOU CAN BREAK LINES.

  YOU CAN ALSO IDENT YOUR TEXT.
"
```

3b. (5pts) Generate samples from the unknown parameters in the model and save them on an object named `stan_samples`

```
# stan_samples = ....
```

3c. (5pts) Plot the 90% credible interval based on the quantiles of the posterior distribution for each one of the four parameters in the model. Don't plot any other interval (the `outer line`).

3d. (10pts) Store the posterior samples of the difference $\mu_2 - \mu_1$ on an object named `mu_diff`, plot its histogram and answer the main question: is it likely that the diet is going to make the cows produce more milk on average?