Bayesian book: problem sets

Ben Lambert

August 3, 2015

Chapter 1

Probability

1.1 The Bayesian game show

This question hinges on deriving the distribution of outcomes under either staying put, or changing, after the game show host has opened the empty door. This can be answered through application of Bayes' rule, but I prefer here to describe more intuitively what is happening.

Imagine we are considering repeating the show a number of times. We can imagine three possibilities for the initial choice of door:

- $\frac{1}{4}$ of the time, the door hides the car.
- $\frac{1}{2}$ of the time, the door hides a null.
- $\frac{1}{4}$ of the time, the door hides the penalty.

Considering now each of these in turn:

If the door contains the car, then the other three doors are two nulls, and the penalty. The game show host opens one of the nulls, meaning that only one null and the penalty remain. In this circumstance, if you stay put, you definitely obtain the car. If you change, you get a null with probability $\frac{1}{2}$ and similarly for the penalty.

If the door contains a null, then the other three doors are one null, one penalty and one car. When the host opens the remaining null, then the other two doors are the car, and the penalty. This is the key step. By staying put, you gain/lose nothing, whereas if you change you face risk; you get the car with probability $\frac{1}{2}$ and similarly for the penalty. Both of these choices have the same expected payoff, but the latter increases risk.

Finally, if the door contains the penalty, then the other three doors are two nulls, and the car. The game show host opens one of the nulls, meaning that only one null and the car remain. In this circumstance, if you stay put, you definitely obtain the penalty. If you change, you get a null with probability $\frac{1}{2}$ and similarly for the car.

We can now write down probability distributions for the outcomes given each possible action. For staying put, the probabilities are what they were if you hadn't received any new information, in other words $(p(car), p(null), p(penalty) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. Whereas if you change:

$$\begin{aligned} p(car) &= \frac{1}{4} \times 0 + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{3}{8} \\ p(null) &= \frac{1}{4}\frac{1}{2} + \frac{1}{2} \times \times 0 + \frac{1}{4} \times \frac{1}{2} = \frac{2}{8} \\ p(penalty) &= \frac{1}{4} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times 0 = \frac{3}{8} \end{aligned} \tag{1.1}$$

So in summary, we get face $(p(car), p(null), p(penalty) = (\frac{3}{8}, \frac{2}{8}, \frac{3}{8})$ by changing.

Both of these two outcomes face the same expected return of \$0:

$$\begin{aligned} \mathrm{E}[return|stay] &= \frac{1}{4} \times \$1,000 + \frac{1}{2} \times \$0 + \frac{1}{4} \times -\$1,000 \\ &= \$0 \end{aligned} \tag{1.2}$$

$$\begin{aligned} \mathrm{E}[return|change] &= \frac{3}{8} \times \$1,000 + \frac{2}{8} \times \$0 + \frac{3}{8} \times -\$1,000 \\ &= \$0 \end{aligned} \tag{1.3}$$

However, what is different is the variance in return from each decision. The variance in return is greater by changing, than it is by staying, since the latter has more weight on the risky outcomes. Obviously, these can be calculated explicitly using the same methodology as above.

If the individual is risk-averse, then he prefers the less risky outcome of *staying put*, given that they both have the same return.

If you still aren't convinced by this reasoning, see the Mathematica file XXX which compares these two strategies over time. If the reward value is increased, then changing becomes relatively more attractive, and vice versa if the cost is increased. This makes sense since we want to choose a riskier option when the reward is higher, and a less riskier option when the cost increases. When we increase the null value we want to choose the option which returns this value more frequently; staying put.

It is also possible to formulate this problem in a Bayesian way. Here we call E_2 the variable which indicates whether the host opens door 2; showing it to be empty. I_1 represents the initial choice of the contestant. What we are interested in finding is $Pr(C_1|E_2, I_1)$ to begin with:

$$Pr(C_1|E_2, I_1) = \frac{Pr(E_2|C_1, I_1)Pr(C_1|I_1)}{Pr(E_2|I_1)} \tag{1.4}$$

In order to calculate the terms, it is easiest to enumerate all the possible outcomes:

- CPEE
- CEPE
- CEEP
- PCEE
- PECE
- PEEC
- EECP
- ECEP
- ECEP
- ECPE
- EEPC
- EPEC

- EPCE

We can hence calculate the above:

$$\begin{aligned} Pr(C_1|E_2, I_1) &= \frac{\left(\frac{1}{2} \times \frac{2}{3}\right)\frac{1}{4}}{4 \times \frac{1}{2} \times \frac{1}{12} + 2 \times 1 \times \frac{1}{12}} \\ &= \frac{1}{4} \end{aligned} \tag{1.5}$$

Hence, the probability of winning a car if you stick with your initial choice is remains at $\frac{1}{4}$. Now we need to calculate $Pr(C_3|E_2, I_1)$ - the probability of a change to door 3 resulting in a car:

$$\begin{aligned} Pr(C_3|E_2, I_1) &= \frac{Pr(E_2|C_3, I_1)Pr(C_3|I_1)}{Pr(E_2|I_1)} \\ &= \frac{\left(\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1 + \frac{1}{3} \times 0\right)\frac{1}{4}}{\frac{1}{3}} \\ &= \frac{3}{8} \end{aligned} \tag{1.6}$$

Hence by switching, you are more likely to win the car, but the same is true for the penalty $Pr(C_3|E_2, I_1) = \frac{3}{8}$. Hence, if you want to maximise your chances of winning, then you should change. However, if you are risk averse, then you should stick.

1.2 Blood doping

	Lost	**Won**
Clean	0.7	0.05
Doping	0.15	0.15

Table 1.1: The historical probabilities of behaviour and outcome for professional 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 1.1.

Tables 1.2 and 1.3 can be calculated using conditional probabilities make the questions easier to handle.

	Lost	**Won**
Clean	0.93	0.067
Doping	0.60	0.40

Table 1.2: The probability of outcomes given a drug status.

Lost	**Won**	
Clean	0.82	0.33
Doping	0.18	0.67

Table 1.3: The probability of a drug status given an outcome.

1.2.1 What is the probability that a professional cyclist wins a race?

This is the marginal given by: $p(won) = p(won, dope) + p(won, clean) = 0.05 + 0.1 = 0.15$

1.2.2 What is the probability that a cyclist wins a race given that they have cheated?

This is a conditional distribution given by:

$$\begin{aligned} p(won|doped) &= \frac{p(doped, won)}{p(doped)} \\ &= \frac{0.1}{0.25} = 0.4 \end{aligned} \tag{1.7}$$

1.2.3 What is the probability that a cyclist is cheating given that he wins?

Use Bayes' rule:

$$\begin{aligned} p(doped|won) &= \frac{p(won|doped)p(doped)}{p(won)} \\ &= \frac{0.4 \times 0.25}{0.15} \\ &= \frac{2}{3} \end{aligned} \tag{1.8}$$

Now suppose that drug testing officials have a test which is relatively accurate in finding cheats. It accurately indicates 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:

1.2.4 They care only about the proportion of people correctly identified as dopers

Here we want to compare $p(doped|positive, group)$ across $group \in \{everyone, winners\}$. For everyone, this is simple and given by:

$$\begin{aligned} p(doped|positive) &= \frac{p(positive|doped)p(doped)}{p(positive)} \\ &= \frac{p(positive|doped)p(doped)}{p(positive, doped) + p(positive, clean)} \\ &= \frac{p(positive|doped)p(doped)}{p(positive|doped)p(doped) + p(positive|clean)p(clean)} \\ &= \frac{0.9 \times 0.25}{0.9 \times 0.25 + 0.05 \times 0.75} \\ &\approx 0.86 \end{aligned} \tag{1.9}$$

Whereas for the winners:

$$p(doped|positive, won) = \frac{p(doped, positive|won)}{p(positive|won)} \tag{1.10}$$

We will proceed to calculate each of these bits in turn. Via Bayes' rule:

$$\begin{aligned} p(doped, positive|won) &= \frac{p(won|doped, positive)p(doped, positive)}{p(won)} \\ &= \frac{p(won|doped)p(doped, positive)}{p(won)} \\ &= \frac{p(won|doped)p(positive|doped)p(doped)}{p(won)} \\ &= \frac{0.4 \times 0.9 \times 0.25}{0.15} \\ &= 0.6 \end{aligned} \tag{1.11}$$

We have got the second line from the first, by assuming that there is a conditional independence between winning and testing positive, once we account for their drug status. This is a fairly safe assumption, unless of course winners are more effective at hiding their drug use!

Now for the last bit:

$$\begin{aligned} p(positive|won) &= p(positive, doped|won) + p(positive, clean|won) \\ &= p(positive|doped, won)p(doped|won) + p(positive|clean, won)p(clean|won) \\ &= p(positive|doped)p(doped|won) + p(positive|clean)p(clean|won) \\ &= 0.9 \times \frac{2}{3} + 0.05 \times \frac{1}{3} \\ &\approx 0.62 \end{aligned} \tag{1.12}$$

Combining these two, we have $p(doped|positive, won) = \frac{0.6}{0.62} \approx 0.97$. Hence we should only test the winners. This makes intuitive sense, since they are a group which have a higher than average percentage of dopers.

1.2.5 They want only to minimise the proportion of falsely accused athletes

Here we want to minimise $p(clean|positive, group)$, where again $group \in \{everyone, winners\}$. For everyone, this is achieved through Bayes' rule:

$$\begin{aligned} p(clean|positive) &= \frac{p(positive|clean)p(clean)}{p(positive)} \\ &= \frac{0.05 \times 0.75}{0.9 \times 0.25 + 0.05 \times 0.75} \\ &\approx 0.14 \end{aligned} \tag{1.13}$$

Whereas for the winners' group:

$$\begin{aligned} p(clean|positive, won) &= \frac{p(positive, won|clean)p(clean)}{p(positive, won)} \\ &= \frac{p(positive|clean) \times p(won|clean)p(clean)}{p(positive, won)} \\ &= \frac{p(positive|clean) \times p(won|clean)p(clean)}{p(positive|won)p(won)} \\ &= \frac{0.05 \times 0.07 \times 0.75}{0.62 \times 0.15} \\ &\approx 0.03 \end{aligned} \tag{1.14}$$

Where we have got the second line from the first by assuming conditional independence of testing and outcome given a drug status.

Hence we should choose the winners' group again.

1.2.6 They care only about the *number* of people correctly identified as dopers

Here what we care about is the expected $n(doped|positive, group) = n(group) \times p(doped|positive, group)$ across $group \in \{everyone, winners\}$. For everyone, this is given by:

$$\begin{aligned} n(doped|positive) &= n(total) \times p(doped|positive) \\ &= 0.86n(total) \end{aligned} \tag{1.15}$$

Whereas for the winners:

$$\begin{aligned} n(doped|positive, won) &= n(won) \times p(doped|positive, won) \\ &= n(won) \times p(doped|positive, won) \\ &= n(total) \times p(won) \times p(doped|positive, won) \\ &= n(total) \times 0.15 \times 0.97 \\ &= 0.15n(total) \end{aligned} \tag{1.16}$$

Therefore in this circumstance, we should test everyone! This makes intuitive sense, since the latter is a subset of the former.

1.2.7 They care five times about the *number* of people who are falsely identified as they do about the *number* of people who are correctly identified as dopers

Now we need to specify a utility function of the form:

$$\begin{aligned} U(group) &= n(doped|positive, group) - 5n(clean|positive, group) \\ &= n(group)\left[p(doped|positive, group) - 5p(clean|positive, group)\right] \\ &= n(group)\left[p(doped|positive, group) - 5(1 - p(doped|positive, group))\right] \\ &= n(total)\left[6p(doped|positive, group) - 5\right] \end{aligned} \tag{1.17}$$

Calculating this for everyone, we have:

$$\begin{aligned} U(total) &= n(total)\left[6p(doped|positive) - 5\right] \\ &\approx n(total)\left[6 \times 0.86 - 5\right] \\ &\approx 0.14n(total) \end{aligned} \tag{1.18}$$

For only the winners' group, we have:

$$\begin{aligned} U(won) &= n(total) \times p(won)\left[6p(doped|positive, won) - 5\right] \\ &\approx n(total) \times 0.15\left[6 \times 0.97 - 5\right] \\ &\approx 0.13n(total) \end{aligned} \tag{1.19}$$

So in this case they should test everyone.

1.2.8 What factor would make the officials choose the other group? (By factor, we mean the number 5 in the previous question.)

They want the number α which makes the utility from testing the winners' group exceed that of testing everyone $U(won) > U(total) \implies$

$$n(total)\left[\alpha p(doped|positive) - 5\right] > n(total)p(won)\left[\alpha p(doped|positive, won) - 5\right] \tag{1.20}$$

We can remove the factor of $n(total)$ since it appears on both sides. Rearranging, we are then left with:

$$\alpha > \frac{5\left[1 - p(won)\right]}{p(doped|positive) - p(won)p(doped|positive, won)} \tag{1.21}$$

Which, when you substitute in the numbers yields $\alpha > 5.97$.

Chapter 2

Posterior

2.1 The lesser evil

Suppose that you area 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 - 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 "data_Posterior_PS_tumour.csv".

Suppose that the more brain that is cut, the more the patient is at risk of losing cognitive functions. Additionally, suppose that the damage inflicted varies:

1. Linearly with the distance the surgery starts away from the tumour.
2. Quadratically with the distance the surgery starts away from the tumour.
3. There is no damage if tissue cut is within 1mm of the tumour.

Under each of the three regimes above, find the best position along this axis from which to belong the surgery?

Note to teachers: this question borrows from Bayesian decision theory. However, it is hoped that the students can work through this without a formal introduction, since they have been lead some of the way.

2.1.1 Linearly with the amount of tissue cut.

Here we assume that there is a loss function of the form:

$$L(\hat{\theta}, \theta) = k_1|\hat{\theta} - \theta| \tag{2.1}$$

Minimising the expected loss yields the median. From the data file ”data_Posterior_PS_tumour.csv we find that this is given by approximately 9.5.

2.1.2 Quadratically with the amount of tissue cut,

Here the loss function takes the form:

$$L(\hat{\theta}, \theta) = k_2(\hat{\theta} - \theta)^2 \tag{2.2}$$

Minimising the expected loss yields the mean. From the data file ”data_Posterior_PS_tumour.csv” we find that this is given by approximately 10.4.

2.1.3 There is no damage if tissue cut is within 2cm of the tumour.

This loss function, is of the form:

$$L(\hat{\theta}, \theta) = \begin{cases} 0, & \text{if } |\hat{\theta} - \theta| < 1mm \\ k_3, & \text{otherwise} \end{cases} \tag{2.3}$$

Since the distance, 1mm, is small relative to the length scales of the posterior, the expression is approximately maximised by the mode. The mode is approximately 11.5 from the sample.

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

Since the brain is inherently non-linear, you might expect that if the surgery starts a long way away from the tumour's location, and moves nearer, then it will cause disproportionately more damage than a cut that starts nearer. This might suggest that a quadratic loss might be most appropriate. Alternatively, since the cut will occur along two axes, then the area is really a more appropriate measure of the damage done; again suggesting the mean over the median. The latter suggestion might also suggest that a cubic loss be appropriate.

The step function loss, which results in the mode, is not really appropriate, since the damage surely increases the more tissue is cut.

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

In the majority of cases, then we probably care disproportionately more about estimates that are a long way away from the true value, than we do about those that are nearer. This might suggest that we use a quadratic loss function. A case could be made for the linear loss, although not for the win-loss case for the mode.

Following from the previous point, which measure of posterior centrality might you choose?

The posterior mean.

Chapter 3

Distributions

3.1 Drug trials

We suppose that we are testing the efficacy of a certain drug which aims to cure depression, across two groups, each of size 10, with varying levels of the underlying condition: *mild* and *severe*. We suppose that the success rate of the drug varies across each of the groups, with $\theta_{mild} > \theta_{severe}$. We are comparing this with another group of 10 individuals, which has a success rate equal to the mean of the other two groups $\theta = \frac{\theta_{mild}+\theta_{severe}}{2}$.

3.1.1 Calculate the mean number of successful trials in each of the three groups.

Across each of the three groups:

- $\mathrm{E}[X_{mild}] = 10\theta_{mild}$
- $\mathrm{E}[X_{severe}] = 10\theta_{severe}$
- $\mathrm{E}[X_{homogeneous}] = 20\theta_{homogeneous}$

3.1.2 Compare the mean across the two heterogeneous groups, with that of the single group of 10 homogeneous people.

Combining the two groups we have:

$$\begin{aligned} E[X_{combined}] &= \frac{1}{2} \times 10(\theta_{mild} + \theta_{severe}) \\ &= 10\theta_{homogeneous} \end{aligned}$$

In words, the mean outcome across the two groups is the same.

3.1.3 Calculate the variance of outcomes across each of the three groups.

The variance across each of the three groups is given by:

- $var(X_{mild}) = 10\theta_{mild}(1 - \theta_{mild})$
- $var(X_{severe}) = 10\theta_{severe}(1 - \theta_{severe})$
- $var(X_{homogeneous}) = 10\theta_{homogeneous}(1 - \theta_{homogeneous})$

3.1.4 How does the variance across both heterogeneous studies compare with that of an equivalent 10-person homogeneous group?

Here we need to use the law of total variance. This is because there are two sources of variance: that which is intra-group, and another which is between group.

$$var(X_{combined}) = E[var(X|D)] + var(E[X|D]) \tag{3.1}$$

here D means the depressive status of the particular subgroup.

Using this we have:

$$\begin{aligned} var(X_{combined}) &= \mathrm{E}[var(X|D)] + \mathrm{E}\left(\mathrm{E}[X|D]^2\right) - (\mathrm{E}\,(\mathrm{E}[X|D]))^2 \\ &= \frac{1}{2} \times 10 \times \theta_{mild}(1-\theta_{mild}) + \frac{1}{2}10 \times \theta_{severe}(1-\theta_{severe}) \\ &+ \frac{1}{2} \times 10^2 \times \theta^2_{mild} + \frac{1}{2} \times 10^2 \times \theta^2_{severe} - 10^2 \times \theta^2_{homogeneous} \end{aligned} \tag{3.1}$$

Now supposing that we can write $\theta_{mild} = \theta_{homogeneous} - \epsilon$ and $\theta_{severe} = \theta_{homogeneous} + \epsilon$. We can then substitute this into the above yielding:

$$var(X_{combined}) = n\theta_{homogeneous}(1-\theta_{homogeneous}) + \epsilon^2 n(n-1) \tag{3.1}$$

Here $n = 10$, so the variance is greater than that of the homogeneous group. Note, the latter term disappears if $n = 1$ since there is no between-group variance!

No consider the extension to a large number of trials, with the depressive status of each group unknown to the experimenter, but follows $\theta \sim Beta(\alpha, \beta)$.

3.1.5 Calculate the mean value of the Beta distribution.

This can be calculated straightforwardly, and found to be $\frac{\alpha}{\alpha+\beta}$.

3.1.6 What combinations of α and β would make the mean the same as that of a single study with success probability θ?

Setting these equal:

$$\frac{\alpha}{\alpha+\beta} = \theta \tag{3.1}$$

Rearranging this we get the following relationship:

$$\alpha = \frac{\beta\theta}{1-\theta} \tag{3.1}$$

3.1.7 How does the variance change, as the parameters of the beta distribution are changed, so as to keep the same mean of θ?

The variance of a beta distribution can be calculated as:

$$var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3.1}$$

This can be shown to be equal to:

$$var(\theta) = \frac{\theta(1-\theta)^2}{\beta+1-\theta} \tag{3.1}$$

Therefore, as $\beta \to \infty \implies var(\theta) \to 0$.

3.1.8 How does the variance of the number of disease cases compare to that of the a single study with success probability θ?

It is possible to work out the variance of the beta-binomial distribution, and one finds it equal to:

$$var(X|n,\alpha,\beta) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3.1}$$

By recognising that $\theta = \frac{\alpha}{\alpha+\beta}$, and $1-\theta = \frac{\beta}{\alpha+\beta}$, the above expression can be written as:

$$var(X|n,\alpha,\beta) = n\theta(1-\theta)\frac{\alpha+\beta+n}{\alpha+\beta+1} \tag{3.1}$$

This can then be rearranged to yield:

$$\begin{aligned} var(X|n,\alpha,\beta) &= n\theta(1-\theta)\left[1+\frac{n-1}{\alpha+\beta+1}\right] \\ &= n\theta(1-\theta)+\epsilon \\ &\geq var(X|n,\theta) = n\theta(1-\theta) \end{aligned} \tag{3.1}$$

Therefore the variance of this distribution exceeds that of an equivalent binomial distribution. Hence, why it is called an over-dispersed distribution.