

Stat 544 – Spring 2005

Homework Assignment 2

Due on Friday, February 11 in TA's office by 5:00 pm

Problem 1

- a- Let y_1, y_2, \dots, y_n be an exchangeable sample from a $N(\mu, \sigma^2)$ distribution, where both the mean and the variance are unknown. Derive

$$p(\mu, \sigma^2),$$

the Jeffrey's non-informative prior distribution.

- b- Let n be a negative binomial random variable $\text{Neg-Bin}(x, \theta)$, with x known. Obtain the Jeffreys prior for θ .
- c- If $x \sim \text{Bin}(n, \theta)$ the Jeffreys prior for θ is the $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ distribution (see class notes). Compare this prior to the one you obtained in part (b) and comment on the result relative to the Likelihood Principle. [Hint: the likelihoods of the binomial and the negative binomial are proportional. Consider what happens with the posterior for θ and thus to inferences about θ when you use the Jeffreys prior in each case.]

Problem 2

An instructor for Stat 101 wishes to know whether reinforcing statistical concepts using graphical methods (after they have already been introduced using only formulae) is an effective way to improve student learning. She has 30 students in her class and carries out the following experiment. Students are given a quiz on a set of topics that were introduced using only formulae and scores are recorded. She then repeats the lecture, but using mostly graphical ideas, and gives the students the same quiz. The response variable of interest is the difference in scores on quizzes one and two. She hypothesizes that the average increase in score is $\mu_0 = 10$ points (on a scale of 0 to 100). The data are stored in the file called scores.dat and that is posted on the Homework page of the course web site.

Let y_i denote the difference in the score between quiz two and quiz one for the i th student, $i = 1, \dots, 30$ and assume that $y_i \sim N(\mu, \sigma^2)$ with (μ, σ^2) unknown. The instructor has some information from the results of a similar experiment in other sections of Stat 101, and decides to use a conjugate prior distribution for the parameters of the form

$$\begin{aligned}\mu|m, \sigma^2, \kappa &\sim N(m, \sigma^2/\kappa) \\ \sigma^2|\nu_0, \sigma_0^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2),\end{aligned}$$

with $m = 7$, $\kappa = 2$, $\nu_0 = 3$ and $\sigma_0^2 = 10$.

- a- Write down the joint posterior distribution of (μ, σ^2) .
- b- Explain (in your own words) how you might go about drawing pairs $(\mu^{(k)}, \sigma^{2(k)})$ from the joint posterior distribution.
- c- Obtain a sample of size 1000 from $p(\mu, \sigma^2|y)$ the joint posterior distribution of (μ, σ^2) and draw a scatter plot of the joint distribution using your draws. Explain in words what you see from the scatter plot. Without doing any calculations, what is your rough estimate of the posterior means of (μ, σ^2) ?
- d- Obtain the 2.5th, 25th, 50th, 75th and 97.5th posterior percentiles for σ^2 , for σ and for $\log \sigma$ (watch out for draws for which $\sigma = 0$).
- e- How likely is it that the increase in scores between quizzes one and two is at least 10 points? At least 6 points? Compute the posterior probabilities $Pr(\mu \geq 10|y)$ and $Pr(\mu \geq 6|y)$.
- f- Derive the posterior predictive distribution $p(\tilde{y}|y)$, where \tilde{y} is the difference in quiz scores obtained by a new student in the group. Please summarize the posterior predictive distribution by listing its mean, standard deviation, and a set of percentiles of your choice. Are the observations in the sample consistent with the posterior predictive distribution that you just derived?
- g- Because her sample size is not large, the instructor wonders whether results are sensitive to her choice of prior parameters. She wishes to investigate the effect of the prior on the 95% credible set for the mean change in score μ . Using the following alternative set of prior parameters, obtain 95% credible sets for the mean μ and discuss your results:
 - 1- $m = 7, \kappa = 20, \nu_0 = 3$ and $\sigma_0^2 = 10$
 - 2- $m = 2, \kappa = 2, \nu_0 = 3$ and $\sigma_0^2 = 10$
 - 3- $m = 2, \kappa = 20, \nu_0 = 3$ and $\sigma_0^2 = 10$
 - 4- $m = 7, \kappa = 2, \nu_0 = 9$ and $\sigma_0^2 = 10$

Problem 3

Observations y_1 and y_2 are independently distributed as $N(\theta_i, 1)$ for $i = 1, 2$, given (θ_1, θ_2) . The prior distribution for the means is the improper uniform distribution

$$p(\theta_1, \theta_2) \propto 1$$

but subject to the constraint $\theta_1 > \theta_2$.

- a- Show that the posterior distribution $p(\theta_1, \theta_2|y_1, y_2)$ is a truncated bivariate normal distribution.

b- Show that the marginal posterior distribution of θ_1 is

$$p(\theta_1|y_1, y_2) \propto \phi(\theta_1 - y_1)\Phi(\theta_1 - y_2)$$

where ϕ and Φ are the standard normal p.d.f. and c.d.f. respectively.

c- Show also that

$$E(\theta_1|y_1, y_2) = y_1 + \frac{\phi(d)}{\sqrt{2}\Phi(d)}$$

with $d = (y_1 - y_2)/\sqrt{2}$.

Hint: you may want to work with the following transformations $X_1 = Y_1 - Y_2$ and $X_2 = Y_1 + Y_2$, and then back-transform your results.

Problem 4

Suppose that the p.d.f. of $X|\sigma$ is a (one-dimensional) scale density, i.e. it is of the form

$$\frac{1}{\sigma} f\left(\frac{1}{\sigma}\right)$$

Give a vague prior density for σ . Provide some kind of explanation on how you arrived to such a prior.

Problem 5

Let's consider again exercise 1.12.17. Recall that the problem was "The participant will be offered the choice of three boxes, and after she has chosen the host will open a different box revealing a lesser price. What is the probability that the participant will win if the participant strategy is to switch?" You may have found that the probability of winning by switching was $2/3$, so the participant should always switch to increase her probability of winning.

Now, consider the following changes to our problem: The participant will be offered the choice of three numbered boxes, one contains a prize and the other two are empty. After she has chosen, the host will open one of the remaining boxes. Suppose that the participant has chosen box number 1. So an element of the sample space can be written as:

PEE2 = prize is in box 1, box 2 and 3 are empty and host opens box 2

Further, note that if we let p_{ij} denote the probability that host will open box j given that the prize is in box i we have that $P(\text{PEE2}) = p_{12}/3$.

Suppose now that the host opens box 3 and it is empty. Under these conditions, when should the participant switch to improve her probability of winning if the host had one of the following strategies (unbeknownst to the participant):

- a- The host had probability p of opening box 2 and probability $q = 1 - p$ of opening box 3. Note, that this strategy implies that the host did not use the information about which box contained the price.
- b- The host will open box 3 whenever possible, i.e. the host will open box 3 if and only if he knows that the prize is not in box 3.
- c- If the prize is in box 1, the host will open: box 2 with probability p , box 3 with probability $q = 1 - p$. Otherwise, it will open box 3 whenever possible.

Even though that your answers may depend on the unknown probabilities p_{ij} you always will be able to decide whether the participant should switch.

Problem 6

You are happy as a clam with your WinBUGS¹ program. You can sample from posterior distributions even though that you do not have a clue about how WinBUGS does it. So you have decided to use it again. This time you want to solve the following problem:

The number of hepatitis B cases in 15 regions of Berlin were recorded for 1998. Epidemiologists are interested in estimating the rate of hepatitis B per 10,000 individuals per year. The following table shows the observed number of cases in each of the 15 regions together with the population in each of the regions (per 10,000 individuals) for the year 1998.

Region	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	16	9	12	5	11	20	5	8	12	15	14	6	2	7	7
Pop.	13.5	16	24	5.5	22	40	10	16.5	21.5	23.5	25	8.5	3	8	11.5

Let λ_i denote the unobserved rate in the i th region.

- a- Write down an appropriate sampling model for these data. Justify your choice of a sampling model.
- b- Assuming a Gamma(0.01,0.01) prior distribution and given your sampling model obtain a sample of 10,000 values from $\lambda_i|y$. You do not have to turn in your draws, you should only attach your WinBUGS code.
- c- Using those 10,000 values compute and report the posterior mean, standard deviation, and percentiles 0.025, 0.50, and 0.975 for region 2, 6 and 15. Create also histograms of your 10,000 draws for the aforementioned regions.

¹BTW: BUGS stands for **B**ayesian inference **U**sing **G**ibbs **S**ampling