Chapter 7

The posterior - the goal of Bayesian inference

7.1 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 7.1 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 7.1: 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 7.1.1. Find the minimum-coverage confidence intervals of topics that are at least at 70%.

In both cases (here and the next question) we are looking for sets of the actual words. Frequentists assume that the data we receive (each set of three letters) is a sample from an infinity of such experiments. As such they design their intervals such that at least 70% of such intervals contain the true word searched across all the potential data samples we could receive. This means that here we want to choose sets such that regardless of the three letters typed we get a coverage of at least 70% for each of the columns. These are shown in Table 7.2.

Problem 7.1.2. Find most narrow credible intervals for topics that are at least at 70%.

Bayesians condition of the data we *actually* receive, and derive intervals based on this information.

	Barack Obama	**Baby clothes**	**Bayes**	**Credibility**
Bar	[50%]	30%	30%	45%
Bab	30%	[60%	30%]	75%
Bay	[20%	10%	40%]	100%
Coverage	70%	70%	70%	

Table 7.2: ≥70% confidence intervals.

This means we need to consider the individual row sums; each time making an interval that exceeds at least 70% of that row. The answer for this question is shown in Table 7.3.

	Barack Obama	**Baby clothes**	**Bayes**	**Credibility**
Bar	[50%	30%]	30%	73%
Bab	30%	[60%	30%]	75%
Bay	20%	[10%	40%]	71%
Coverage	50%	100%	70%	

Table 7.3: ≥70% credible intervals.

Now we suppose that your boss gives you the historic search information shown in Table 7.4. 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.

Problem 7.1.3. Do you prefer confidence intervals or credible intervals in this circumstance?

Here all we need to do is work out the total losses under the confidence and credible intervals. For both cases this means we need to work out the expected loss for each of the actual words being searched, using the volumes given in Table 7.4. This is easily done using the coverages at the bottom of each of tables 7.2 and 7.3.

For the confidence intervals we thus get an expected loss:

$$loss = 0.6 \times (1 - 0.7) + 0.3 \times (1 - 0.7) + 0.1 \times (1 - 0.7) = 0.3 \tag{7.1}$$

And for the credible intervals:

$$loss = 0.6 \times (1 - 0.5) + 0.3 \times (1 - 1) + 0.1 \times (1 - 0.7) = 0.33 \tag{7.2}$$

So in this circumstance we prefer the confidence intervals.

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

Table 7.4: The historic search traffic broken down by topic.

Problem 7.1.4. 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?

Now we need to find the loss for each possible three letter search. This requires that we first of all calculate the historic search volumes for these letters using Tables 7.1 and 7.4. Specifically you take the matrix product of the two, yielding a percentage of historical searches of (42%, 39%, 19%) for (bar, bab bay). You then use the credible levels for each of the rows from the confidence and credible interval tables respectively to weight the losses.

For confidence intervals:

$$loss = 0.42 \times (1 - 0.45) + 0.39 \times (1 - 0.75) + 0.19 \times (1 - 1) = 0.33 \tag{7.3}$$

And for credible intervals:

$$loss = 0.42 \times (1 - 0.73) + 0.39 \times (1 - 0.75) + 0.19 \times (1 - 0.71) = 0.27 \tag{7.4}$$

So in this case we prefer the credible intervals.

7.2 GDP versus infant mortality

The data in `posterior_gdpInfantMortality.csv` contains the GDP per capita (in real terms) and infant mortality across a large sample of countries in 1998.

Problem 7.2.1. A simple model is fit to the data of the form:

$$M_i \sim \mathcal{N}(\alpha + \beta GDP_i, \sigma) \tag{7.5}$$

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

Graph the data first! I have perhaps been a bit misleading here asking the student to fit the model *before* graphing the data. However, this is sort of the point. You should never - blindly - fit a model to data. If you graph the data, you see that a linear model is not well suited to the data at all; a power law is better-suited. You can also use AIC/BIC/adjusted-R^2 etc. to compare between these models, but really the graphical explanation is the gold standard.

Problem 7.2.2. An alternative model is:

$$log(M_i) \sim \mathcal{N}(\alpha + \beta log(GDP)_i, \sigma) \tag{7.6}$$

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

See above - this model is much better suited!

Problem 7.2.3. Construct 80% confidence intervals for (α, β) for the log-log model.

Take the point estimates of the parameters and add the relevant critical values of a *standardised* Student T distribution with $n-2$ degrees of freedom (here the population standard deviation is unknown so we need to use a T rather than a normal) multiplied by the parameter's standard error. The 10% critical value (we need 10% values because we are using a two-sided test) of a T with the relevant degrees of freedom is 1.28. We therefore obtain the following 80% confidence intervals:

$$6.8 \leq \alpha \leq 7.3$$
$$-0.53 \leq \beta \leq -0.46$$

Problem 7.2.4. We have fit the log-log model to the data using MCMC. Samples from the posterior for (α, β, σ) are contained within the file `posterior_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?

Using the "quantile" function these can be estimated:

$$6.8 \leq \alpha \leq 7.3$$
$$-0.53 \leq \beta \leq -0.46$$
$$0.56 \leq \sigma \leq 0.64$$

The first two are, to the accuracy shown, indistinguishable from the Frequentist estimates. The point estimate for σ for the two approaches is:

$$\hat{\sigma}_F = 0.59$$
$$\hat{\sigma}_B = 0.60$$

where I have used the posterior mean for the Bayesian estimate.

Problem 7.2.5. The following priors were used for the three parameters:

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

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

Here the priors are very diffuse over the range of possible range of the parameters. To a (rough) approximation this is equivalent to a flat prior on the parameters. This means from Bayes' rule we have (approximately):

$$p(\theta|X) \propto p(X|\theta) \tag{7.7}$$

Therefore the confidence and credible intervals are going to be largely similar here.

Problem 7.2.6. How are the estimates of parameters (α, β, σ) correlated? Why?

α and β are negatively correlated. This is because we want a line that goes through the centre of the data: if the y intercept increases then the slope must decrease.

Problem 7.2.7. Generate samples from the prior predictive distribution. How does the min and max of the prior predictive distribution compare with the actual data?

The prior predictive distributions show about two orders of magnitude greater variation in data compared to the actual data.

Problem 7.2.8. Generate samples from the posterior predictive distribution, and compare these with the actual data. How well does the model fit the data?

There are a number of ways to compare the model vs the data here. I have just used the min and max as a point of comparison. What we see with these is that the minimum is captured well by the model, but the max isn't. In particular the variation seen in fitted model is *greater* than that in the data. This is because at low values of GDP there could be a deviation from the log-log model (or it's just due to sampling variation, of course).

7.3 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 `posterior_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 this part of the question.)

Problem 7.3.1. Under each of the three regimes above, find the best position along this axis to cut.

Surgeon 1:

A **quadratic** loss function has the form: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, where $\hat{\theta}$ is the point estimate of the parameter, and θ is the actual value. We can find the expected loss:

$$\begin{aligned}
\mathrm{E(L)} &= \int (\hat{\theta} - \theta)^2 p(\theta|x)\mathrm{d}\theta \\
&= \hat{\theta}^2 \int p(\theta|x)\mathrm{d}\theta - 2\hat{\theta} \int \theta p(\theta|x)\mathrm{d}\theta + \int \theta^2 p(\theta|x)\mathrm{d}\theta \\
&= \hat{\theta}^2 - 2\hat{\theta} < \theta|x > + < \theta|x >^2
\end{aligned}$$

which is minimised if $\hat{\theta} =< \theta|x >$. From the data the mean is 6.1.

Surgeon 2:

A **binary** loss function has the form: $L(\hat{\theta}, \theta) = 1 - \delta_{\theta=\hat{\theta}}$, where the δ is a Dirac delta at $\theta = \hat{\theta}$. Finding the expected loss:

$$\begin{aligned}
\mathbb{E}(L) &= \int (1 - \delta_{\theta=\hat{\theta}}) p(\theta|x)\mathrm{d}\theta \\
&= 1 - p(\theta = \hat{\theta}|x)
\end{aligned}$$

which is minimised when $\hat{\theta} = \arg\max_\theta p(\theta|x)$; in other words the MAP estimator, which from a histogram of the data is at around 4.5.

Surgeon 3:

A **linear** loss is of the form: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, resulting in an expected loss of:

$$\begin{aligned}
\mathbb{E}(L) &= \int |\hat{\theta} - \theta| p(\theta|x)\mathrm{d}\theta \\
&= \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta|x)\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta|x)\mathrm{d}\theta \\
&= \hat{\theta} \left(\int_{-\infty}^{\hat{\theta}} p(\theta|x)\mathrm{d}\theta - \int_{\hat{\theta}}^{\infty} p(\theta|x)\mathrm{d}\theta \right) - \int_{-\infty}^{\hat{\theta}} \theta p(\theta|x)\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} \theta p(\theta|x)\mathrm{d}\theta
\end{aligned}$$

Differentiating the above we get (using Feynman's differentiation under the equals sign):

$$\begin{aligned}\frac{\mathrm{d}L}{\mathrm{d}\hat{\theta}} &= 2\hat{\theta}p(\hat{\theta}|x) + \int_{-\infty}^{\hat{\theta}} p(\theta|x)\mathrm{d}\theta - \int_{\hat{\theta}}^{\infty} p(\theta|x)\mathrm{d}\theta - 2\hat{\theta}p(\hat{\theta}|x) \\ &= \int_{-\infty}^{\hat{\theta}} p(\theta|x)\mathrm{d}\theta - \int_{\hat{\theta}}^{\infty} p(\theta|x)\mathrm{d}\theta = 0\end{aligned}$$

which is true only when $\int_{-\infty}^{\hat{\theta}} p(\theta|x)\mathrm{d}\theta = 0.5$; in other words $\hat{\theta}$ is the median. For the data the median is 5.19.

Problem 7.3.2. Which of the above loss functions do you think is most appropriate, and why?

I would say that either the quadratic or linear loss functions are more appropriate than the binary loss. One could argue that losses to brain function are likely the proportional to the *volume* of tissue lost. Since a quadratic loss is closer to this (a cubic loss), then we might suppose that this is most appropriate.

Problem 7.3.3. Which loss function might you choose to be most robust to any situation?

Either the quadratic or linear losses since most problems exhibit a loss that increases in the distance away from the true value.

Problem 7.3.4. Following from the previous point, which type of posterior point measure might be most widely applicable?

Either the posterior mean or median.

Problem 7.3.5. Using the data estimate the loss under the three different regimes assuming that the true loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^3$.

The mean loss from the mean is 151; from the mode it is 245; from the median it is 175.

Bibliography