Chapter 5

Exercises

Assignment questions

You will be told in class (and via email) which questions to answer for each of the assignments. Solutions to these questions should be placed in the homework submission letterbox in the foyer of the Maths & Stats General Office (Herschel Building 3rd floor) by no later than 4pm on the due date of the assignments.

Computing questions

This course will make some use of the statistical computing package R, and there will be occassional computing labs. You should bring this booklet to all lectures and computing labs. In the labs, you will be told which questions to work through. Some of these computing questions will be used for the assignments.

Non–assessed questions

We will work through most of the non–assessed questions in the problems classes; solutions to *all* questions will be made available through Blackboard.

Pencil-and-paper exercises

1. (a) Mark on a $0-1$ probability scale the approximate positions of the following events. Undertake a simple investigation if necessary. Explain the values you choose and state, in each case, whether you are using the frequency, classical or subjective interpretation of probability.

   (i) A student taking MAS2903 has a body mass index (BMI) of at least 22.5.

   (ii) Newcastle United will finish in a top 6 position in the Premier League this year.

   (iii) A randomly selected number between 10 and 99 (inclusive) is divisible by 7.

   (b) For each event in part (a), give a range which you believe contains the 'true' value with probability 0.95.

2. An insurance company classifies car drives as X, Y or Z. Experience indicates that the probability that a class X driver has at least one accident in any one year is 0.02, whilst the corresponding probabilities for class Y and class Z drivers are 0.04 and 0.10 respectively. They have also found that, of the drivers who apply to them for cover, 30% are class X, 60% are class Y and 10% are class Z.

   (i) Mr Smith is a new client and, within 12 months, he has an accident. What is the probability that he is a class Z risk?

   (ii) If Mrs Brown goes for n years without an accident, and we assume the number of accidents in different years to be independent, how large must n be before the company considers that she is more likely to belong to class X than to class Y?

3. Suppose that 1% of the population have a disease D. A diagnostic test S, designed to detect the disease, has the following accuracy:

$$\begin{aligned}\Pr(S \text{ positive}|D) &= 0.95 \qquad \text{and}\\ \Pr(S \text{ positive}|D^c) &= 0.05.\end{aligned}$$

   If 100 people were tested at random, how many would we expect to test positive and of those that do, about how many would we expect to have the disease?

4. On June 1st 2009, Air France flight 447 disappeared en-route to Paris Charles de Gaulle Airport about four hours after leaving Rio de Janeiro. Locating the aircraft's flight data recorder (the *black box*) was key to understanding the cause of one of the deadliest accidents in the history of Air France.

   After considering various oceanographic characteristics in the Atlantic at the time the flight disappeared, initial searches for the data recorder focussed on a large area to the north of the point at which it was believed the aircraft made impact with the ocean. In fact, it was believed that there was an 80% chance of finding the data recorder in this area, as opposed to the south of the impact zone. After almost two years of searching in the area mainly to the north of the point of impact, the elusive black box had still not been found.

   In April 2011 *Metron, Inc.* were hired to launch a Bayesian review of the search effort. The following probabilities of black box "identifiers" were estimated, based on the black box being in the areas defined as "North" and "South":

	Identifier			
	x_1	x_2	x_3	x_4
North	0.3	0.1	0.1	0.5
South	0.05	0.7	0.2	0.05

   These "identifiers" were:

   x_1 : "Fresh" debris found (since December 2010)

   x_2 : Signal detected from the black boxes

   x_3 : Both "fresh" debris found *and* signal from black boxes detected

   x_4 : Neither "fresh" debris found *nor* signals detected from the black boxes

   (a) Construct rules for differential diagnosis of the most plausible search area for the flight's data recorder, on the basis of the posterior assessments for each type of identifier.

   (b) According to your set of rules in part (a), what is the probability of 'misdiagnosis'?

   (c) In March 2011 a weak signal from the black box was detected. Quantify the shift in odds in favour of searching in the area south of the impact zone, having observed this identifier.

5. An experiment can result in four possible outcomes x_1, x_2, x_3 and x_4. The probability of each of these outcomes is affected by a parameter θ which can take one of three values θ_1, θ_2 and θ_3:

$p(x\|\theta)$	x_1	x_2	x_3	x_4
θ_1	0.1	0.2	0.3	0.4
θ_2	0.3	0.3	0.2	0.2
θ_3	0.5	0.1	0.1	0.3

The prior plausibilities attaching to θ_1, θ_2 and θ_3 are 0.1, 0.4 and 0.5. Find the posterior probability function for θ after observing x_1. Repeat the calculations for x_2, x_3 and x_4.

6. Extensive manuscripts of two scribes show that they differ on the relative frequencies with which they use two alternative vowel forms *e* and *oe*, as in the following table:

	e	*oe*
Scribe A	0.3	0.7
Scribe B	0.5	0.5

A new manuscript fragment, known to be by one of the scribes, uses the vowel form five times, three times as *e* and twice as *oe*. If an independent assessment of similar fragments favours Scribe A as author with probability 0.7, show that the evidence from the vowel form is sufficient to turn the odds in favour of Scribe B.

7. If $X_1, X_2, \ldots, X_n$ are independent random variables with probability function

$$f(x|\theta) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots,$$

show that $T = \sum_{i=1}^n X_i$ is sufficient for θ.

8. If $X_1, X_2, \ldots, X_n$ are independent $N(5, \sigma^2)$ random variables, show that

$$T = \sum_{i=1}^{n} (X_i - 5)^2$$

is sufficient for σ.

9. Hurricane–strength wind speeds (X miles per hour) for locations in the Gulf of Mexico are modelled using a distribution with probability density function

$$f(x|\sigma) = \frac{x}{\sigma^2}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}, \quad x > 0, \quad \sigma > 0.$$

During a hurricane, you collect wind speeds at n randomly chosen locations in the Gulf of Mexico, giving $x_1, x_2, \ldots, x_n$. Find the likelihood function for σ, and obtain a sufficient statistic for σ.

10. 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 $5.45\,g/cm^3 < \mu < 5.55\,g/cm^3$.

11. Suppose that a random sample $x_1, x_2, \ldots, x_n$ is obtained. Derive the posterior distributions for the following models:

(a) $f(x|\theta) = \theta^{x-1}(1-\theta)$, $x = 1, 2, \ldots$ with a $U(0,1)$ prior distribution for θ.

(b) $f(x|\theta) = \dfrac{e^{-\theta}\theta^x}{x!}$, $x = 0, 1, \ldots$ with a $Exp(1)$ prior distribution for θ.

(c) $f(x|\theta) = \dbinom{7}{x}\theta^x(1-\theta)^{7-x}$, $x = 0, 1, \ldots, 7$ with a $Beta(3,2)$ prior distribution for θ.

12. 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:

$$\pi(\theta) = \frac{100^6}{120}\theta^5 e^{-100\theta}. \tag{5.1}$$

(a) Identify fully the distribution in Equation (5.1).

(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.

13. Consider an experiment with a possibly loaded six-sided die. Let $\theta = \Pr(\text{rolling a six})$. 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 $E[\theta|x=3]$?

14. *[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

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

15. 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, $X \sim bin(5, \theta)$.

   (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 $Beta$(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, $\pi(\theta|x = 1)$.

16. 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.

17. 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.

18. A random sample of size n is to be taken from a $N(\theta, 2^2)$ distribution. The prior for θ is $N(b, 1/d)$.

    (a) 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?

    (b) 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.

19. 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?

20. 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*.

21. 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 42 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.

22. 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 21?

23. 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.]*

24. 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.

25. 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:

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

    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.

26. 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:

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

    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.

27. 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?

28. Suppose that the random variable Y follows a $Ga(a, b)$ distribution. Determine the probability density function for the random variable $X = 1/\sqrt{Y}$.

29. 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.

30. 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.

31. Use techniques from MAS2901 to obtain 95% frequentist confidence intervals for θ_{C} and θ_{I} in question 44.

32. Compare the frequentist and Bayesian confidence intervals for θ_{C} and θ_{I} as obtained in questions 31 and 44 (respectively). Comment on any numerical differences, as well as differences in interpretation.

33. 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 $a < b < c < d < e < f$.

34. 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 $m \to \infty$?

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

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

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

35. 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 $W \sim NegBin(r, p)$?

Computing exercises

Open R by clicking on `Start` → `All Programs` → `Statistical Software` → `RStudio`. You will 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/mas2903/`

Some questions will also make use of the *MATCH Uncertainty Elicitation Tool*, available at:

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

36. 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.

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

Assuming that $\theta \sim U(0,1)$, we found that $\theta|x = 1 \sim Be(2,5)$. 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 $Be(2,5)$ posterior for θ, we apply the R function `dbeta` to `x`:

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

This evaluates the $Be(2,5)$ 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.

37. 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

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

He assumes a beta prior for θ: $\theta \sim Be(2.5,12)$, which results in $\theta|x=4 \sim Be(6.5,13)$. Following carefully what you did in question 36, 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.

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

$$X|\theta \sim Exp(\theta).$$

An expert tells us that $\theta \sim Ga(10,4000)$. We then combine our observations with the expert's prior beliefs to obtain $\theta|\boldsymbol{x} \sim Ga(30,13633)$. 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 questions 36 and 37 to produce a plot comparing your prior and posterior distributions for the rate of earthquakes θ. 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.

39. 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 36,

(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.

Questions 40 and 41 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`.

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

    (a) Download the reference data from the MAS2903 course webpage by executing the following code in R:

    ```
    > ref = read.table("http://www.mas.ncl.ac.uk/~nlf8/teaching/
      + mas2903/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)))
```

41. This question will use your solutions to question 40 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 40(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 40(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 14 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 14(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 36–38, 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) in the case study, or question 14(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.

42. 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.

$$\begin{aligned}
\Pr(12 < \theta < 16) &= 0.8a\\
\Pr(20 < \theta < 24) &= 0.8a\\
\Pr(8 < \theta < 12) &= 0.5a\\
\Pr(24 < \theta < 28) &= 0.2a\\
\Pr(28 < \theta < 32) &= 0.2a\\
\Pr(32 < \theta < 36) &= 0.1a
\end{aligned}$$

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.

(a) 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*.

(b) 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.

(c) 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

$\theta \sim$

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

$E(\theta) =$

$SD(\theta) =$

(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 (see overleaf):

Feedback after refinement: 2nd fit

$\theta \sim$

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

$E(\theta) =$

$SD(\theta) =$

$\Pr(\theta <$ ________ $) = 0.01$

$\Pr(\theta >$ ________ $) = 0.01.$

(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 36–38]*

(ii) To confirm the feedback percentiles you obtained from your second fit in *MATCH*, and include any relevant R code in your solutions.

43. In this question we return to the scenario in question 21. 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.

44. 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 Meierdin (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.

Tennessee
North Carolina
Georgia
Chattanooga
South Carolina
Atlanta
Alabama
Macon
Columbus
Savannah
Albany
ATLANTIC OCEAN
Florida
Tallahassee
Jacksonville

(a) Download the `granite-coastal.txt` dataset into R by typing:

```
>coastal=read.table('http://www.mas.ncl.ac.uk/~nlf8/teaching/
+  mas2903/problems/granite-coastal.txt')
```

Make sure you do something similar for the `granite-inland.txt` dataset. You can the 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
$0 \to 0.1$	ε	a
$0.1 \to 0.2$	ε	$b = 0.4a$
$0.2 \to 0.3$	ε	b
$0.3 \to 0.4$	ε	ε
$0.4 \to 0.5$	ε	ε
$0.5 \to 0.6$	$d = 0.1c$	ε
$0.6 \to 0.7$	$2d$	ε
$0.7 \to 0.8$	$4d$	ε
$0.8 \to 0.9$	c	ε
$0.9 \to 1$	$8d$	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 $\pi(\theta_{\text{C}})$ and $\pi(\theta_{\text{I}})$ (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. *Journnal of Geological Education*, **19**, pp, 112—118.