

# Computational Statistics Assignment 2

by James D. Wilson (University of San Francisco)

1. Suppose that we model observed data  $\mathbf{y}$  with the joint probability density function  $f(\mathbf{y} \mid \theta)$ .
  - (a) Describe the assumptions about  $\theta$  from both the *frequentist* and *Bayesian* standpoint.
  - (b) Suppose that we would like to determine whether or not  $\theta = 0$ . Describe how to go about this from both a *frequentist* and *Bayesian* standpoint.
  - (c) When do *frequentist* and *Bayesian* inference provide the same conclusions about  $\theta$  from your inferential tasks described in (b)?
  - (d) What is the difference between a 95% confidence interval and a 95% credible interval for  $\theta$ ?
2. State which methods you would use to perform the following computations. If more than 1 method is needed, provide pseudo-code to describe how to go about the computation. Further, if there are multiple options, please write all options that you can think of down.
  - (a) Estimate  $\mathbb{E}[\log(|\theta|)|y]$  when

$$p(\theta \mid y) = \frac{1}{\pi y} \left( 1 + \left( \frac{\theta - 1}{y} \right)^2 \right)$$

- (b) Estimate  $\mathbb{E}[\log(|\theta|)|y]$  when  $p(\theta \mid y)$  is unknown, but we know that  $p(\theta \mid y) \propto q(\theta \mid y)$
  - (c) Simulate from an unknown  $p(\theta \mid y)$
  - (d) Simulate from a known but unrecognizable  $p(\theta \mid y)$
3. **Creating animations with MCMC** Read the blog post about creating animations with MCMC here: <https://jankrepl.github.io/creating-animations-with-MCMC/>. Using the source code from the author, create GIF animations for two different images of your choice (example - the USF logo). For each image, run rejection sampling, Gibbs, and Metropolis Hastings on each image with the following parameters
  - (a) Run 1, 2, 3, and 4 chains for each with different colors for each chain and run each with 10000 samples.
  - (b) Repeat (a) but with 100, 500, 1000, and 3000 samples.

Now answer the following questions:

- (a) Based on the animation, what is the importance of running multiple chains? What happens if we only use say 1 chain?
- (b) Based on your sampling procedures and sample sizes, which method seems to converge to the distribution of the image the fastest? Which is the slowest?
- (c) Based on your sample size analysis, how many samples and chains would you suggest someone uses for each sampling procedure and each image?

4. In 2015, ABC News conducted a survey of 600 people before a presidential debate between Trump and Clinton, and an independent group of 600 people were polled after the presidential debate. The results were recorded as follows:

Survey	Trump	Clinton	Other
pre-debate	215	310	75
post-debate	280	290	30

Let  $\alpha_1$  be the proportion of people that supported Trump before the presidential debate. And let  $\alpha_2$  be the proportion of people that supported him after the debate. Construct a Bayesian model from which you analyze the value  $\alpha_2 - \alpha_1$ . State your data generating distribution, and your prior and conduct MCMC to sample from the posterior of  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_2 - \alpha_1$ . What is the probability that there was a shift away from Clinton after the debate?