Stat/Econ 574B: Bayesian Statistics, HW 1

- Due Wed Sept 13 in class. HW needs to word-processed (i.e. Word or Latex). Make the presentation nice. You might intersperse code with your answers, or put all code in an Appendix. You need to be very clear on what you have done and answer questions with complete, grammatically appropriate (mathematical), sentences. Don't just boldface or italicize some numbers in the output.

Problem 1

You might find it convenient to download and install the **binom** package from CRAN. You may use this package for parts a and b, but not for the other parts except to check results. Refer to the scanned PDF on page 3 which describes diagnostic testing for Lyme disease. We will consider the specificity of the test θ, which is the probability the test is negative when given to a Lyme negative individual. In the study, 25 patients with no history of Lyme disease were tested and 19 tested negative, for a sample proportion of $19/25 = 0.76$.

a. Using the **binom.test** function in the R stats package (or the appropriate function in **binom**), compute an exact (frequentist) 95% confidence interval for θ (the so-called Clopper-Pearson interval) and compare it to the large sample confidence interval given with the problem description.

b. The exact Binomial CI can be conservative, due to discreteness of the Binomial distribution. The large sample CI can have poor coverage properties, due to discreteness and accuracy of the normal approximation. A popular alternative frequentist method was proposed by Agresti and Coull (1998), who suggested that the large sample interval be computed after augmenting the data with 2 "successes" and 2

"failures". Compute their interval and compare it to the two intervals computed in a. The **binom** package computes a slight variation on Agresti-Coull. Either version is OK.

c. Assuming a Binomial model, consider these 4 choices for a Beta(α, β) prior: (1) $\alpha = \beta = 2$; (2) $\alpha = 2; \beta = 8$; (3) $\alpha = 8; \beta = 2$ and (4) $\alpha = \beta = 11$. Using simple properties of the Beta distribution, qualitatively describe what effect you expect the prior to have on the posterior. Then make and label 4 plots, each showing the likelihood, prior and posterior. What do you see relative to your expectations? There are a several ways to plot all three on the same axes in R. If you are unsure ask me, or a colleague.

d. Assume $\theta \sim$ Beta(2,2). What is the (exact) posterior distribution, mean and standard deviation of θ? Using the **hpd** function described in class, compute a 95% HPD for θ. Using the **qbeta** function, compute an equal-tails 95% CI and compare the two. Which of the frequentist intervals computed above is closest to the Bayesian equal-tails interval? Is there a reasonable explanation why?

e. Specify your own prior for θ, and appropriately summarize the posterior. You might consider using the BetaBuster sliders, which can be accessed at

https://gjones.shinyapps.io/priorapp/

f. Consider a non-conjugate prior for θ with density $p(\theta) = 4\theta$ for $\theta < 0.50$ and $p(\theta) = 4(1 - \theta)$ for $\theta > 0.50$. Devise an importance sampling (IS) strategy (not hard here) to estimate the posterior density of θ. In particular, use SIR after IS to generate a posterior sample. Plot the SIR samples in a histogram, then smooth them using the **density** function in R. Also, estimate the posterior mean and standard deviation, and give a 95% equal-tails posterior interval. Write a short summary of your analysis.

Testing for Lyme Disease

Lyme disease is a tick-borne infection caused by the spirochete *Borrelia burgdorferi*, that, when left untreated, can cause arthritis, nervous system damage, and damage to the heart. Treatment using antibiotics in the early stages of Lyme disease is generally highly effective, but treatment is more difficult in later stages. The standard test for infection requires a time lag of up to several weeks, and can be inconclusive. In February 1999, the Centers for Disease Control and Prevention (CDC) conducted a small study of the effectiveness of a simple diagnostic immunologic assay test for infection with *B. burgdorferi* (Schriefer et al., 2000; Schutzer et al., 1999). In this study, samples were taken from 13 patients with early, culture-confirmed Lyme disease, and from 25 patients with no history of Lyme disease, and the assay test was applied to the samples. The estimated sensitivity for the immune complex assay test (the proportion of tests on patients with Lyme disease that came back positive) was $8/13 = 61.5\%$. A large-sample 95% confidence interval for the sensitivity of this test is thus

$$.615 \pm (1.96)\sqrt{\frac{(.615)(.385)}{13}} = (.351, .880).$$

Similarly, the estimated specificity for the immune complex assay test (the proportion of tests on patients without Lyme disease that came back negative) was $19/25 = 76\%$, yielding a large-sample 95% confidence interval for the specificity of

$$.76 \pm (1.96)\sqrt{\frac{(.76)(.24)}{25}} = (.593, .927).$$

The score intervals for these data are (.355, .823) and (.566, .885), respectively, and are noticeably different from the Wald intervals, especially at the upper end. These are relatively low sensitivity and specificity values; for example, the Wellcome Elisa test, the standard test for HIV, has sensitivity roughly .993 and specificity roughly .9999.

Problem 2

The table below, from Gelman et al., p.59-60, gives the number of fatal accidents and deaths on scheduled airline flights per year over a 10 year period.

Table 1: Worldwide airline fatalities. Death rate is passenger deaths per 100 million passenger miles.

Year	Fatal Accidents	Fatalities	Death Rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

a. Assume that the numbers of fatal accidents in each year are independent with a Poisson(θ) distribution, that is $Y_i|\theta_i$ are independent Poisson(θ) for $i = 1, 2, ..., 10$ where Y_i is the count for year i. Specify a Gamma prior for θ supposing that a subject matter expert told you that her best guess for θ was 20 and that she would be surprised if $\theta > 32$. To get a prior, you might read through the Jones and Johnson paper who discuss a similar problem, and use their BetaBuster app.

b. Interpret this prior in terms of "prior data": how many prior observations is the prior worth and what is the sum of the prior observations. As the number of prior observations in likely to be a non-integer, this is mostly a mental exercise to get a sense as to how informative the prior will be relative to posterior inferences. Would you expect the prior to significantly influence posterior inferences? Explain.

b. Determine (explicitly) the posterior distribution based on the data from 1976-85.

c. Plot the prior, likelihood and posterior on one set of axes and discuss what you see, especially as it relates to the impact of the prior on the posterior.

d. Compute and plot the exact predictive distribution for the number of fatal accidents in 1986. Give a 95% predictive interval for the number of fatal accidents in 1986. You might it useful to use the R function **dnbinom**. Now simulate the predictive distribution and a 95% predictive interval. Compare the exact and approximate summaries.

e. Our count model treats the years as "exchangeable" even though the numbers of flights per year varies, which likely has a bearing on the number of fatal accidents. This is a common issue with analyzing counts when the "duration of exposure" varies. Given the available data, we define the duration of exposure as the number of passenger miles flown per year and refine the model to be $Y_i|(\theta,\, t_i) \sim \text{Poisson}(\theta t_i)$, where $10^8 t_i$ is the number of passenger miles flown in year i and θ is an assumed constant mean rate per 10^8 miles flown. As before, Y_i is the number of fatal accidents. Note that t_i is treated as fixed in this analysis and can be computed (approximated) from the observed rate $r_i = Z_i/t_i$ in the table, where Z_i is the number of deaths. Assuming $\theta \sim \text{Gamma}(\alpha, \beta)$ show that the posterior of θ is $\text{Gamma}(\alpha + \sum_i y_i, \beta + \sum_i t_i)$.

f. Specify a Gamma prior for θ, but note that θ has a different interpretation from the earlier analysis. l leave it you about how to make the specification - you might consider whether the prior set earlier together with the passenger mile information in the table is useful or not. Appropriately summarize the posterior, including a 95% posterior CI. Also, give a 95% predictive interval for the number of fatal accidents in 1986 under the assumption that 8×10^{11} passenger miles will be flown that year.