EXERCISES

ALAN HEAVENS

1. The Monty Hall Problem

1. Solve the 'Monty Hall' problem given in the lectures, using Bayes' theorem.

Answer: Let the doors be labelled a, b, c, where a is the door you choose initially, and b is the door which is opened. Many, if not all, of the probabilities below should be interpreted as 'given that you have chosen a', but for clarity we won't write this explicitly.

Let $p(a)$ = probability that a leads to the desired prize, etc.

Let B be the event that door b gets opened and leads to worthless junk.

What you want is the probability that a leads to the prize, given that b is opened and leads to junk. i.e. the aim is to calculate

$$p(a|B).$$

We can use Bayes' theorem for this:

$$p(a|B) = \frac{p(a,B)}{p(B)} = \frac{p(B|a)p(a)}{p(B)}$$

Now, clearly $p(a) = p(b) = p(c) = 1/3$ (all doors are equally likely, before any experiment is done).

$p(B|a)$ = probability that door b is opened, given that a leads to the prize. Evidently

$$p(B|a) = \frac{1}{2} :$$

Alan could have opened either door b or c, since they both lead to junk.

What about $p(B)$? It is the sum of all the joint probabilities:

$$p(B) = p(B,a) + p(B,b) + p(B,c) = p(B|a)p(a) + p(B|b)p(b) + p(B|c)p(c),$$

each of which we can calculate. $p(a) = p(b) = p(c) = 1/3$, as before, and $p(B|a) = 1/2$ as before. Now

$$p(B|b) = 0 :$$

Alan will not open b since it leads to the prize in this case.

$p(B|c)$ is the most interesting. Given that you have chosen a (remember this is implicit throughout), then if c leads to the prize, then Monty Hall *must* open door b, i.e.

$$p(B|c) = 1$$

So the probability that your original choice a leads to the prize is

$$\begin{aligned} p(a|B) &= \frac{p(B|a)p(a)}{p(B|a)p(a) + p(B|b)p(b) + p(B|c)p(c)} \\ &= \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + \left(0 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{3}\right)} \\ &= \frac{1}{3} \end{aligned} \tag{1}$$

So you would double your chances (from 1/3 to 2/3) if you switch to the other door.

2. The slope of the number counts of radio sources.

2. The distribution of flux densities of extragalactic radio sources are distributed as a power-law with slope $-\alpha$, say. In a non-evolving Euclidean universe $\alpha = 3/2$ (can you prove this?) and departure of α from the value 3/2 is evidence for cosmological evolution of radio sources. This was the most telling argument against the steady-state cosmology in the early 1960s (even though they got the value of α wrong by quite a long way).

Given error-free observations of radio sources with flux densities $S_i; i = 1, \ldots, n$ above a known, fixed measurement limit S_0, what is the posterior for α? What is the MAP (maximum a posteriori) value of α?

If a single source is observed with flux $S_1 = 2S_0$, what is the most probable value of α?

Can you deduce the uncertainty on α?

Hints:
1. You will use the distribution as a probability distribution function (pdf), $p(S)$ and will have to normalise it.
2. A pdf is not a probability, but $p(S)\,\Delta S$ is, for a (small) interval ΔS. You will need to introduce an arbitrary (but small) interval around each source.

Answer:
The model probability distribution for S is

$$p(S)dS = (\alpha - 1)\left(\frac{S}{S_0}\right)^{-\alpha}\frac{dS}{S_0}$$

where the factors $\alpha - 1$ and powers of S_0 arise from the normalization requirement

$$\int_{S_0}^{\infty} dS\, p(S) = 1.$$

The likelihood function L for n observed sources is obtained the following argument. Divide the flux range into a very large number N small bins of width ΔS, so that each bin has zero or 1 source in them. The probability of having zero sources is $\exp(-\lambda)$, and the probability of having 1 source is $\lambda\exp(-\lambda)$, where $\lambda = p(S)\Delta S$ is the expected number. If $\lambda \ll 1$ (take ΔS vanishingly small), then the probabilities are $1-\lambda$ and λ respectively. The joint probability of getting the data observed is

$$\prod_{\text{empty cells}} (1-\lambda) \prod_{\text{filled cells}} \lambda.$$

Ignoring the first product (of 1s, essentially), this is

$$\prod_{\text{filled cells}} p(S_i)\Delta S.$$

Hence the likelihood is (the ΔS terms just affect the constant of proportionality)

$$L(\alpha) \propto \prod_{i=1}^{n} (\alpha-1) S_0^{\alpha-1} S_i^{-\alpha}$$

with logarithm (ignoring an addictive constant)

$$\ln L = \sum_{i=1}^{n} \left[\ln(\alpha-1) + (\alpha-1)\ln S_0 - \alpha \ln S_i\right].$$

Maximising $\ln L$ with respect to α:

$$\frac{\partial}{\partial\alpha} \ln L = \sum_{i=1}^{n} \left(\frac{1}{\alpha-1} + \ln S_0 - \ln S_i\right) = 0$$

we find the minimum when

$$\alpha = 1 + \frac{n}{\sum_{i=1}^{n} \ln \frac{S_i}{S_0}}.$$

Suppose we only observe one source with flux twice the cut-off, $S_1 = 2S_0$, then

$$\alpha = 1 + \frac{1}{\ln 2} = 2.44$$

but with a large uncertainty. More useful in this case would be to present the full posterior, which, for a constant prior on α, is shown in Fig. 1, showing that it is quite skewed.

How can we infer the slope from only one object? The key is that we also know S_0, so we have two pieces of information. e.g. if the slope is steep, we will expect S_1 to be close to S_0.

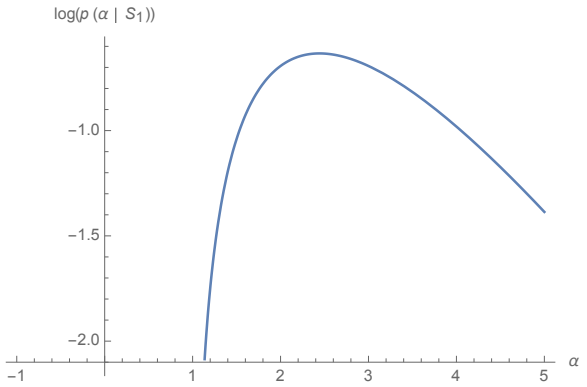

FIGURE 1. Posterior for the slope α if one source is observed with $S = 2S_0$.

3. You are the barrister

3. A known petty thief was lodging in a house when a theft of a piece of cheese takes place. The defence lawyer has argued that only 1/2500 of known thieves (T) who are in lodgings steal cheese from their hosts, *so the information that he is a known thief is irrelevant and must be ignored.* $p(S|T) = 0.0004$. You are the prosecution barrister, and you are fairly sure that the thief has stolen the cheese (this possible event we call S). How do you counter this argument? You are in possession of the knowledge that the probability of a theft of cheese (C) from lodgings is 1/20000.

What is the real probability, on the basis only of the information supplied here, that the lodger thief was the culprit? It is not 0.0004... Hint: what piece of information has the defence lawyer ignored (probably deliberately)?

Answer: The key extra ingredient is that in this case, cheese has definitely been stolen. The maths is basically the same as for the allergy problem in the lecture. Rule 1 says we want $p(S|C,T)$. All the probabilities are conditional on T:

$$p(S|C,T) = \frac{p(C|S,T)p(S|T)}{p(C|S,T)p(S|T) + p(C|\sim S,T)p(\sim S|T)}$$

Now $p(C|S,T) = 1$, $p(S|T) = 0.0004$ and $p(\sim S|T) = 0.9996$. $p(C|\sim S,T)$ is the probability that there is a cheese theft given that the perpetrator is not the thief. For this we take the probability of a theft in the general population, i.e. $p(C|\sim S,T) = 0.00005$, so

$$p(S|C,T) = \frac{1 \times 0.0004}{1 \times 0.0004 + 0.00005 \times 0.9996} = 0.89.$$

A very different conclusion. Put another way, for the landlord/ladies of every 100,000 petty thieves, there are 40 cheese thefts by them, and only 5 by strangers.

4. The Lighthouse Problem

4. This problem was set by Steve Gull to first year Cambridge students many years ago. It contrasts the Bayesian approach with an estimator-based approach.

A lighthouse is situated at unknown coordinates (x_0, y_0) with respect to a straight coastline $y = 0$. It sends a series of N flashes in random directions, and these are recorded on the coastline at positions $x_i;\ i = 1 \ldots N$. Only the positions of the arrivals of the flashes, not the directions, nor the intensities, are recorded. Using a Bayesian approach, find the posterior distribution of x_0, y_0.

Now focus only on the unknown x_0. Define a suitable estimator, $\hat{x}$, for x_0 from the observed x_i. Work out the probability distribution for $\hat{x}$. You may need to refer to a proof of the Central Limit Theorem, for the pdf of repeated trials of the same experiment. You may also find this useful:

$$\int_{-\infty}^{\infty} e^{ikx} \frac{1}{\left[1 + \frac{(x-x_0)^2}{y_0^2}\right]}\, dx = e^{ikx_0 - |k|y_0}$$

Comment.

You may like to simulate this process, compute the posterior distribution, and also show the estimator.

Answer: First, apply Rule 1. We want to know

$$p(x_0, y_0|\{x_i\})$$

Using Bayes, we write this as

$$p(x_0, y_0|\{x_i\}) \propto p(\{x_i\}|x_0, y_0)p(x_0, y_0) \propto \prod_i p(x_i|x_0, y_0)$$

if we assume a uniform prior for x_0, y_0.

Let the angle of the direction of the flash to the normal to the coastline be ψ. Then by trigonometry, the position that the flash arrives at is given by

$$\frac{x_i - x_0}{y_0} = \tan\psi_i.$$

So

$$p(x_i|x_0, y_0) = p(\psi_i|x_0, y_0) \left|\frac{d\psi_i}{dx_i}\right|$$

and for signals that are received on the shore, ψ is uniformly distributed in $-\pi/2 < \psi < \pi/2$, so $p(\psi_i) = 1/\pi$ in this range, independent of x_0, y_0. Also

$$\sec^2\psi_i \frac{d\psi_i}{dx_i} = \frac{1}{y_0} \Rightarrow \left[1 + \frac{(x_i - x_0)^2}{y_0^2}\right] \frac{d\psi_i}{dx_i} = \frac{1}{y_0}$$

and the likelihood of x_i is a Cauchy distribution:

$$p(x_i|x_0,y_0) = \frac{1}{\pi y_0\left[1+\frac{(x_i-x_0)^2}{y_0^2}\right]}.$$

Hence the (unnormalised) posterior for x_0, y_0 is

$$p(x_0,y_0|\{x_i\}) \propto \prod_{i=1}^{N} \frac{1}{\pi y_0\left[1+\frac{(x_i-x_0)^2}{y_0^2}\right]},$$

which is our desired outcome.

Estimator

A sensible-sounding estimator for x_0 is simply the average of the x_i:

$$\hat{x}_0 = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

For large N, one might hope that it gives a precise estimate of x_0. What is its distribution? We can use characteristic functions, where the characteristic function $\Phi(k)$ for the sum $(= N\bar{x}_0)$ is the product of the individual characteristic functions $\phi(k)$, so the characteristic function of the sum $N\bar{x}_0$ is

$$\Phi(k) = \phi^N(k)$$

where

$$\phi(k) = \int_{-\infty}^{\infty} e^{ikx} \frac{1}{\left[1+\frac{(x-x_0)^2}{y_0^2}\right]}\,dx = e^{ikx_0-|k|y_0}.$$

Hence

$$\Phi(k) = e^{iNkx_0-N|k|y_0}$$

which we can invert (by noting that x_0 gets replaced by Nx_0, and y_0 by Ny_0), to get the pdf of $(N\times)$ the estimator,

$$p(N\hat{x}_0) = \frac{1}{\pi N y_0\left[1+\frac{(N\hat{x}_0-Nx_0)^2}{N^2y_0^2}\right]}$$

and a simple change of variable gives

$$p(\hat{x}_0) = \frac{1}{\pi y_0\left[1+\frac{(\hat{x}_0-x_0)^2}{y_0^2}\right]}$$

so the distribution of the estimator is the same as for any individual x_i! Nothing is to be gained by averaging them, and it is no better than having one measurement! This seems to violate the CLT, but it does not, since the Cauchy distribution has infinite variance.