MRES IN MACHINE LEARNING AND BIG DATA
STATISTICS FOR EXPERIMENTAL PHYSICS
PART II: BAYESIAN STATISTICS
ASSESSED PROBLEM SHEET 1
DUE 1 DEC 2022

ALAN HEAVENS. A.HEAVENS@IMPERIAL.AC.UK

1. The Monty Hall Problem

Monty Hall hosted a game show called *Let's make a deal*.

You, the contestant, are given the choice of three doors: Behind one door is a car; behind the others, a goat. You pick a door, say Door a, and the host, who knows what's behind the doors, opens another door, say Door b, revealing a goat. He then says to you, "Do you want to change your choice?" Is it to your advantage to switch?

Solve this problem using Bayesian reasoning.

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 car, etc.

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

What you want is the probability that a leads to the car, given that b is opened and leads to the goat. 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 car. Evidently

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

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

Date: 24 November 2022.

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 car 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 car, then Monty Hall *must* open door b, i.e.

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

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

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

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

2. Estimator for the mean of a distribution

Data $\{x_i\}, i = 1, \ldots n$ are drawn independently from gaussian distributions with common mean value μ, and known variances σ_i^2. Show that

$$\hat{\mu} = \sum_{i=1}^{n} w_i x_i$$

is an unbiased estimate of μ, for any weights w_i which satisfy $\sum_{i=1}^{n} w_i = 1$.

Show that the variance of $\hat{\mu}$ is minimised if

$$w_i = \frac{\sigma_i^{-2}}{\sum_{i=1}^{n} \sigma_i^{-2}}.$$

Hint: variances of independent data add, and the variance of ax is a^2 times the variance of x. Write down the variance of $\hat{\mu}$, $\sigma_{\hat{\mu}}^2$, and minimise it with respect to a particular w_k. Note that you need to minimise the variance subject to the constraint on the sum of the weights. This needs a Lagrange multiplier.

You may wish to prove that the weighted sum of gaussian-distributed variables is gaussian (hint, use the mathematics used for proof of the Central Limit Theorem), but it is not asked

for here. Assume it, and hence show that the distribution of the minimum variance estimator is

$$p(\hat{\mu}|\mu) \propto \exp\left[-\frac{(\hat{\mu}-\mu)^2}{2\sigma_{\hat{\mu}}^2}\right].$$

This is the same formula as given in the lectures for a Bayesian posterior. Discuss the differences of interpretation of this formula in the frequentist and Bayesian case.

Taking the expectation value of $\hat{\mu}$, and noting that $\langle x_i \rangle = \mu$,

$$\langle \hat{\mu} \rangle = \sum_{i=1}^{n} w_i \langle x_i \rangle = \sum_{i=1}^{n} w_i \mu = \mu \sum_{i=1}^{n} w_i = \mu.$$

so $\hat{\mu}$ is an unbiased estimator of μ.

The variance of $\hat{\mu}$ is (from the rules given)

$$\sigma_{\hat{\mu}}^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2.$$

We minimise it, subject to the constraint $\sum_i w_i = 1$ (without it, we would just get $w_i = 0$), so we introduce a Lagrange multiplier λ and minimise

$$\sum_{i=1}^{n} w_i^2 \sigma_i^2 - \lambda \left(\sum_{i=1}^{n} w_i - 1 \right).$$

Differentiating w.r.t. w_k for some arbitrary k and setting to zero, we find

$$2 w_k \sigma_k^2 - \lambda = 0$$

Hence $w_i \propto \sigma_i^{-2}$ and the condition that the weights sum to unity gives the result.

With the weights determined, the variance of the estimator is

$$\sigma_{\hat{\mu}}^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{n} \sigma_i^{-4} \sigma_i^2}{\left(\sum_{i=1}^{n} \sigma_i^{-2}\right)^2} = \frac{1}{\sum_{i=1}^{n} \sigma_i^{-2}}.$$

Since the expectation value of $\hat{\mu}$ is μ, the minimum variance estimator (you are told it's gaussian) has a distribution

$$p(\hat{\mu}|\mu) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\mu}}} \exp\left[-\frac{(\hat{\mu}-\mu)^2}{2\sigma_{\hat{\mu}}^2}\right].$$

On the right is exactly the same mathematical expression as the Bayesian posterior for μ, given a set of data $\{x_i\}$, for a uniform prior on μ. The interpretation is entirely different though. The posterior gives the probability of the parameter μ, from a single (fixed) set of data. The frequentist expression is the probability distribution of the estimator $\hat{\mu}$, under repeated trials of data drawn from gaussians with the same μ.

3. The Lighthouse Problem

This problem was set by Steve Gull. It contrasts the Bayesian approach with an estimator-based frequentist approach, and is plausibly engineered to make the Bayesian approach look good.

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. Defining a suitable estimator, $\hat{x}$, for x_0 from the observed x_i (take the average). 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:

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

Comment on the mean and variance of the distribution of the estimator.

You may like to simulate this process, compute the posterior distribution, and also show the estimator. But you may not.

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

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

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,

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

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.