# HW 01: Inference using Bayesian statistics

## United States Births in 2014

> **!** Due date
>
> This assignment is due on **Thursday, January 29 at 11:45am.** To be considered on time, the following must be done by the due date:
>
> - Final `.qmd` and `.pdf` files pushed to your GitHub repo
> - Final `.pdf` file submitted on Gradescope

## Getting started

- Go to the biostat725-sp26 organization on GitHub. Click on the repo with the prefix **hw-01**. It contains the starter documents you need to complete the homework.
- Clone the repo and start a new project in RStudio. See the AE 01 instructions for details on cloning a repo and starting a new project in R.

## Packages

The following packages are used in this assignment:

```
library(dplyr)
library(ggplot2)
library(openintro)
library(knitr)
library(mvtnorm)

# load other packages as needed
```

## Introduction

Every year, the United States releases to the public a large dataset containing information on births recorded in the country. This dataset has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. A random sample of 1,000 cases from the dataset released in 2014 can be found in the `births14` data set in the **openintro** R package.

We will focus on the following variables:

- `visits`: Number of hospital visits during pregnancy
- `weight`: Weight of the baby at birth in pounds
- `habit`: Status of the mother as a `nonsmoker` or a `smoker`

For this homework, work with the complete case dataset.

```
births14 <- births14[complete.cases(births14), ]
glimpse(births14)
```

```
Rows: 794
Columns: 13
$ fage           <int> 34, 36, 37, 32, 32, 37, 29, 30, 29, 30, 34, 28, 32, 24,~
$ mage           <dbl> 34, 31, 36, 31, 26, 36, 24, 32, 26, 34, 27, 22, 25, 20,~
$ mature         <chr> "younger mom", "younger mom", "mature mom", "younger mo~
$ weeks          <dbl> 37, 41, 37, 36, 39, 36, 40, 39, 39, 42, 40, 40, 34, 37,~
$ premie         <chr> "full term", "full term", "full term", "premie", "full ~
$ visits         <dbl> 14, 12, 10, 12, 14, 10, 13, 15, 11, 14, 16, 20, 20, 10,~
$ gained         <dbl> 28, 41, 28, 48, 45, 20, 65, 25, 22, 40, 30, 31, 25, 70,~
$ weight         <dbl> 6.96, 8.86, 7.51, 6.75, 6.69, 6.13, 6.74, 8.94, 9.12, 8~
$ lowbirthweight <chr> "not low", "not low", "not low", "not low", "not low", ~
$ sex            <chr> "male", "female", "female", "female", "female", "female~
$ habit          <chr> "nonsmoker", "nonsmoker", "nonsmoker", "nonsmoker", "no~
$ marital        <chr> "married", "married", "married", "married", "married", ~
$ whitemom       <chr> "white", "white", "not white", "white", "white", "white~
```

### Exercises 1-7

Define a random variable $Y_i$ that represents the number of hospital visits during pregnancy for each woman $i$, for $i = 1, \ldots, n$. Assume that this random variable follows a Poisson distribution with rate $\lambda$, such that $Y_i \overset{iid}{\sim} \text{Poisson}(\lambda)$. For a Poisson distribution, the mean and variance are equal to $\lambda$. We are interested in performing statistical inference on $\lambda$ using a Bayesian

approach. A frequently used prior for $\lambda$ is Gamma(shape $= a$, rate $= b$), where $\mathbb{E}[\lambda] = a/b$ and $\mathbb{V}(\lambda) = a/b^2$.

### Exercise 1

The researchers have prior knowledge that leads them to believe that $\lambda$ should have mean 8 and variance 4. What values of $a$ and $b$ should they specify?

### Exercise 2

Using the prior specified in **Exercise 1**, compute the probability that $\lambda$ is greater than 11? For this computation compute the exact probability using the `pgamma` function in R. This is equivalent to computing $P(\lambda > 11)$.

### Exercise 3

Compute the same probability as in **Exercise 2**, this time using Monte Carlo sampling. Report your Monte Carlo standard error and make sure it is less than 0.01.

### Exercise 4

Suppose the researchers are interested in the quantity, $\alpha = \sqrt{\lambda}$. Compute the probability that $\alpha$ is greater than 2.5. Use the same number of Monte Carlo samples as in **Exercise 3** and describe why Monte Carlo sampling makes this computation much more efficient than computing the exact probability.

### Exercise 5

Using the prior specified in **Exercise 1**, compute the posterior distribution for $\lambda$, $f(\lambda|\mathbf{Y})$, where $\mathbf{Y}_i = (Y_1, \ldots, Y_n)$. Recall that the Gamma prior for $\lambda$ is a conjugate prior, so that the posterior is given by: $f(\lambda|\mathbf{Y}) \sim Gamma(a+\sum_{i=1}^{n} Y_i, b+n)$. Visualize the posterior distribution and report the posterior mean and a 95% credible interval. Provide an interpretation of the posterior summaries within the context of the US births data.

### Exercise 6

What is the posterior probability that $\lambda$ is greater than 11? This is equivalent to computing $P(\lambda > 11|\mathbf{Y})$. Again, use Monte Carlo sampling. Provide an interpretation for this probability in the context of hospital visits.

### Exercise 7

Create a figure that includes both the prior and posterior distributions for $\lambda$. Also, include a figure of the observed data. Use these figures to make a comparison of the prior and posterior probabilities found in **Exercise 3** and **Exercise 6**, respectively. Describe any changes in these two probabilities and how they relate to the observed data.

## Exercise 8-10

Define a random variable $weight_i$ that represents the weight of the baby at birth in ounces for pregnancy $i$. We are interested in learning the association between birth weight and the smoking habit, $habit_i$, of the mother. Fit the following Bayesian linear regression model using Gibbs sampling,

$$\begin{aligned} weight_i &= \beta_0 + \beta_1 \times 1(habit_i = \text{smoker}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \\ &= \mathbf{x}_i\beta + \epsilon_i, \\ \beta &\sim N(\mathbf{0}, 100\mathbf{I}) \\ \sigma^2 &\sim \text{Inv-Gamma}(3, 1). \end{aligned}$$

### Exercise 8

Obtain samples from the posterior distribution of $(\beta, \sigma^2)$ given the observed data. Visualize the posterior distributions and provide justification that the Gibbs sampler has converged.

### Exercise 9

Report the posterior mean, standard deviation, and 95% credible intervals for each parameter.

### Exercise 10

If someone were to fit the same regression using a frequentist approach the resulting model would look like the following.

```
mod <- lm(weight ~ habit, data = births14)
res <- summary(mod)
print(res)
```

```
Call:
lm(formula = weight ~ habit, data = births14)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4965 -0.6865  0.0635  0.8150  3.1135

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.30654    0.04586 159.317  < 2e-16 ***
habitsmoker -0.75203    0.16412  -4.582 5.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.241 on 792 degrees of freedom
Multiple R-squared:  0.02583,   Adjusted R-squared:  0.0246
F-statistic:    21 on 1 and 792 DF,  p-value: 5.345e-06
```

Suppose researchers are interested in testing the following hypothesis test: $H_0 : \beta_1 = 0, H_1 : \beta_1 < 0$. We can compute this p-value from the frequentist model.

```
pvalue <- pt(coef(res)[, 3], mod$df, lower = TRUE)[2]
```

The resulting p-value is <0.001. Compute the Bayesian p-value that corresponds to the same hypothesis test, $P(\beta_1 < 0|\mathbf{Y})$. Interpret both p-values at a Type-I error rate of 0.05 and compare and contrast their interpretations in the context of the association between smoking and low birth weight.

You're done and ready to submit your work! render, commit, and push all remaining changes. You can use the commit message "Done with Homework 1!", and make sure you have pushed all the files to GitHub (your Git pane in RStudio should be empty) and that all documents are updated in your repo on GitHub. The PDF document you submit to Gradescope should be identical to the one in your GitHub repo.

## Submission

You will submit the PDF documents for homeworks, and exams in to Gradescope as part of your final submission.

> **Warning**
>
> Before you wrap up the assignment, make sure all documents are updated on your GitHub repo. We will be checking these to make sure you have been practicing how to commit and push changes.
> Remember – you must turn in a PDF file to the Gradescope page before the submission deadline for full credit.

To submit your assignment:

- Access Gradescope through the menu on the BIOSTAT 725 Canvas site.
- Click on the assignment, and you'll be prompted to submit it.
- Mark the pages associated with each exercise. All of the pages of your homework should be associated with at least one question (i.e., should be "checked").
- Select the first page of your .PDF submission to be associated with the *"Workflow & formatting"* section.

## Grading

| Component | Points |
|---|---|
| Ex 1 | 3 |
| Ex 2 | 3 |
| Ex 3 | 3 |
| Ex 4 | 5 |
| Ex 5 | 7 |
| Ex 6 | 4 |
| Ex 7 | 7 |
| Ex 8 | 8 |
| Ex 9 | 3 |
| Ex 10 | 3 |
| Workflow & formatting | 4 |
| Total | 50 |

The "Workflow & formatting" grade is to assess the reproducible workflow and document format. This includes having at least 3 informative commit messages, a neatly organized document with readable code and your name and the date updated in the YAML.