Name: Key

STAT 535 — Intro to Bayesian Data Analysis
Test 1 – Spring 2024

1. Fill in the blanks: Bayesian posterior inference is based on a combination of prior information and data/sample/likelihood information.

2. A quality control employees at a citrus packing plant wishes to estimate the mean number of blemishes per orange for a population of oranges to be shipped. She will take a random sample of five oranges and count the blemishes on each. Assume the five counts $Y_1, \ldots, Y_5$ can be modeled as iid Poisson random variables with unknown mean parameter λ, which the employee believes to be around 3. The standard deviation of her prior is judged to be around $\sqrt{1.5} \approx 1.22$.

(a) Given the prior knowledge, what would be a reasonable choice of a prior distribution for λ? Include hyperparameter values. Explain your choice. The gamma is the natural choice since it is the conjugate prior here.

Set $\frac{s}{r} = 3$ and $\frac{s}{r^2} = 1.5 \Rightarrow \frac{3r}{r^2} = 1.5 \Rightarrow \frac{3}{r} = 1.5$

$\Rightarrow r = 2 \Rightarrow$ Gamma$(6, 2)$

(b) Based on your prior distribution and the data model here, state the form of the posterior distribution for λ, including expressions for the parameter values. (Just state it, no need to derive it.)

Gamma$(\sum y_i + 6, n+2)$

$\Rightarrow$ Gamma$(\sum y_i + 6, 7)$

(c) If we observe sample values of 4, 0, 3, 5, 1, then write the posterior distribution for λ, specifying actual numerical parameter values.

$\sum y_i = 13$

$\Rightarrow$ Gamma$(19, 7)$

(d) If possible, give a Bayesian point estimate for λ using your posterior you found in (c).

posterior mean $= \hat{\lambda}_B = \frac{19}{7} \approx 2.71$

3. Suppose we have iid observations $Y_1, \ldots, Y_n$ that follow a distribution with pdf:

$$f(y|\theta) = 2\theta y e^{-\theta y^2}$$

where $y > 0$ and the unknown parameter is $\theta > 0$.

(a) Suppose you choose as a prior distribution for θ a gamma(s, r) distribution. Briefly explain why the gamma is a reasonable choice as a prior here.

The support of a gamma is $(0, \infty)$ which makes it reasonable since $\theta > 0$.

(b) Write (and simplify as much as possible) the likelihood function $L(\theta|y_1, \ldots, y_n)$.

$$L(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} 2\theta y_i e^{-\theta y_i^2}$$

$$= 2^n \theta^n \left(\prod_{i=1}^{n} y_i \right) e^{-\theta \sum y_i^2}$$

(c) Based on your prior distribution and the likelihood here, derive the form of posterior distribution for θ, including formulas for the posterior parameters.

$$p(\theta|\underline{y}) \propto \theta^n e^{-\theta \sum y_i^2} \theta^{s-1} e^{-r\theta}$$

$$= \theta^{n+s-1} e^{-\theta(\sum y_i^2 + r)}$$

which is the kernel of a

$$\boxed{\text{gamma}(n+s, \textstyle\sum y_i^2 + r)}$$

(d) Give a general formula for the posterior mean here, based on your answer to (c).

$$\hat{\theta}_B = \frac{n+s}{\sum y_i^2 + r}$$

(e) If we choose a gamma($s = 3, r = 15$) prior for θ and we observe a sample of $n = 3$ values which are 1.2, 4, and 2.5, then write the posterior distribution for θ, specifying actual numerical parameter values.

$n = 3$, $\sum y_i^2 = 1.2^2 + 4^2 + 2.5^2 = 23.69$

$\Rightarrow$ gamma$(6, 38.69)$

(f) Based on what you know about the form of the posterior distribution here, give a Bayesian point estimate for θ using your posterior, using the specific prior and data in (e). Your answer should be an actual number. Show work.

Posterior mean

$$\hat{\theta}_B = \frac{3+3}{23.69+15} = \frac{6}{38.69} = 0.155$$

(g) Note that, for $n = 3$ observations, the MLE of θ is

$$\hat{\theta}_{ML} = \frac{3}{y_1^2 + y_2^2 + y_3^2}$$

Briefly discuss how the posterior mean compares numerically to the prior mean and the MLE for this data set.

Prior mean $= \frac{3}{15} = 0.2$

MLE $= \frac{3}{23.69} = 0.1266$

Posterior mean is smaller than prior mean but larger than MLE.

(h) Write the general formula for the posterior mean as a weighted average of the MLE and the prior mean.

$$\frac{n+s}{\sum y_i^2 + r} = \frac{n}{\sum y_i^2 + r} + \frac{s}{\sum y_i^2 + r}$$

$$= \left(\frac{\sum y_i^2}{\sum y_i^2 + r}\right)\left(\frac{n}{\sum y_i^2}\right) + \left(\frac{r}{\sum y_i^2 + r}\right)\left(\frac{s}{r}\right)$$

4. A college student was planning to make point-spread bets on the NFL playoff games and wanted to do a Bayesian data analysis of his performance. In particular, he was interested in doing inference about his probability of winning a bet on a randomly chosen NFL playoff game.

(a) He chose a Beta(3, 2) prior for his probability of winning a bet. Based on this, before collecting the data, what is his best guess for his probability of winning a bet?

$$\frac{3}{3+2} = \frac{3}{5} = 0.6$$

(b) He bet on the 13 NFL playoff games and he won 5 of those bets. What is the posterior distribution for his probability of winning a bet? (Just state the distribution, including parameter values; you don't have to derive it mathematically).

$\pi | y \sim \text{Beta}(\alpha + y, \beta + n - y) \Rightarrow \text{Beta}(3+5, 2+13-5)$

$\Rightarrow \text{Beta}(8, 10)$

(c) Using your answer to part (b), what is a point estimate for his probability of winning a bet? Indicate how you got your answer.

$$\frac{8}{8+10} = \frac{8}{18} \approx 0.444$$

(d) Recall that the variance of a Beta(α, β) random variable is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Which sentence best reflects how the student's prior beliefs have been updated into the posterior information?

(A) After seeing the data, he has become more optimistic about his chance to win a bet, and he has become more certain about his belief.
(B) After seeing the data, he has become more optimistic about his chance to win a bet, and he has become less certain about his belief.
(C) After seeing the data, he has become less optimistic about his chance to win a bet, and he has become more certain about his belief.
(D) After seeing the data, he has become less optimistic about his chance to win a bet, and he has become less certain about his belief.

Can verify posterior variance is smaller than prior variance

5. Suppose 15 percent of all people in the population lack health insurance. Among people who lack health insurance, 5 percent of them are senior citizens. Also note that senior citizens make up 17 percent of the population.

(a) What is the probability that a randomly selected person is a senior citizen, given that the person does NOT lack health insurance? Show work. S = senior citizen, L = lacks insurance

$P(S) = P(S|L)P(L) + P(S|L^c)P(L^c)$

$\Rightarrow 0.17 = (0.05)(0.15) + P(S|L^c)(0.85)$

$\Rightarrow P(S|L^c) = \frac{(0.17) - (0.05)(0.15)}{0.85} = \boxed{0.1912}$

(b) If a randomly selected person in the population is a senior citizen, then use Bayes' Rule to find the probability that the person lacks health insurance. Show work.

$P(L|S) = \frac{P(S|L)P(L)}{P(S)} = \frac{(0.05)(0.15)}{0.17}$

$= \boxed{0.044}$

6. Consider the posterior distribution (for some parameter θ) pictured in the plot in Figure 1 at the end of the exam.

If the 95% (equal-tail) quantile-based credible interval for θ here is $(\hat{\theta}_L^Q, \hat{\theta}_U^Q)$, and the the 95% HPD credible interval for θ here is $(\hat{\theta}_L^H, \hat{\theta}_U^H)$, then which of the following can we conclude? (Feel free to draw on the plot to illustrate your reasoning.)

(A) $\hat{\theta}_L^Q < \hat{\theta}_L^H$ and $\hat{\theta}_U^Q < \hat{\theta}_U^H$ (B) $\hat{\theta}_L^Q < \hat{\theta}_L^H$ and $\hat{\theta}_U^Q > \hat{\theta}_U^H$

(C) $\hat{\theta}_L^Q > \hat{\theta}_L^H$ and $\hat{\theta}_U^Q < \hat{\theta}_U^H$ (D) $\hat{\theta}_L^Q > \hat{\theta}_L^H$ and $\hat{\theta}_U^Q > \hat{\theta}_U^H$

7. A survey in 1997 obtained lead concentration levels (in mg/kg) at 37 sampled stations in Kenya. Some R output giving the results of the analysis is given in Appendix 1 at the end of the exam.

(a) Briefly explain why the prior distribution for σ^2 in this problem is called a *conjugate* prior.

The prior is inverse gamma and the posterior is also inverse gamma but with different parameter values.

(b) Note that the prior for μ was:

$$\mu|\sigma^2 \sim N(\delta, \sigma^2/s_0)$$

Refer to the R code and the form of the posterior for μ. Comment on exactly how our prior knowledge about μ has been altered by observing the sample data, specifically referencing relevant numbers in the R code and output.

Before seeing the data, our best guess was that μ was near 30.
After seeing the data, we believe μ is around 37.05

(c) What do the two given point estimates for σ^2 indicate about the symmetry/skewness of its posterior distribution?

it is skewed (to the right) since the posterior mean is greater than the posterior median.

(d) What do the two given point estimates for μ indicate about the symmetry/skewness of its posterior distribution?

it is symmetric since posterior mean = posterior median

(e) Carefully interpret, in the context of the variable in the problem, what the given 95% credible interval for μ tells you about the mean lead concentration level in Kenya.

With posterior probability 0.95, the mean lead concentration level is between 27.838 and 46.267 mg/kg.

Appendix 1

```
> lead <- c(48,53,44,55,52,39,62,38,23,27,41,37,41,46,32,17,
+ 32,41,23,12,3,13,10,11,5,30,11,9,7,11,77,210,38,112,52,10,6)
>
> y <- lead
>
> ybar <- mean(y); n <- length(y)
>
> ybar
[1] 37.24324
> n
[1] 37
> sum(y^2)
[1] 100936
>
> # prior parameters:
>
> my.alpha <- 12; my.beta <- 110
>
> my.delta <- 30; s0 <- 1
>
> library(pscl)  # loading pscl package
>
> ### Point estimates:
>
> p.mean.sig.sq <- (my.beta + 0.5*(sum(y^2) - n*(ybar^2)) ) / (my.alpha + n/2 - 0.5 - 1)
>
> p.median.sig.sq <- qigamma(0.50, my.alpha + n/2 - 0.5, my.beta + 0.5*( sum(y^2) - n*(ybar^2) ) )
>
> print(paste("posterior.mean for sigma^2=", round(p.mean.sig.sq,3),
+       "posterior.median for sigma^2=", round(p.median.sig.sq,3) ))
[1] "posterior.mean for sigma^2= 859.221 posterior.median for sigma^2= 839.894"
>
> p.mean.mu <- ((sum(y)+my.delta*s0)/(n+s0))
>
> p.median.mu <- qnorm(0.50, mean=((sum(y)+my.delta*s0)/(n+s0)), sd=sqrt(p.median.sig.sq/(n+s0)) )
>
> print(paste("posterior.mean for mu=", round(p.mean.mu,3),
+       "posterior.median for mu=", round(p.median.mu,3) ))
[1] "posterior.mean for mu= 37.053 posterior.median for mu= 37.053"
>
> ### Marginal Interval estimates:
>
> hpd.95.sig.sq <- hpd(qigamma, alpha=my.alpha + n/2 - 0.5, beta=my.beta + 0.5*( sum(y^2) - n*(ybar^2) ) )
>
> round(hpd.95.sig.sq, 3)
[1]  570.083 1184.742
>
> hpd.95.mu <- hpd(qnorm, mean=((sum(y)+my.delta*s0)/(n+s0)), sd=sqrt(p.median.sig.sq/(n+s0)) )
>
> round(hpd.95.mu, 3)
[1] 27.838 46.267
>
```

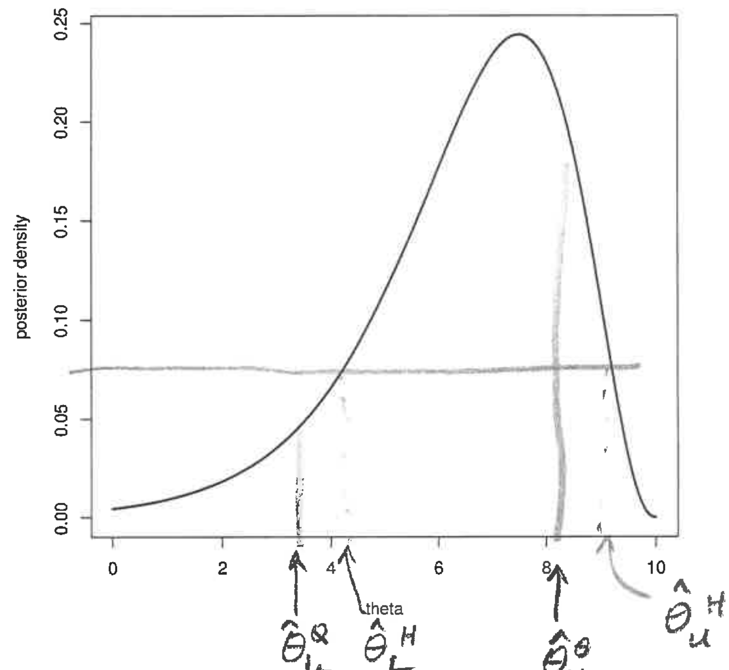

Figure 1: Posterior distribution for a parameter θ.