

MAS2317/3317: Problems sheet 4

Introduction to Bayesian Statistics

PROBLEMS SHEET 4 is to be used in this week's problems class (4pm, Thurs 27th Mar). Further help will be available in the drop-in session after Easter (4pm, 1st May). There will also be an optional practical session that week for extra help with question 9 (5pm, Thurs 1st May). Questions 7 and 8 will be discussed in the problems class in week 11 (4pm, Thurs 8 May).

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 6th May**.

1. Suppose that the random variable Y follows a $Ga(a, b)$ distribution. Determine the probability density function for the random variable $X = 1/\sqrt{Y}$.
2. Using a random sample from a $Bin(k, \theta)$ (with k known), determine the posterior distribution for θ assuming
 - (i) vague prior knowledge;
 - (ii) the Jeffreys prior distribution;
 - (iii) a very large sample.
- 3.* Using a random sample from a $Ga(k, \theta)$ distribution (with k known), determine the posterior distribution for θ assuming
 - (i) vague prior knowledge;
 - (ii) the Jeffreys prior distribution;
 - (iii) a very large sample.
4. Use techniques from MAS2302 to obtain 95% frequentist confidence intervals for θ_C and θ_I in question 9.
- 5.* Compare the frequentist and Bayesian confidence intervals for θ_C and θ_I as obtained in questions 4 and 9 (respectively). Comment on any numerical differences, as well as differences in interpretation.
6. Sketch a probability density function for which the highest density interval (HDI) takes the form
 - (i) (a, b)
 - (ii) $(a, b) \cup (c, d)$
 - (iii) $(a, b) \cup (c, d) \cup (e, f)$.where $a < b < c < d < e < f$.

7. The weights of items from a certain production processes are independently and identically distributed, each with a $N(\theta, 4)$ distribution. The production manager believes that θ varies from batch to batch according to a $N(110, 0.4)$ distribution. A sample of 5 items is randomly selected from a batch, giving data: 108, 109, 107.4, 109.6, 112. Determine the posterior distribution for θ .

- (i) Determine the predictive distribution for the weight of one further item from the batch.
- (ii) Calculate the (predictive) probability that the weight of this item exceeds 110.
- (iii) Determine the predictive distribution for the sample mean of the weights of m further items from the batch. What happens as $m \rightarrow \infty$?

Hint: if the posterior distribution is $\theta|\mathbf{x} \sim N(B, 1/D)$ then the predictive distribution for a future item y , where $Y|\theta \sim N(\theta, 1/k)$, is

$$f(y|\mathbf{x}) = \frac{f(y|\theta)\pi(\theta|\mathbf{x})}{\pi(\theta|\mathbf{x}, y)} = \left\{ \frac{kD}{2\pi(D+k)} \right\}^{1/2} \exp \left\{ -\frac{kD}{2(D+k)}(y-B)^2 \right\},$$

that is, $Y|\mathbf{x} \sim N\left(B, \frac{D+k}{kD}\right)$.

8. The distribution of flaws along the length of an artificial fibre follows a Poisson process, and the number of flaws in a length ℓ is $Po(\ell\theta)$. Very little is known about θ . The number of flaws in five fibres of lengths 10, 15, 25, 30 and 40 metres were found to be 3, 2, 7, 6, 10 respectively. We are interested in finding the predictive distribution for the number of flaws in another piece of length 60 metres.

Consider first the general case.

- (i) Suppose the number of flaws are independent with $X_i \sim Po(\ell_i\theta)$. Determine the posterior distribution for θ .
- (ii) Determine the predictive distribution for the number of flaws Y in another piece of length ℓ .
- (iii) Use the information in the data to determine the number of flaws in another piece of length 60 metres.

Hint for (ii): what is the probability function for $T = W + r$ when $W \sim NegBin(r, p)$?

Practical work

Open R by clicking on Start → All Programs → Statistical Software → R → R. You will also need access to the *MATCH Uncertainty Elicitation Tool* – this can be found at

<http://optics.eee.nottingham.ac.uk/match/uncertainty.php>

- 9.* Gravestones provide a useful means of observing and measuring the degradation of different types of rock over time. One method of measuring the extent of degradation is *Lettering Alteration*, as outlined in Rahn (1971) and Meierding (1981).

In a recent study, researchers were interested in the degradation of granite. To investigate, a random sample of granite gravestones was taken from a cemetery in Savannah, a city along the southeast coast of the USA; the method of lettering alteration was used to assess the degradation of these gravestones. These data can be found in the file **granite-coastal** on the course website.

A random sample of granite gravestones was also taken from a cemetery in Macon, Georgia, a town located 170 miles inland from Savannah. These data can be found in the file **granite-inland** on the course website.



- (a) Download the two datasets by right-clicking on their links; each time, you should select **Save Target As** and save the files anywhere that is convenient for you. You can then read the datasets into R by carefully typing in the following code:

```
> coastal=read.table('granite-coastal.txt')
> inland=read.table('granite-inland.txt')
```

Alternatively, datasets can be read directly into R. For example:

```
>coastal=read.table('http://www.mas.ncl.ac.uk/~nlf8/teaching/
+ mas2317/problems/granite-coastal.txt')
```

You can then look at the data by typing **coastal** and **inland** at the R command prompt. Notice that each dataset contains two columns: for each gravestone, we have the age of the gravestone (the difference between the current year and the year the person died), and the depth of degradation as determined by the lettering alteration method. For each dataset, you should obtain the rate of degradation per year, by dividing the depth of degradation by the age of the gravestone. For example, for the coastal site:

```
> rate.coastal=coastal[,2]/coastal[,1]
```

Use R to find the mean and standard deviation annual rate of granite degradation for each dataset, and show these values in your solutions.

- (b) Researchers are interested in rock which degrades at a rate of more than 0.12mm per year. From your data in R, estimate θ_C and θ_I , the proportion of gravestones with an annual degradation rate of at least 0.12mm at the coastal and inland locations, respectively. Also find the likelihood functions for θ_C and θ_I .
- (c) A geologist is asked about the proportion of granite gravestones she would expect to see with a rate of degradation of at least 0.12mm per year. The geologist explains that gravestones in coastal locations are prone to more severe degradation due to the salt content in the air; the table below summarises the expert's beliefs about θ_C and θ_I . An entry of ε indicates that the geologist thinks this range for θ is extremely unlikely.

θ	Coastal	Inland
$0 \rightarrow 0.1$	ε	a
$0.1 \rightarrow 0.2$	ε	$b = 0.4a$
$0.2 \rightarrow 0.3$	ε	b
$0.3 \rightarrow 0.4$	ε	ε
$0.4 \rightarrow 0.5$	ε	ε
$0.5 \rightarrow 0.6$	$d = 0.1c$	ε
$0.6 \rightarrow 0.7$	$2d$	ε
$0.7 \rightarrow 0.8$	$4d$	ε
$0.8 \rightarrow 0.9$	c	ε
$0.9 \rightarrow 1$	$8d$	ε

- (i) Given this information from the expert, which method of prior elicitation might be suitable to obtain prior distributions for θ_C and θ_I ?
- (ii) Use this method to obtain $\pi(\theta_C)$ and $\pi(\theta_I)$ (rounding the parameters of these priors to the nearest integer).
- (iii) Briefly explain how these prior distributions reflect the expert's beliefs about the differences between θ_C and θ_I .
- (d) Now combine the geologist's beliefs with the data to obtain the posterior distributions for θ_C and θ_I .
- (e) Use R to obtain plots of the prior and posterior for θ_C and θ_I , superimposing the prior and posterior on the same graphs in each case. Compare and contrast:
- (i) Prior beliefs with posterior beliefs, for both θ_C and θ_I ;
- (ii) Posterior beliefs about the rate of granite degradation at the coast and inland.
- (f) Obtain the 95% highest density interval for θ_I . [*Hint: You may use the results from Example 4.1 in the lecture notes to help you here.*]
- (g) Why is the 95% highest density interval for θ_C more awkward to find? Use the R function `optim` to obtain this interval, and include your R code in your solutions.
- (h) In light of the data, would you say there is a significant difference in the rate of granite degradation between coastal and inland locations? Explain your answer.

References

- Meierdin, T.C. (1981). Marble weathering rates: a transect of the United States. *Physical Geography*, **2**, pp. 1—18.
- Rahn, T. (1971). The weathering of tombstones and its relation to the topography of New England. *Journal of Geological Education*, **19**, pp. 112—118.