ST440/540 Applied Bayesian Analysis
Lab activity for 2/5/2024

A. HOMEWORK AND QUIZ SOLUTIONS

Q3: Say the mean house price is μ_1 for houses within a 100 ft of a greenspace and μ_2 for other houses. Based on a survey, you run a Bayesian analysis to test the hypotheses:

H1: $\mu_1 \leq \mu_2$
H2: $\mu_1 > \mu_2$

You are able (using methods we will discuss in a few weeks) to draw S samples from the posterior distribution of each parameter, $\mu_1^{(1)},\ldots, \mu_1^{(S)}$ and $\mu_2^{(1)},\ldots, \mu_2^{(S)}$. Describe how you would carry out this test (including an equation or two).

I would approximate the posterior probability of H2 with the sample proportion from the MC draws with $\mu_1 > \mu_2$, that is,

$$P(H2|Y) = \sum_{s=1}^{S} I(\mu_1^s > \mu_2^s)/S,$$

and conclude H2 if the probability exceeded 0.95 (where I(x) = 1 if x is true and 0 otherwise).

B. DISCUSSION QUESTIONS

(1) What is the difference between a confidence interval and a credible interval?

A confidence interval tells you that if you repeat the procedure many times the interval will include the truth 95% of the time.

A credible interval says we are 95% certain the true value in is the interval.

(2) You are managing a factory and a collectively-bargained agreement says the injury rate should not exceed 4 injuries per month. In this past 6 months there have been 25 injuries, 4, 5, 2, 7, 1 and 6 respectively. Outline an analysis to determine whether the injury rate exceeds the agreed upon level. Select and justify your choice of likelihood and prior, and describe how you would summarize the posterior.

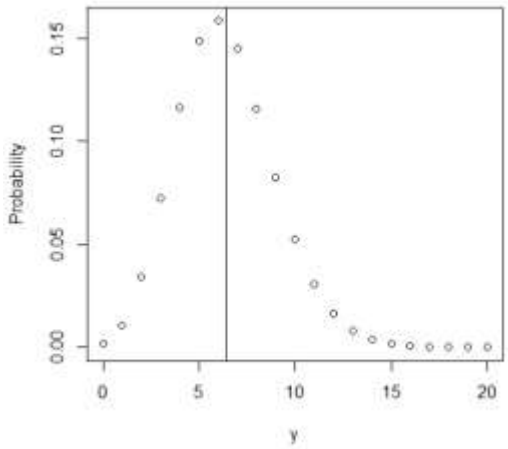
Likelihood: Since the data are counts, we assume Y_t ~ Poisson(lambda) for month t=1,…,6.

Prior: Since lambda is a positive real number we could pick a uniform or gamma prior, e.g., lambda ~ Uniform(0,25) or lambda ~ gamma(0.01,0.01).

Summary of the posterior: A summary of the posterior is the probability the injury rate exceeds 4, P(lambda>4|Y).

(3) A survey of n=10 people found that participants had a sample mean of $\bar{Y}$ = 6.4 pairs of shoes. The analysis (trust me, this is important) wanted to compute the probability that the next survey participant would have at least 10 pairs of shoes. Denote the next participant's number of shoes as Y*. They assumed Y* ~ Poisson(6.4) and got P(Y*>9) = 0.114 (code below).

```
> y <- seq(0,20,1)
> plot(y,dpois(y,6.4),ylab="Probability")
> abline(v=6.4)
> 1-ppois(9,6.4)
[1] 0.1142008
```



(a) Would you expect the probability to be higher or lower than 0.114 if it was computed using the posterior predictive distribution? Why?

It could depend on the prior, but assuming it is uninformative, because of uncertainty about the true value of the mean parameter, the PPD will be wider than the Poisson distribution giving more probability above 9.

(b) Would you expect the probability computed using the posterior predictive distribution to be more similar to 0.114 if the sample size increased or decreased? Why?

As the sample size increases the posterior of the true mean will be 6.4 with no uncertainty and so the PPD will be exactly Poisson(6.4) and the PPD prob will be 0.114.

(c) Describe how to compute probabilities from the posterior predictive distribution using R.

We would use Monte Carlo sampling to get a bunch of samples for Y*, and then compute the proportion of the MC samples that are at least 10.

(4) Say Y|θ ~ Binomial(n, θ) and θ ~ beta(a,b). The derivation of the posterior in the notes was

$$p(\theta|Y) \propto f(Y|\theta)\pi(\theta) \propto \theta^{(Y+a)-1}(1-\theta)^{(n-Y+b)-1}$$

and then we concluded that the posterior was Beta(Y+a,n-Y+b). Explain why we never had to compute the marginal distribution of the data, m(y).

The marginal m(y) as a function of theta is a constant, so we can just say it's proportional to the term that include theta.

(5) Assume Y1|θ ~ Binomial(n1,θ) independent of Y2|θ ~ Binomial(n2,θ), and the prior is θ ~ Beta(a,b). Derive the posterior distribution of θ|Y1,Y2.

Since Y1 and Y2 are independent the likelihood is f(Y1,Y2|theta) = f(Y1|theta)f(Y2|theta). Thus

$p(\theta|Y1, Y2) \propto f(Y1, Y2|\theta)\pi(\theta)$

$\propto f(Y1|\theta)f(Y2|\theta)\pi(\theta)$

$\propto [\theta^{Y1}(1-\theta)^{n1-Y1}][\theta^{Y2}(1-\theta)^{n2-Y2}][\theta^{a-1}(1-\theta)^{b-1}]$

$\propto \theta^{(Y1+Y1+a)-1}(1-\theta)^{(n1-Y1+n2-Y2+b)-1}$

Therefore, the posterior is Beta(Y1+Y2+a, n1-Y1+n2-Y2+b), or Beta(Y+A, n+b) where Y=Y1+Y2 and n=n1+n2.

(6) You are tasked with writing the R package to conduct a Bayesian analysis of a proportion.

(a) What are the inputs to your function and what (if any) are the default values?

The model is Likelihood Y ~ Binomial (n,theta) and theta ~ beta(a,b). So we would have to have inputs Y, n, a and b. Y and n have no defaults, but maybe a=b=1 is a reasonable default.

(b) What are the outputs?

A table with the posterior mean, mode and 95% credible interval. Other options are a plot of the prior and posterior PDFs and the probability above a threshold so the users could test say Ho: theta<0.5 versus Ha: theta>0.5.