

MAS2317/3317: Problems sheet 2

Introduction to Bayesian Statistics

PROBLEMS SHEET 2 is to be used in this week's practical session (5pm, Thurs 20th Feb), and next week's problems class (4pm, Thurs 27 Feb). Further help will be available in the drop-in session the week after (4pm, Thurs 6 Mar).

Answers to the starred (*) questions should be placed in the homework submission box in the foyer of the Maths & Stats General Office (Herschel Building 3rd floor) by no later than **4pm, Tuesday 11th March**.

In order to complete some of this work, you will need the handout for **Case Study 1: Speed Cameras and Regression To the Mean**.

Questions marked with a P are to be used in student presentations.

- 1.* The 18th century physicist Henry Cavendish made 23 experimental determinations of the earth's density, and these data (in g/cm^3) are given below.

5.36	5.29	5.58	5.65	5.57	5.53	5.62	5.29
5.44	5.34	5.79	5.10	5.27	5.39	5.42	5.47
5.63	5.34	5.46	5.30	5.78	5.68	5.85	

Suppose that Cavendish asserts that the error standard deviation of these measurements is $0.2 g/cm^3$; he also assumes that they are normally distributed with mean equal to the true earth density μ . Using a normal prior distribution for μ with mean $5.41 g/cm^3$ and standard deviation $0.4 g/cm^3$,

- (a) obtain the posterior distribution, and compare this to the prior;
 - (b) find the posterior probability that $5.45 g/cm^3 < \mu < 5.55 g/cm^3$.
2. Suppose that a random sample $x_1, x_2, \dots, x_n$ is obtained. Derive the posterior distributions for the following models:
- (a) $f(x|\theta) = \theta^{x-1}(1 - \theta)$, $x = 1, 2, \dots$ with a $U(0, 1)$ prior distribution for θ .
 - (b) $f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$, $x = 0, 1, \dots$ with a $Exp(1)$ prior distribution for θ .
 - (c) $f(x|\theta) = \binom{7}{x} \theta^x (1 - \theta)^{7-x}$, $x = 0, 1, \dots, 7$ with a $Beta(3, 2)$ prior distribution for θ .

- 3.* [It might help if you have the handout for case study 1 handy when completing this question] Suppose the number of casualties Y_i at location i , due to road traffic accidents, is assumed Poisson with rate θ_i , i.e.

$$Y_i|\theta_i \sim Po(\theta_i).$$

Suppose further that we assume the following gamma prior for θ_i :

$$\theta_i \sim Ga\left(g, \frac{g}{\mu_i}\right).$$

- (a) Show that the prior mean and standard deviation for θ_i are μ_i and $\mu_i/\sqrt{g}$ respectively.
- (b) Suppose that we observe $Y_i = y_i$ casualties at site i during some observation period.
 - (i) Derive the posterior distribution $\pi(\theta_i|y_i)$, i.e. confirm the result given by Equation (3) in the case study.
 - (ii) Show that the posterior mean is given by the following Bayes linear rule:

$$E(\theta_i|y_i) = \alpha_i\mu_i + (1 - \alpha_i)y_i,$$

where

$$\alpha_i = \frac{g}{g + \mu_i}.$$

- 4.^P The time until a bicycle tyre gets punctured is exponentially distributed with mean $1/\theta$ weeks. From a cyclist, we are able to elicit the following prior for θ :

$$\pi(\theta) = \frac{100^6}{120}\theta^5 e^{-100\theta}. \quad (1)$$

- (a) Identify fully the distribution in Equation (1).
 - (b) The cyclist records the times that two tyres last, and his readings (in weeks) are 33 and 41. Find how his beliefs about θ change in light of these data.
- 5.^P Consider an experiment with a possibly loaded six-sided die. Let $\theta = \text{Pr}(\text{rolling a six})$. Before conducting the experiment, we believe that all values of θ are equally likely and so we use a $U(0, 1)$ random variable as the prior distribution for θ . We then roll the die 10 times and observe 3 sixes.
- (a) Obtain the likelihood function and use it to obtain the posterior distribution, $\pi(\theta|x = 3)$.
 - (b) What is $E[\theta|x = 3]$?
- 6.^P Richard plays “flip to win” with a student. The game is very simple – Richard tosses a coin and if the outcome is a head, the student gets a fiver. If the coin lands tails, the student must buy Richard a pint. Five students play this game with Richard, and only one head is observed. Let X be the number of heads observed; thus, $X \sim \text{bin}(5, \theta)$.
- (a) Find the likelihood function $f(x = 1|\theta)$.
 - (b) The students are suspicious of Richard’s coin. In fact, Daniel, an insider, tells them that he believes the coin is biased and thus specifies a $\text{Beta}(6, 46)$ distribution as the prior for θ . Why might this be a sensible prior distribution in light of these suspicions?
 - (c) Obtain the posterior distribution for θ , $\pi(\theta|x = 1)$.

7. The data in the table below are measurements of the logarithm of *ornithine carbonyltransferase* (a liver enzyme) in patients suffering from acute viral hepatitis.

2.66	2.38	2.37	2.50	1.96	2.85	2.68	1.76	2.36
2.56	2.09	2.85	2.37	2.40	2.79	1.82	3.00	2.50
2.36	2.48	2.60	2.42	2.51	2.51	2.80	2.50	2.57

Assume that the enzyme measurement varies according to a $N(\mu, 0.27^2)$ distribution, and that an expert says his beliefs about the mean level μ are described by a $N(2.7, 0.3^2)$ distribution. After combining both sources of information, what is the probability that this mean level is less than 2.5? Sketch the prior and posterior distributions for μ on the same graph. Describe the main changes in beliefs after incorporating the data.

8. The number of defects in a 1200 foot roll of magnetic recording tape has a $Po(\theta)$ distribution. The prior distribution for θ is $Ga(3, 1)$. When 5 rolls of this tape are selected at random and inspected, the number of defects found on the rolls are 2, 2, 6, 0 and 3. Determine the posterior distribution of θ .
9. A random sample of size n is to be taken from a $N(\theta, 2^2)$ distribution. The prior for θ is $N(b, 1/d)$.
- If $d = 1$, what is the smallest number of observations that must be included in the sample if the standard deviation of the posterior distribution of θ is to be reduced to 0.1?
 - If $n = 100$, show that no matter how large the value of d , the standard deviation of the posterior distribution is less than $1/5$.
10. In viewing a section through the pancreas, doctors see what are called “islands”. The numbers of islands seen in 900 patients are summarised below.

Number of islands	0	1	2	3	4	5	6	≥ 7
Frequency	327	340	160	53	16	3	1	0

Assuming that the numbers of islands are Poisson distributed with mean θ , obtain the posterior mean and standard deviation for θ from several gamma $Ga(g, h)$ prior distributions (with g and h in the range 1 to 50). Why is the posterior distribution quite insensitive to variations in the prior distribution?

Practical work

Open R by clicking on **Start** → **All Programs** → **Statistical Software** → **RStudio**. You will also need to access data for some of these questions from the course website, so have this open:

<http://www.mas.ncl.ac.uk/~nlf8/teaching/mas2317/>

11. Recall Example 2.1 in the lecture notes. We have an experiment with a possibly biased coin, and $\theta = \Pr(\text{Head})$. The number of heads observed, X , from 5 flips of the coin, is therefore a binomial random variable, i.e.

$$X|\theta \sim \text{Bin}(5, \theta).$$

Assuming that $\theta \sim U(0, 1)$, we found that $\theta|x = 1 \sim \text{Be}(2, 5)$. Let's now re-produce the plots in Figure 2.3, comparing prior with posterior.

In R, type the following:

```
> x=seq(0,1,0.01)
```

You can now look at **x** by typing

```
> x
 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13 0.14
 0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
 0.30 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44
 0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74
 0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89
 0.90 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```

Thus, **x** is just a sequence of values between 0 and 1, in steps of 0.01. This covers the range of possible for θ .

Now, for the $U(0, 1)$ prior for θ , we apply the R function **dunif** to **x**:

```
> prior=dunif(x,0,1)
```

This evaluates the $U(0, 1)$ pdf at each of our **x** values.

For the $\text{Be}(2, 5)$ posterior for θ , we apply the R function **dbeta** to **x**:

```
> posterior=dbeta(x,2,5)
```

This evaluates the $\text{Be}(2, 5)$ pdf at each of our **x** values.

We will now plot the prior and posterior in R on the same graph, as in Figure 2.3 of the lecture notes. In R, type

```
> plot(x,prior,type='l',xlab='theta',main='Example 2.3',lty=2)
```

This should plot your $U(0,1)$ prior for θ . The arguments `xlab=` and `main=` specify labels for your x -axis and the plot as a whole, respectively. Using `type='l'` uses a line to join the points in the plot, and setting `lty=2` uses a dashed line effect.

Now to superimpose the posterior. Type

```
> lines(x,posterior)
```

Notice that when you do this, your beta posterior is superimposed (the `lines` command superimposes, instead of making a brand new plot); however, the range on the y -axis is not large enough to see this properly. We can fix this by using the command `ylim=c(lower,upper)` to specify a more suitable range. For example, inserting `ylim=c(0,max(posterior))` along with `type='l'`, `xlab='theta'` etc. within the plotting command `plot(...)` starts the plot again, but now the y -axis will go from 0 to the largest value in our posterior distribution. Now

```
> lines(x,posterior)
```

should produce the desired effect. Check this against Figure 2.3 in the lecture notes.

12. Recall Example 2.3 in the lecture notes. Max, the video game pirate, is interested in the proportion θ of customers who will buy *Call of Duty* from him. X is the number of people who say they *would* buy this game; if he asks 5 people, then we have

$$X|\theta \sim \text{Bin}(5, \theta).$$

He assumes a beta prior for θ : $\theta \sim \text{Be}(2.5, 12)$, which results in $\theta|x = 4 \sim \text{Be}(6.5, 13)$. Following carefully what you did in question 11, produce a plot in R which compares the prior and posterior distributions for θ . Compare your plot to that in Figure 2.7 of the lecture notes.

13. Recall the earthquakes example in the lecture notes: Example 2.4. Here, we assume the times between earthquakes X are exponentially distributed, i.e.

$$X|\theta \sim \text{Exp}(\theta).$$

An expert tells us that $\theta \sim \text{Ga}(10, 4000)$. We then combine our observations with the expert's prior beliefs to obtain $\theta|x \sim \text{Ga}(30, 13633)$. Recall that our strong expert opinion suggest that only very small values of θ are likely (as earthquakes are relatively rare); thus, set up your x values in R by typing

```
> x=seq(0,0.01,0.001)
```

Of course, the gamma distribution is defined over the *entire* positive real line, but we only look at very small values of θ here. Now use the R command `dgamma` in exactly the same way you used `dunif` and `dbeta` in parts (a) and (b) to produce a plot comparing your prior and posterior distributions for the rate of earthquakes θ . Again, you might have to change the range of the y -axis in your plot to accommodate the posterior. Compare you plot to that in Figure 2.9 of the lecture notes.

14.* In R, type

```
> par(mfrow=c(2,2))
```

You are required to produce four plots in this question; the above command splits the plotting window into four components (2 by 2), one for each of the plots you will produce. You should include this single panel showing all four plots in your solutions to this question. Using the same approach as in question 11,

- (a) produce a plot of the $Be(10, 37)$ density for values on the x -axis between 0 and 1;
- (b) produce a plot of the $Be(20, 77)$ density for values on the x -axis between 0 and 1;
- (c) produce a plot of the $Ga(15, 0.625)$ density for values on the x -axis between 0 and 100;
- (d) produce a plot of the $Ga(9, 0.36)$ density for values on the x -axis between 0 and 100.

You should now have a single panel of four plots, one each for parts (a), (b), (c) and (d). Now read the following scenarios.

Scenario 1

A football fan tells us it is quite unlikely that her team will get relegated this year. Let $\theta = \Pr(\text{her team will be relegated})$. We are able to elicit from her that there is less than a 1 in 100,000 chance that $\theta > 0.5$.

Scenario 2

We are interested in the rate of retreat θ (feet per year) of the *Zachriae Isstrøm* glacier in Greenland, as this could be an indicator of climate change (see the recent BBC *Frozen Planet* series). An expert glaciologist tells us that this year, he expects this glacier to retreat by about 24 feet, although there is some uncertainty here. In fact, he thinks that there is 1/100 chance that the glacier might retreat by less than 10 feet.

- (e) Match θ in each of the scenarios above with the most suitable distribution from parts (a)—(d). You can use the R commands `pbeta(q,a,b)` and `pgamma(q,a,b)` to find $\Pr(\theta < q)$, for $\theta \sim Be(a, b)$ and $\theta \sim Ga(a, b)$, respectively.

Please turn over for questions 15 and 16

Questions 15 and 16 relate to **case study 1: speed cameras and regression to the mean**. Recall that there were 17 road traffic accident blackspot sites identified in South Tyneside (see Section 5 of the case study). The file `blackspots.txt` on the course website has data relating to the average observed speed of vehicles, and the number of casualties in the periods *before* and *after* the introduction of speed cameras, at each of these sites.

Recall also that Northumbria Police collected information on the average observed vehicle speed and the observed number of casualties (throughout the same time as the *before* period for the blackspot sites) from a further 67 “reference” sites; these data can be found on the course website in the file `reference.txt`.

15.* This question relates to data collected at the **reference sites**.

- (a) Download the reference data from the MAS2317 course webpage by executing the following code in R:

```
> ref = read.table("http://www.mas.ncl.ac.uk/~nlf8/teaching/  
+ mas2317/case/reference.txt")
```

Do not enter the + sign. I have done this because the code would not fit on a single line in this handout. You should enter the code in a single line!

- (b) Look at the data by typing

```
> ref
```

The first column contains the average observed speed of vehicles at each of the 67 reference sites, and the second column contains the number of casualties at each of these sites, during an observation period. Typing

```
> AvSp = ref[,1]  
> Cas = ref[,2]
```

will store the data in column 1 of `ref` in the object `AvSp`, and the data in column 2 of `ref` in the object `Cas`.

- (c) Use the `plot` command in R to produce a scatterplot of the number of casualties (Y) against the average observed speed (X). You should make sure you have appropriately labelled axes and a title on your plot; if in doubt, typing

```
> ?plot
```

will bring up a help screen for the command `plot`. Comment on the relationship you see, but at this stage *do not include this plot in your solutions*.

- (d) Perform a simple linear regression of casualties on the average observed speed by typing

```
> lm(Cas~AvSp)
```

In your solutions, write down the resulting regression equation, and make sure this confirms what you see in Equation (2) of the case study.

- (e) You should now superimpose the regression line from part (d) onto your plot in part (c) and include this updated plot in your solutions. This can be done by typing:

```
> lines(AvSp,fitted(lm(Cas~AvSp)))
```

16.* This question will use your solutions to question 15 and the equations given in the case study handout, as well as data collected at the **blackspot sites**.

- (a) Some results based on analyses of blackspot sites 1 and 10 were given in the case study (see Section 5.5). Now, each of you have been randomly allocated to one of the other 15 blackspot sites; go to the “Case Studies” section of the course website to find out which site you have been allocated to, and make a note of this.
- (b) Download the blackspot dataset in exactly the same way you downloaded the reference dataset in question 15(a), and store this in the object **blackspot** in R (as you stored the reference data in **ref**).
- (c) Look at the data. As with the reference data, the first column relates to the average observed vehicle speed at each site. The second column relates to the number of casualties at each blackspot site in the period *before* the installation of speed cameras, and the third column gives the corresponding number in the period *after* the cameras had been installed. As before, each row is for one of the sites. Look at the row for the blackspot site you have been allocated to; in your solutions, write down your blackspot site number ($i =$), the average observed vehicle speed for your site ($x_i =$), the number of observed casualties at this site in the *before* period ($y_i =$) and the number of casualties in the *after* period ($y_{i,\text{after}} =$).
- (d) Estimate the mean casualty frequency μ_i at your blackspot site by substituting your average observed vehicle speed x_i into the regression equation obtained in question 15(d) from the reference data. Make sure you write this in your solutions.
- (e) In the case study, you are told that a road safety expert has suggested that $g = 2.5$ in the following distribution for the casualty rate θ_i at blackspot site i :

$$\theta_i \sim Ga\left(g, \frac{g}{\mu_i}\right)$$

(see question 3 and Equation (1) in the case study). Use your answer to part (d) to write down the distribution for θ_i for *your* blackspot site.

- (f) Assuming that Y_i is the number of casualties at blackspot i , and that

$$Y_i|\theta_i \sim Po(\theta_i),$$

use your answer to question 3(b)(i), or Equation (3) in the case study, to write down the posterior distribution for $\theta_i|y_i$ for *your* blackspot site.

- (g) Using similar commands in R to those used in questions 11–14, produce a plot of the prior and posterior distribution of θ_i for *your* blackspot site, similar to that shown for site 1 in Figure 4 of the case study. Make sure these two plots are given on the same graph (use the **lines** command in R for the second graph). Also, show the position of your observed casualty frequency by using the command

```
> abline(v=*)
```

where you would replace $*$ with *your* observed casualty frequency y_i .

- (h) With reference to Equation (4) the case study, or question 3(b)(ii), calculate the posterior mean $E(\theta_i|y_i)$ for *your* blackspot site. Does this give more weight to the prior mean μ_i or the observed casualty frequency y_i ?
- (i) After accounting for the phenomenon of *Regression To the Mean* (RTM), compute the percentage reduction in casualty frequency attributed to the speed camera at *your* blackspot site.