

MAS2317/3317: Problems sheet 3

Introduction to Bayesian Statistics

PROBLEMS SHEET 3 is to be used in this week's practical session (5pm, Thurs 6th Mar), and next week's problems class (4pm, Thurs 13th 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 25th March**.

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

- 1.* The *United States Geological Survey* (USGS) continuously tracks tropical depressions over the Atlantic Ocean – such systems can develop into hurricanes, and these storms can make landfall in eastern/southeastern states of the USA. You are a statistician working for the USGS, and in question 9 you have already used the *MATCH Uncertainty Elicitation Tool*, and R, to elicit a prior distribution for the rate of occurrence, θ , of three tropical depressions during the Atlantic hurricane season.

Let X be the number of tropical depressions during the Atlantic hurricane season; we assume a Poisson distribution for X , i.e. $X \sim P(\theta)$. In 2010, 2011 and 2012, the USGS observe 15, 6 and 10 tropical depressions, respectively. Combine the prior opinion elicited from the meteorologist with the data to obtain the posterior distribution for θ , and write a short report for the meteorologist summarising how beliefs about the rate of occurrence of tropical depressions have changed in light of the data.

2. Suppose we have a random sample from a Poisson distribution, that is $X_i|\theta \sim P(\theta)$, $i = 1, \dots, n$ (independent). Determine the posterior distribution assuming a vague prior for θ . What would be the resulting posterior if we had used a vague prior in question 1?
3. Read through case study 2, and then answer the following questions.
 - (a) Produce a list of benefits of working within the Bayesian framework over a standard frequentist analysis.
 - (b) What difficulties arose as a result of working within the Bayesian framework?
 - (c) How did we attempt to overcome these difficulties?
 - (d) The resulting posterior distribution for θ was *improper*. What does this mean?
 - (e) Name the simulation technique that can be used to obtain approximate draws from the posterior distribution when our prior is *non-conjugate*.

- 4.^P The number of admissions, per day, to a busy Accident & Emergency (A&E) department is assumed *Poisson* with rate θ .

This A&E department can cope with, at most, $\theta = 200$. The A&E consultant at this hospital tells us that he often sees between 80 and 100 patients admitted to A&E in any one day, although this can vary. Let $\Pr(80 < \theta < 100) = a$. We then elicit the following information from the consultant:

$$\begin{aligned}\Pr(40 < \theta < 60) &\approx 0.5a; \\ \Pr(60 < \theta < 80) &\approx 0.8a \quad (= b); \\ \Pr(100 < \theta < 120) &\approx b; \\ \Pr(120 < \theta < 140) &\approx 0.2a \quad (= c); \\ \Pr(140 < \theta < 160) &\approx c; \quad \text{and} \\ \Pr(160 < \theta < 180) &\approx 0.5c.\end{aligned}$$

In his time at this hospital, the consultant has never witnessed more than 180 casualties in any one day; only on very few occasions have there been less than 60 casualties.

- (a) Use the trial roulette method to obtain a suitable prior for θ , clearly demonstrating each step.
 - (b) Is your prior in (a) *conjugate*? Explain.
 - (c) Produce a feedback summary for the A&E consultant which gives:
 - the prior mode for θ ;
 - the prior standard deviation for θ ;
 - the 1% and 99% prior percentiles.
- 5.^P The number of casualties observed at a known traffic accident blackspot, in a single year, is assumed *Poisson* with rate λ . Using the bisection method, we are able to elicit the following quantities from a road safety expert:

$$Q_1(\lambda) = 2, \quad Q_2(\lambda) = 4 \quad \text{and} \quad Q_3(\lambda) = 8,$$

where Q_1 , Q_2 and Q_3 are the first, second and third quartiles (respectively). The expert also tells us that the mean casualty rate at this site is extremely unlikely to exceed 20 per year.

- (a) Use *MATCH* to elicit an appropriate prior distribution for θ , demonstrating this clearly.
 - (b) Given that we observed 6 casualties at this location in 2011, obtain the posterior distribution for λ .
- 6.^P A “degassing burst” is a seismic event that can occur just before a major earthquake. Let θ be the chance that a major earthquake will occur after a degassing burst has taken place. From discussions with an seismologist, we are able to elicit the following:

$$\Pr(\theta < 0.6) = 10^{-3}.$$

The seismologist also tells us that she thinks the most likely value for θ will be about 0.9. Use this information to justify using a *Beta*(22.6, 3.4) prior for θ , clearly explaining your method. [*Hint: You may use **R** to help you.*]

- 7.^P Let θ be the chance that a major earthquake will occur after a “degassing burst” has taken place (see previous question). Six of the last ten degassing bursts at locations along the edge of the Eurasian plate were followed by major earthquakes. Update the beliefs of the seismologist from the previous question with this information to form the posterior for θ , and compare prior and posterior means.
- 8.^P The proportion θ of defective items in a large shipment is unknown. However, experience suggests that a $Beta(2, 200)$ prior is appropriate.
- (a) Suppose that 100 items are selected at random from the shipment and that three are found to be defective. What is the posterior distribution of θ ?
 - (b) Suppose that another statistician, having observed the three defectives, said that his posterior distribution for θ was a beta distribution with mean $4/102$ and variance 0.0003658 . What prior distribution had that statistician used?

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.* The *United States Geological Survey* (USGS) continuously tracks tropical depressions over the Atlantic Ocean – such systems can develop into storm systems or hurricanes, and these storms can make landfall in eastern/southeastern states of the USA. You are a statistician working for the USGS, and you are interested in the rate of occurrence, θ , of these tropical depressions during the Atlantic hurricane season (June–November).

After a consultation with a meteorologist on the team, you find out that, in any year, she would probably expect to see between 16 and 20 tropical depressions over the Atlantic ocean. Let $\Pr(16 < \theta < 20) = a$.

You then ask her to specify how likely she thinks some other ranges for θ are, relative to the range she specified as most likely (given above). The results of your consultation are shown below.

$$\Pr(12 < \theta < 16) = 0.8a$$

$$\Pr(20 < \theta < 24) = 0.8a$$

$$\Pr(8 < \theta < 12) = 0.5a$$

$$\Pr(24 < \theta < 28) = 0.2a$$

$$\Pr(28 < \theta < 32) = 0.2a$$

$$\Pr(32 < \theta < 36) = 0.1a$$

The meteorologist tells you that she has never seen more than 40 tropical depressions in any one year; she also tells you that during a “quiet” hurricane season there might no tropical depressions observed at all. However, she assigns values of θ outside the range specified above negligible probabilities.

- In *MATCH*, enter the **Lower Limit** and **Upper Limit** as specified by the meteorologist. Then select the **Roulette Input Mode** to perform a trial roulette elicitation. A grid should appear; the x -axis should be split into equal-width *bins* in which you can place your *chips*.
- Place 10 chips in the 16→20 bin; this will represent the meteorologist’s probability a . Now use the information you elicited from the meteorologist to place the remaining chips.
- Now click **Fitting & Feedback** on the toolbar at the top of the screen. In your solutions, copy and complete the following, specifying the parameters of your fitted distribution to two decimal places.

Feedback: first fit

$$\theta \sim$$

$$\pi(\theta) = \quad , \quad \theta > 0$$

$$E(\theta) =$$

$$SD(\theta) =$$

- (d) Use the **Fitting & Feedback** box in *MATCH* to find the 1st and 99th percentiles of your prior distribution, and write these values down as probability statements in your solutions.
- (e) Phrase a question that you could ask the meteorologist using your answers to part(d), as part of the feedback stage necessary in any elicitation process.
- (f) In response to your question in part (e), the meteorologist would like you to refine the fitted prior for θ ; in particular, she believes the 99th percentile is too high, and that

$$\Pr(\theta > 32) \leq 0.01.$$

In response to the meteorologist, remove a single chip from bins $24 \rightarrow 28$ and $28 \rightarrow 32$. Also remove two chips from bin $20 \rightarrow 24$. You should see that the meteorologist's request has now been met. Now provide her with a new summary sheet, by copying and completing the following in your solutions:

Feedback after refinement: 2nd fit

$$\theta \sim$$

$$\pi(\theta) = \quad , \quad \theta > 0$$

$$E(\theta) =$$

$$SD(\theta) =$$

$$\Pr(\theta < \underline{\hspace{2cm}}) = 0.01$$

$$\Pr(\theta > \underline{\hspace{2cm}}) = 0.01.$$

(g) Now use R:

- (i) To produce a plot of $\pi(\theta)$ obtained from your second fit in *MATCH*, and include this in your solutions. *[Hint: you may find it useful to refer to questions 11–14 in problems sheet 2]*
- (ii) To confirm the feedback percentiles you obtained from your second fit in *MATCH*, and include any relevant R code in your solutions. *[Hint: you could use the command `pgamma` as you did in question 14 in problems sheet 2]*

10.* In this question we return to the scenario in question 9. Suppose the USGS are also interested in the probability, λ , that a tropical depression over the Atlantic develops into a major hurricane. After another consultation with the meteorologist, we are able to elicit that the mode of any distribution for λ should be around 0.7 and that $\Pr(\lambda < 0.3) = 0.001$.

- (a) From the following list, choose the most appropriate distribution for λ : $U(a, b)$, $N(a, b)$, $Be(a, b)$, $Ga(a, b)$.
- (b) For the model you selected in (a), write down an expression for the mode, and use the information elicited from the meteorologist to find an expression for a in terms of b .
- (c) Now write a function in R like that at the top of page 54 in the lecture notes, which evaluates the cumulative distribution function for your chosen distribution, minus 0.001, at b .
- (d) Now use the R command `uniroot`, as we do on page 54 of the lecture notes, to find b .
- (e) Now find a , and write down your elicited prior for λ .