Bayesian course - problem set 1 (lectures 1& 2)

November 2, 2016

1 Dodgy coins

Suppose there are three coins in a bag. The first coin is biased towards heads, with a 75% probability of a heads occurring if the coin is flipped; the second is fair – so a 50% chance of heads occurring; the third coin is biased towards tails, and so has a 25% probability of coming up heads. Assume that it is impossible to identify which coin is which from looking/touching the coins.

Problem 1.1 *Suppose we put our hand into the bag, and pull out a coin. We then flip the coin and find it comes up heads. Let the random variable* $C = \{1, 2, 3\}$ *denote the identity of the coin, where the probability of heads is* $(0.75, 0.50, 0.25)$ *respectively. Obtain the likelihood by using the equivalence relation (that a likelihood of a parameter value given data is equal to the probability of data given a parameter value), and show that the sum of the likelihood over all parameter values is 1.5.*

Problem 1.2 *What is the maximum likelihood estimate of the coin's identity?*

Problem 1.3 *Use Bayes' rule to prove that,*

$$Pr(C = c|X = H) \propto Pr(X = H|C = c) \times Pr(C = c) \tag{1}$$

where $c = 1, 2, 3$.

Problem 1.4 *Assume that since we cannot visually detect the coin's identity we use a uniform prior* $Pr(C = c) = \frac{1}{3}$ *for* $c = 1, 2, 3$*. Use this to complete Table 1 and determine the (marginal) probability of the data.*

Problem 1.5 *By completing Table 1 find the posterior, and confirm that it is a proper probability distribution.*

Problem 1.6 *Now assume that we throw the same coin twice, and find that it lands heads up on both occasions. By using a Table of form similar to Table 1 (known as a "Bayes Box"), or otherwise, determine the new posterior distribution.*

parameter C	likelihood $Pr(X = H \| C = c)$	prior $Pr(C = c)$	likelihood × prior $Pr(X = H \| C = c) \times Pr(C = c)$	posterior $Pr(C = c \| X = H)$
1				
2				
3				
			$Pr(X = H) =$	

Table 1: A Bayes box for the coins example.

Problem 1.7 *Now assume that you believe that the tails-biased coin is much more likely to be drawn from the bag, and thus specify a prior:* $Pr(C = 1) = 1/20$, $Pr(C = 2) = 5/20$ *and* $Pr(C = 3) = 14/20$. *What is the posterior probability that* $C = 1$ *now?*

Problem 1.8 *Continuing on from the previous example calculate the posterior mean, MAP (maximum a posteriori), and Maximum Likelihood estimates.*

Problem 1.9 *For the case when we flip the coin once and obtain* $X = H$*, using the uniform prior on* C*, determine the posterior predictive distribution for a new coin flip with result* $\tilde{X}$*, using the below expression,*

$$Pr(\tilde{X}|X = H) = \sum_{C=1}^{3} Pr(\tilde{X}|C) \times Pr(C|X = H) \qquad (2)$$

Remembering that your result should be a proper probability distribution.

Problem 1.10 *(Optional) Justify the use of the expression in the previous question.*

2 The epidemiology of Lyme disease

Lyme disease is a tick-borne infectious disease spread by bacteria of species *Borrelia*, which are transmitted to ticks when they feed on animal hosts. Whilst fairly common in the US, this disease has recently begun to spread throughout Europe.

Imagine you are researching the occurrence of Lyme disease in the UK. As such, you begin by collecting samples of 10 ticks from fields and grasslands around Oxford, and counting the occurrence of the *Borrelia* bacteria.

Problem 2.1 *We start by assuming that the occurrence of Borrelia bacteria in one tick is independent of that in other ticks. In this case, why is it reasonable to assume a binomial likelihood?*

Problem 2.2 *Suppose the number of Borrelia-positive ticks within each sample* i *is given by the random variable* X_i*, and that the underlying prevalence (amongst ticks) of this disease is* θ*. Using the equivalence principle write down the likelihood for sample* i*.*

Problem 2.3 *Suppose that you collect one sample and find $X_1 = 1$. Graph the likelihood here and hence – by eye – determine the maximum likelihood estimate of θ.*

Problem 2.4 *By numerical integration show that the area under the likelihood curve is about 0.09. Comment on this result.*

Hint: use R's "integrate" function.

Problem 2.5 *Assuming that $\theta = 10\%$, graph the probability distribution (also known as the sampling distribution). Show that this distribution – in contrast to the likelihood – is a proper probability distribution.*

Note: this distribution, where we vary the data not the parameter, is a discrete distribution unlike the continuous likelihood.

Problem 2.6 *(Optional) Now assume that we do not know θ. Use calculus to show that the Maximum Likelihood estimator of the parameter, for a single sample of size 10 where we found X ticks with the disease is given by,*

$$\hat{\theta} = \frac{X}{10} \tag{3}$$

Hint: maximise the log-likelihood rather than the likelihood.

Problem 2.7 *A colleague mentions that a reasonable prior to use for θ is a $Beta(a, b)$ distribution. Graph this for $a = 1$ and $b = 1$.*

Problem 2.8 *How does this distribution change as you vary a and b?*

Hint: either in R or Google, lookup its mean.

Problem 2.9 *Prove that a $Beta(a, b)$ prior is conjugate to the Binomial likelihood, showing that the posterior distribution is given by a $Beta(X+a, 10-X+b)$ distribution.*

Problem 2.10 *Graph the posterior for $a = 1$ and $b = 1$. How does the posterior distribution vary as you change the mean of the Beta prior? (In both cases assume that $X = 1$.)*

Problem 2.11 *You now collect a larger dataset (encompassing the previous one) that has a sample size of 100 ticks in total; of which you find 7 carry Borrelia. Find the new posterior using the conjugate prior rules for a $Beta(1, 1)$ prior and binomial likelihood.*

Hint: if you did not work out these rules previously, Google "conjugate priors".

Problem 2.12 *You collect a second dataset of 100 ticks; this time finding that 4 carry the disease. Find the new posterior (across both datasets) using the conjugate prior rules for a $Beta(1, 1)$ prior and binomial likelihood. How does it compare to the previous one?*

Problem 2.13 *Now we are going to use sampling to estimate the posterior predictive distribution for a sample size of 100, using the posterior distribution obtained from the entire sample of 200 ticks (11 of which were disease positive). To do this we will first sample a random value of* θ *from the posterior: so* $\theta_i \sim p(\theta|X)$*. We then sample a random value of the data* X *by sampling from the binomial sampling distribution* $X_i \sim binomial(100, \theta_i)$*. We repeat this process a large number of times to obtain samples from this distribution. Follow the previous rules to produce 10,000 samples from the posterior predictive distribution, which we then graph using a histogram.*

Problem 2.14 *Does our model adequately fit the data?*

Problem 2.15 *Indicate whether you expect this model to hold across future sampling efforts.*

Problem 2.16 *If we assume a uniform prior on* θ *- the probability that a randomly sampled tick carries Lyme disease - what is the shape of the prior for* θ^2*? (This is the probability that 2/2 ticks carry Lyme disease.)*

Hint: either do this using Jacobians (hard-ish), or by sampling (easy-ish).

3 GDP vs infant mortality

The data in "prob1_gdpInfantMortality.csv" contain the GDP per capita (in real terms) and infant mortality across a large sample of countries in 1998.

Problem 3.1 *A simple model is fit to the data of the form:*

$$M_i \sim N(\alpha + \beta GDP_i, \sigma) \quad (4)$$

Fit this model to the data using a Frequentist approach. How well does the model fit the data?

Problem 3.2 *An alternative model is:*

$$log(M_i) \sim N(\alpha + \beta log(GDP)_i, \sigma) \quad (5)$$

Fit this model to the data using a Frequentist approach. Which model do you prefer, and why?

Problem 3.3 *Construct 80% confidence intervals for* (α, β)*.*

Problem 3.4 *I have fit the log-log model to the data using MCMC. Samples from the posterior for* (α, β, σ) *are contained within the file "prob1_posteriorsGdpInfantMortality.csv". Using this data find the 80% credible intervals for all parameters (assuming these intervals to be symmetric about the median). How do these compare with the confidence intervals calculated above for* (α, β)*? How does the point estimate of* σ *from the Frequentist approach above compare?*

Problem 3.5 *I used the following priors for the three parameters:*

$$\alpha \sim N(0,10)$$
$$\beta \sim N(0,10)$$
$$\sigma \sim N(0,5), \textit{where } \sigma \geq 0$$

Explain any similarity between the confidence and credible intervals in this case.

Problem 3.6 *How are the estimates of parameters* (α, β, σ) *correlated? Why?*

Problem 3.7 *Generate samples from the prior predictive distribution. How does the min and max of the prior predictive distribution compare with the actual data?*

Problem 3.8 *Generate samples from the posterior predictive distribution, and compare these with the actual data. How well does the model fit the data?*

4 Googling

Suppose you are chosen, for your knowledge of Bayesian statistics, to work at Google as a search traffic analyst. Based on historical data you have the data shown in table 2 for the actual word searched, and the starting string (the first three letters typed in a search). It is your job to help make the search engines faster, by reducing the search-space for the machines to lookup each time a person types.

	Barack Obama	**Baby clothes**	**Bayes**
Bar	50%	30%	30%
Bab	30%	60%	30%
Bay	20%	10%	40%

Table 2: The columns give the historic breakdown of the search traffic for three topics: Barack Obama, Baby clothes, and Bayes; by the first three letters of the user's search.

Problem 4.1 *Find the minimum-coverage confidence intervals of topics that are at least at 70%.*

Problem 4.2 *Find most narrow credible intervals for topics that are at least at 70%.*

Now we suppose that your boss gives you the historic search information shown in table 3. Further, you are told that it is most important to correctly suggest the actual topic as one of the first auto-complete options, *irrespective* of the topic searched.

	Barack Obama	Baby clothes	Bayes
Search volume	60%	30%	10%

Table 3: The historic search traffic broken down by topic.

Problem 4.3 *Do you prefer confidence intervals or credible intervals in this circumstance?*

Problem 4.4 *Now assume that it is most important to pick the correct actual word across all potential sets of three letters. Which interval do you prefer now?*

5 Epilepsy

In the data file "prob1_epil.csv" there is a count of seizures for 112 patients with epilepsy who took part in a study [1]. Assume a) the underlying rate of seizures is the same across all patients, and b) the event of a seizure occurring is independent of any other seizures occurring.

Problem 5.1 *Under these assumptions what model might be appropriate for this data?*

Problem 5.2 *Write down the likelihood for the data.*

Problem 5.3 *Show that a Gamma prior is conjugate to this likelihood.*

Problem 5.4 *Assuming a* $\Gamma(4, 0.25)$ *(with a parameterisation such that it has mean of 16) prior. Find the posterior distribution, and graph it.*

Problem 5.5 *Find/look-up the posterior predictive distribution, and graph it.*

Problem 5.6 *Comment on the suitability of the model to the data.*

6 Bayesian neurosurgery

Suppose that you are a neurosurgeon and have been given the unenviable task of finding the position of a tumour within a patient's brain, and cutting it out. Along two dimensions - vertical height and left-right axis - the tumour's position is known to a high degree of confidence. However, along the remaining axis - front-back - the position is uncertain, and cannot be ascertained without surgery. However, a team of brilliant statisticians has already done most of the job for you, and has generated samples from the posterior for the tumour's location along this axis, and is given by the data contained within the data file "prob1_brainData.csv".

Suppose that the more brain that is cut, the more the patient is at risk of losing cognitive functions. Additionally, suppose that there is uncertainty over the amount of damage done to the patient during surgery. As such, three different surgeons have differing views on the damage caused:

1. *Surgeon 1:* Damage varies quadratically with the distance the surgery starts away from the tumour.
2. *Surgeon 2:* There is no damage if tissue cut is within 0.0001mm of the tumour; for cuts further away there is a fixed damage.
3. *Surgeon 3:* Damage varies linearly with the absolute distance the surgery starts away from the tumour. (Hard - use fundamental theorem of Calculus for the question below.)

Problem 6.1 *Under each of the three regimes above, find the best position along this axis to cut.*

Problem 6.2 *Which of the above loss functions do you think is most appropriate, and why?*

Problem 6.3 *Which loss function might you choose to be most robust to any situation?*

Problem 6.4 *Following from the previous point, which type of posterior point measure might be most widely applicable?*

Problem 6.5 *Using the data estimate the loss under the three different regimes assuming that the true loss* $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^3$.

7 Blood doping in cyclists

Suppose as a benign omniscient observer, we tally up the historical cases where professional cyclists used/didn't-use blood doping, and either won or lost a particular race. This results in the probability distribution shown in table 4.

Problem 7.1 *What is the probability that a professional cyclist wins a race?*

Problem 7.2 *What is the probability that a cyclist wins a race given that they have cheated?*

Problem 7.3 *What is the probability that a cyclist is cheating given that they win?*

	Lost	**Won**
Clean	0.70	0.05
Doping	0.15	0.10

Table 4: The historical probabilities of behaviour and outcome for professional cyclists.

Now suppose that drug testing officials have a test that can accurately identify a blood-doper 90% of the time. However, it incorrectly indicates a positive for clean athletes 5% of the time.

Should the officials test all the athletes or only the winners, for the cases when:

Problem 7.4 *They care only about the proportion of people correctly identified as dopers.*

Problem 7.5 *They care five times as much about the number of people who are falsely identified as they do about the number of people who are correctly identified as dopers.*

Problem 7.6 *What factor would make the officials choose the other group? (By factor, we mean the number 5 in the previous problem.)*

8 Light speed

In "prob1_newcomb.csv" there are Newcomb's (1882) measurements of the passage time (in millionths of seconds) it took light to travel from his lab, to a mirror on the Washington Monument, and back again. The distance of the path travelled is ca.7.4 km. The primary goal of this experiment is to determine the speed of light, and to quantify the uncertainty of the measurement. We assume there are a multitude of factors that additively result in measurement error for the passage time.

Problem 8.1 *Why might a normal distribution be appropriate here?*

Problem 8.2 *Write down the likelihood for all the data.*

Problem 8.3 *Derive the maximum likelihood estimators of all parameters.*

Problem 8.4 *Based on the likelihood function what functional form for the prior* $p(\mu, \sigma^2)$ *would make it conjugate?*

Problem 8.5 *Assuming a decomposition of the prior* $p(\mu, \sigma^2) = p(\sigma^2) \times p(\mu|\sigma^2)$*, what priors might we use?*

Problem 8.6 *(Difficult) Using these priors, find the parameters of the posterior distribution.*

Problem 8.7 *Comment on the suitability of the model to the data. (You can use the ML estimates here, or if you're feeling ambitious, the full posterior predictive distribution.)*

9 Game theory

A game show presents contestants with four doors: behind one of the doors is a car worth $1000; behind another is a forfeit whereby the contestant must pay $1000 out of their winnings thus far on the show. Behind the other two doors there is nothing.

The order of the game is as below:

1. The contestant chooses one of four doors.
2. The game show host opens another door; revealing that there is 'nothing' behind it.
3. The contestant is given the option of changing their choice to one of the two remaining unopened doors.
4. The contestant's final choice door is opened, either to their delight (a car!), dismay (a penalty), or indifference (nothing).

Assuming that:

- The contestant wants to maximise their expected wealth.
- The contestant is risk averse.

What is the optimal strategy for the contestant?

10 Breast cancer revisited

Suppose that the prevalence of breast cancer for a randomly-chosen 40 year old woman in the UK population is about 1%. Further suppose that mammography has a relatively high sensitivity to breast cancer, where in 90% of cases the test shows a positive result if the individual has the disease. However the test also has a rate of false positives of 8%.

Problem 10.1 *Show that the probability that a woman tests positive is about 9%.*

Problem 10.2 *A woman tests positive for breast cancer. What is the probability she has the disease?*

Problem 10.3 *Draw a graph of the probability of having a disease given a positive test, as a function of a. the test specificity (true positive), b. the false positive rate, and c. the disease prevalence. Graph these variables for rare (1% prevalence) and common (10% prevalence) diseases. What does this tell you about the importance of the various elements of medical tests?*

Problem 10.4 *Assume the result of a mammography is independent when retesting an individual (probably a terrible assumption!). How many tests (assume a positive result in each) would need to be undertaken to ensure that the individual has a 99% probability that they have cancer?*

Problem 10.5 *Now we make the more realistic assumption that the probability of testing positive in the* n*th trial depends on whether positive tests were achieved in the* $(n-1)$*th trials, for both individuals with cancer and those without. For a cancer status* $\kappa \in \{C, NC\}$*,*

$$p(n+|(n-1)+,\kappa) = 1 - (1 - p(+|\kappa))e^{-(n-1)\epsilon} \tag{6}$$

where $n+$ *indicates testing positive in the* n*th trial,* $p(+|\kappa)$*, and* $\epsilon \geq 0$ *determines the persistence in test results. Assume that* $p(+|C) = 0.9$ *and* $p(+|NC) = 0.08$*. For* $\epsilon = 0.15$ *show that we now need at least 17 positive test results to conclude that a patient has cancer.*

References

[1] Peter F Thall and Stephen C Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, pages 657–671, 1990.