Homework 7
Statistics 200B
Due March 18

1. Consider a Bayesian model in which, conditional on unknown parameter λ, $X_1, \ldots, X_n$ are iid with exponential PDF

$$f(x|\lambda) = \frac{1}{\lambda} e^{-x/\lambda}$$

for $x > 0$, and the prior distribution is $InverseGamma(a, b)$, with PDF

$$f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{-a-1} e^{-b/\lambda}$$

for $\lambda > 0$.

(a) Find the posterior distribution for λ, conditioning on $X_1, \ldots, X_n$. It is fine to write the family and specify its parameters; you do not need to write out the CDF or PDF.

(b) Show that the posterior mean can be written as a weighted average of the prior mean and the MLE for λ. You may use the fact that the mean of an Inverse Gamma distribution with parameters a and b is $\frac{b}{a-1}$. What happens as $n \to \infty$?

Solutions:

(a) By Bayes rule,

$$\begin{aligned} f(\lambda|X^n) &\propto f(X^n|\lambda) f(\lambda) \\ &\propto \lambda^{-n} e^{-\sum X_i/\lambda} \lambda^{-a-1} e^{-b/\lambda} \\ &\propto \lambda^{-(a+n)-1} e^{-(b+\sum X_i)/\lambda}, \end{aligned}$$

which we recognize as the kernel of an $InverseGamma(a^*, b^*)$ distribution for λ, where $a^* = a+n$ and $b^* = b+\sum_{i=1}^n X_i$. Since $f(\lambda|X^n)$ must integrate to one, we have shown that $InverseGamma(a^*, b^*)$ is the posterior.

(b) The MLE for λ is $\bar{X}_n$, which you may derive or state based on earlier problems. The prior mean is $\frac{b}{a-1}$, and the posterior mean is

$$\begin{aligned}\frac{b^*}{a^*-1} = \frac{b+\sum X_i}{a+n-1} &= \frac{b}{a+n-1}\frac{a-1}{b}\frac{b}{a-1} + \frac{\sum X_i}{a+n-1}\frac{1}{\bar{X}_n}\bar{X}_n \\ &= \frac{a-1}{a-1+n}\frac{b}{a-1} + \frac{n}{a-1+n}\bar{X}_n\end{aligned}$$

and this is a weighted average of the prior mean and the MLE. As $n \to \infty$, $\frac{a-1}{a-1+n} \to 0$ and $\frac{n}{a-1+n} \to 1$, so the posterior mean approaches $\bar{X}_n$.

2. Here is an example for which we can use the model in problem 1. Let λ represent the average time (in units of days) between earthquakes in the Berkeley area. To make this more precise, let's consider only earthquakes with magnitude 3 or greater on the Richter scale, and whose epicenter is within a 10 mile radius of downtown Berkeley, whose coordinates I have as $37°52'18''N$ and $122°16'22''W$.

   (a) Consider using an Inverse Gamma prior for λ. We need to choose the parameters a and b. You may have some prior knowledge about λ, but it may be difficult to translate this into a choice of a and b. To facilitate this, write expressions for a and b in terms of the prior mean m and the prior variance v, using that $m = \frac{b}{a-1}$ and $v = \frac{b^2}{(a-1)^2(a-2)}$ when $a > 2$.

   (b) Based on your current knowledge, choose parameters a and b, and make a plot of the prior PDF. You may find it useful here and in the rest of the problem to modify the `R` code in the file `BetaBinomial.R`, which is on bSpace under Resources>Lecture Notes. (There is a `dinvgamma` function in the `R` package `MCMCpack`, which you can install and load using `install.packages("MCMCpack")` and then `library(MCMCpack)`, or you can just code the mathematical form of the prior PDF directly.) Turn in a sentence of explanation with your plot regarding how your prior knowledge (or lack of it) informed your choice of prior distribution.

   (c) The file `BerkeleyEarthquakes.RData` on bSpace contains a data frame called `earthquakes` with information about earthquakes within a 10 mile radius of Berkeley, from 1969-2008. Load it in and take a look at the first few lines, then extract the waiting time between each earthquake using

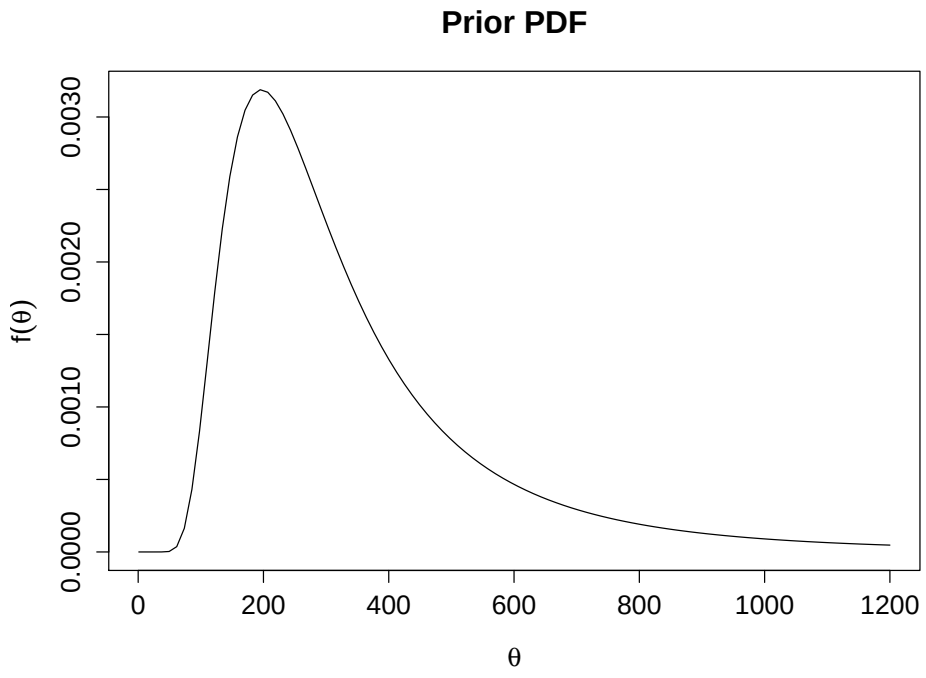
```
load("BerkeleyEarthquakes.RData")
head(earthquakes)
X <- earthquakes$Lag[-1] # First element is NA
```

Using the results you found in problem 1, calculate the posterior distribution for λ, conditional on the observed waiting times. Make a plot comparing your posterior PDF to your prior PDF. Turn in a sentence of explanation with your plot regarding any changes in your knowledge about λ after seeing the data.

Solutions:

(a) $m = \frac{b}{a-1}$ and $v = \frac{b^2}{(a-1)^2(a-2)}$, so $v = \frac{m^2}{a-2} \Rightarrow a = \frac{m^2}{v} + 2$. Therefore $b = m(a-1) = m(\frac{m^2}{v} + 2)$. Note that since $m, v > 0$, $a > 2$, so by construction we have also chosen a and b so that the distribution has finite variance.

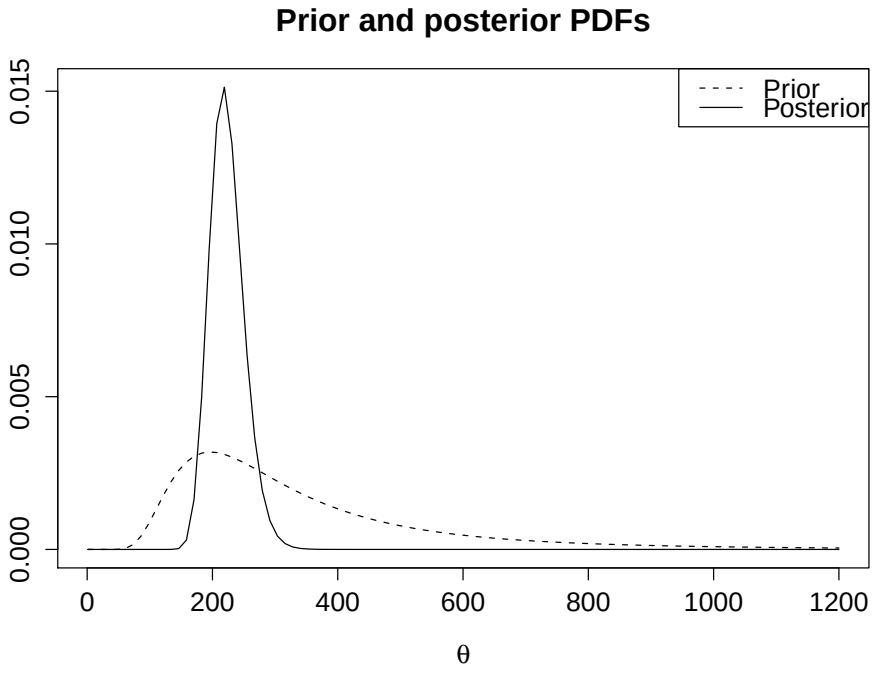
(b) I chose $m = 365$ and $v = 100{,}000$ as my prior mean and variance. This is because I would guess 3+ magnitude earthquakes occur near Berkeley about once per year, but I have quite a bit of uncertainty about that quantity. For example, I'm not sure what a 3+ magnitude earthquake feels like, or how close the epicenter has to be to be able to feel it. Your prior may differ based on your experience. The corresponding prior parameters were $a = 3.33$ and $b = 851.27$.



(c) From Problem 1 we have that the posterior is also Inverse Gamma, with parameters

$$\begin{aligned} a^* &= a + n = 3.33 + 66 = 69.33 \\ b^* &= b + \sum_{i=1}^{n} X_i = 851.27 + 14424.79 = 15276.06 \end{aligned}$$

Again, your posterior parameters may differ from mine. The plot of the prior and posterior PDFs below shows that my beliefs about the mean waiting time between 3+ earthquakes near Berkeley (parameter λ) after seeing the data is much more concentrated. The posterior mean for the average waiting time is smaller than my prior mean ($\frac{b^*}{a^*-1} = 223.55 < 365 = \frac{b}{a-1}$), so earthquakes are occurring more frequently than I thought.



3. Suppose $X_1, \ldots, X_n|\theta \overset{iid}{\sim} Unif(0, \theta)$. Show that the Pareto distribution is the conjugate prior distribution for θ. The PDF for a random variable $Y \sim Pareto(\psi, \alpha)$ is

$$f(y; \psi, \alpha) = \frac{\alpha \psi^\alpha}{y^{\alpha+1}} I\{y > \psi\}.$$

Solution:

The prior PDF for θ is

$$f(\theta) = \frac{\alpha\psi^\alpha}{\theta^{\alpha+1}} I\{\theta > \psi\},$$

and the likelihood is

$$\begin{aligned} f(X^n|\theta) &= \prod_{i=1}^{n} f(X_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} I\{X_i < \theta\} \\ &= \frac{1}{\theta^n} I\{max\{X_1, \ldots, X_n\} < \theta\}. \end{aligned}$$

Therefore, Bayes rule gives that

$$\begin{aligned} f(\theta|X^n) &\propto f(X^n|\theta) f(\theta) \\ &\propto \frac{1}{\theta^n} I\{max\{X_1, \ldots, X_n\} < \theta\} \frac{1}{\theta^{\alpha+1}} I\{\theta > \psi\} \\ &= \frac{1}{\theta^{(\alpha+n)+1}} I\{\theta > max\{X_1, \ldots, X_n, \psi\}\} \end{aligned}$$

which we recognize as the kernel of another Pareto distribution, with parameters $\alpha^* = \alpha + n$ and $\psi^* = \max\{X_1, \ldots, X_n, \psi\}$. Since $f(x^n|\theta)$ must integrate to one, this is the posterior distribution. Since it is also a member of the Pareto family, we have shown that the Pareto distribution is conjugate for θ when the data have $Unif(0, \theta)$ distribution.

4. Consider rejection sampling when the target density $h(\theta) = f(\theta|x^n)$. In class we considered taking the proposal density $g(\theta) = f(\theta)$, i.e., the prior PDF. We set $M = \mathcal{L}(\hat{\theta}_n)$, where $\hat{\theta}_n$ is the MLE for θ. Explain why we should not take M to be any larger than this. Explain why we should not take it to be any smaller.

Solution:

As in class, define $k(\theta) = \mathcal{L}_n(\theta) f(\theta)$, and note that $h(\theta) = f(\theta|x^n) = k(\theta)/\int k(\theta) d\theta$. we need to choose M such that

$$\frac{k(\theta)}{Mg(\theta)} = \frac{\mathcal{L}_n(\theta) f(\theta)}{Mf(\theta)} \leq 1$$

for all θ. Note that $M = \mathcal{L}(\hat{\theta}_n)$ achieves this, where $\hat{\theta}_n$ is the MLE. This is because

$$\sup_{\theta} \mathcal{L}_n(\theta) = \mathcal{L}_n(\hat{\theta}_n)$$

by definition of the MLE. If we chose instead $M > \mathcal{L}_n(\hat{\theta}_n)$, our acceptance probabilities would be lower, since a given θ^{cand} is accepted with probability

$$\frac{\mathcal{L}_n(\theta^{cand})}{M} < \frac{\mathcal{L}_n(\theta^{cand})}{\mathcal{L}_n(\hat{\theta}_n)}$$

if $M > \mathcal{L}_n(\hat{\theta}_n)$. This means it would take more tries on average to obtain a sample of size B. However, we should also not choose $M < \mathcal{L}_n(\hat{\theta}_n)$, because then the envelope condition

$$\frac{k(\theta)}{Mg(\theta)} = \frac{\mathcal{L}_n(\theta)}{M} \leq 1 \text{ for all } \theta$$

required for rejection sampling may no longer be true.

5. Suppose θ has truncated normal distribution with parameters μ, σ^2, α, and β. That is, when $\alpha < \theta < \beta$,

$$f(\theta; \mu, \sigma^2, \alpha, \beta) \propto (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \mu)^2\right\}$$

and $f(\theta; \mu, \sigma^2, \alpha, \beta) = 0$ otherwise.

   (a) Consider using rejection sampling to sample from this distribution, using the PDF of a $Normal(\mu, \sigma^2)$ distribution as your proposal density. What are the steps of the algorithm in obtaining B iid samples?

   (b) Write down an expression for the probability of acceptance in any iteration of the algorithm. When will the algorithm be efficient (i.e., have high probability of acceptance)?

   (c) Write R code to generate 1000 iid samples from the truncated normal distribution with $\mu = 10$, $\sigma^2 = 4$, $a = 9$, and $\beta = 13$. Run it and make a histogram of the samples. (Turn in both your code and your plot.)

Solution:

(a) Define

$$k(\theta) = \begin{cases} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\theta-\mu)^2\right\} & \alpha < \theta < \beta \\ 0 & \text{otherwise} \end{cases},$$

$$g(\theta) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\theta-\mu)^2\right\}.$$

Note that $f(\theta; \mu, \sigma^2, \alpha^2, \beta) = k(\theta)/\int k(\theta)d\theta$, and

$$\frac{k(\theta)}{g(\theta)} = \begin{cases} 1 & \alpha < \theta < \beta \\ 0 & \text{otherwise} \end{cases} \Rightarrow \frac{k(\theta)}{g(\theta)} \leq 1, \forall\theta,$$

and we may set $M = 1$.

Then we accept a candidate with probability

$$\frac{k(\theta^{cand})}{g(\theta^{cand})} = \begin{cases} 1 & \alpha < \theta < \beta \\ 0 & \text{otherwise} \end{cases}$$

That is, we simply accept θ^{cand} if it's in the interval and reject it otherwise (no need to draw $U \sim Unif(0,1)$).

So the algorithm is

i. Draw $\theta^{cand} \sim N(\mu, \sigma^2)$.

ii. Accept θ^{cand} if $\alpha < \theta^{cand} < \beta$; otherwise reject.

Repeat 1 and 2 until B values have been accepted.

(b)

$$\begin{aligned} P(\theta^{cand}\text{accepted}) &= P(\alpha < \theta^{cand} < \beta) = P\left(\frac{\alpha-\mu}{\sigma} < \frac{\theta^{cand}-\mu}{\sigma} < \frac{\beta-\mu}{\sigma}\right) \\ &= P\left(\frac{\alpha-\mu}{\sigma} < Z < \frac{\beta-\mu}{\sigma}\right), \text{ where } Z \sim N(0,1) \\ &= \Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right). \end{aligned}$$

(c) The `R` code is as follows.

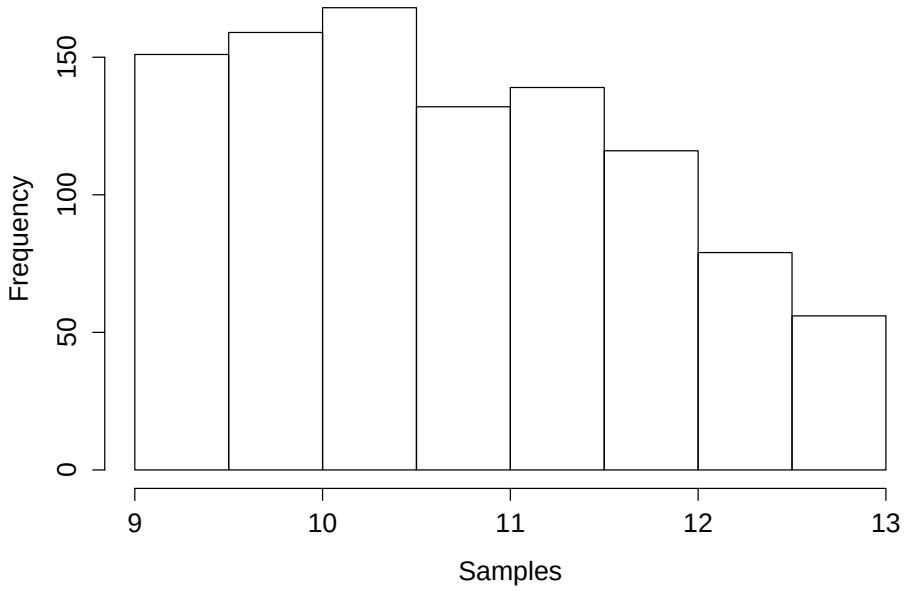
```
mu <- 10
sigma2 <- 4
alpha <- 9
```

```
beta <- 13

samps <- rep(NA, 1000)
acc <- 0
while(acc < 1000){
  cand <- rnorm(1, mu, sqrt(sigma2))
  if(cand > alpha & cand < beta){
    acc <- acc +1
    samps[acc] <- cand
  }
}

hist(samps, main = "", xlab = "Samples")
```

The histogram of samples is



Since $\Phi(x)$ is increasing in x, this probability will be high when $\frac{\beta-\mu}{\sigma}$ is large and $\frac{\alpha-\mu}{\sigma}$ is small. That is efficiency is high when there is little truncation.

6. Suppose you are in the following setting, which is a simple but realistic clinical

trial. 100 people in a control group receive a placebo (or standard treatment), and 100 receive a new experimental treatment. In the control group, $x_1 = 33$ survive at least one year, while in the experimental group, $x_2 = 38$ survive at least one year. Let p_1 be the probability of one-year survival under control and p_2 under treatment in the hypothetical population from which these patients were drawn.

(a) Find the posterior distribution for p_1 and p_2, assuming the two groups of patients are independent and that we use prior distribution $p_1, p_2 \overset{iid}{\sim} Unif(0, 1)$. You may use the result in Example 11.1 of Wasserman, but make sure you understand the steps in obtaining it.

(b) Using the method suggested on page 406 of Wasserman, sample from the posterior distribution of $\delta = p_2 - p_1$. Make a histogram of the sampled values.

(c) Suppose a doctor tells you that a value of δ greater than 0.04 would represent a breakthrough in treatment. Use your sample to approximate the posterior probability that this drug represents a breakthrough.

Solution:

(a) If we assume the two patient groups are independent, we have

$$f(p_1, p_2|x_1, x_2) = f(p_1|x_1)f(p_2|x_2).$$

Since

$$f(p_1|x_1) \propto f(x_1|p_1)f(p_1) \propto p_1^{x_1}(1 - p_1)^{100-x_1},$$
$$\Rightarrow p_1|x_1 \sim Beta(x_1 + 1, 100 - x_1 + 1) = Beta(34, 68).$$

Similarly,

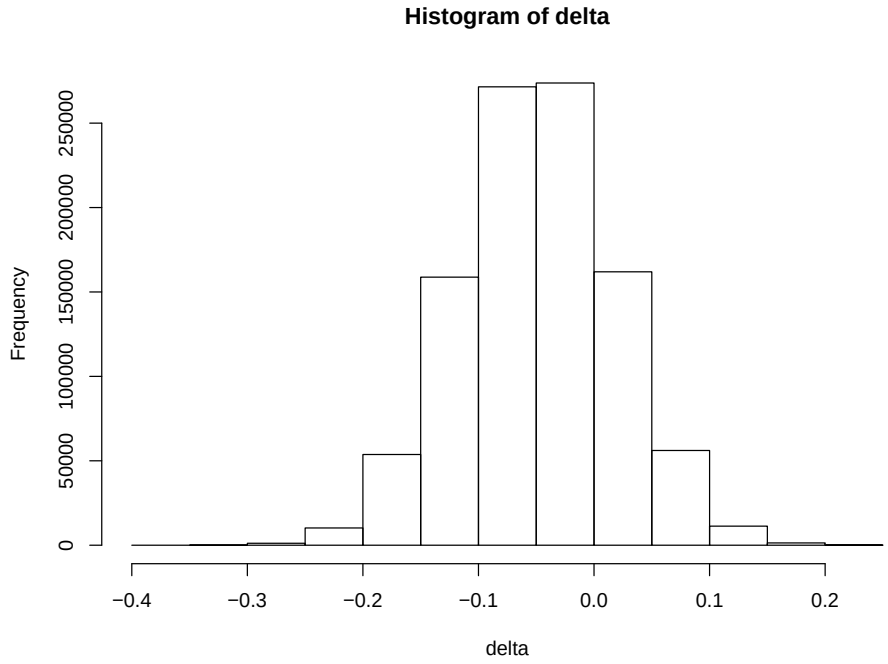
$$p_2|x_2 \sim Beta(x_2 + 1, 100 - x_2 + 1) = Beta(38, 63).$$

(b) The sample is generated from the following `R` code:

```
B <- 1e6
p1 <- rbeta(B, 34, 68)
p2 <- rbeta(B, 39, 63)
delta <- p1-p2

hist(delta)
```

The histogram of the sample is below.



(c) For the sample of δ in (b) generated from the posterior distribution, the proportion of $\delta > 0.04$ in that sample is approximately 0.09, which is calculated in `R` as

```
print(mean(delta>0.04))
```

This is approximately the posterior probability that the drug represents a breakthrough.