

STAT 238 - Bayesian Statistics

Lecture Five

Spring 2026, UC Berkeley

Aditya Guntuboyina

30 Jan 2026

In the last lecture, we started discussed the following problem.

1 Example 4: Coin Fairness Testing

Problem 1.1. *Suppose a coin is tossed 12 times leading to the outcome: TTTTHTHTTTTH (this has 3 heads and 9 tails). What is your assessment of the fairness of the coin?*

We looked at frequentist solutions in the last lecture. These are generally based on p -values. Calculation of these p -values require consideration of alternative datasets that could have appeared (but did not actually appear). As we discussed last time, such calculations would violate the likelihood principle.

Here is a Bayesian solution to this problem. The goal is to calculate:

$$\mathbb{P}\{\text{fairness} \mid \text{data}\}$$

where data refers to TTTTHTHTTTTH. By the Bayes rule, we can write

$$\mathbb{P}\{\text{fairness} \mid \text{data}\} = \frac{\mathbb{P}\{\text{data} \mid \text{fairness}\}\mathbb{P}\{\text{fairness}\}}{\mathbb{P}\{\text{data} \mid \text{fairness}\}\mathbb{P}\{\text{fairness}\} + \mathbb{P}\{\text{data} \mid \text{not fair}\}\mathbb{P}\{\text{not fair}\}}$$

We clearly have

$$\mathbb{P}\{\text{data} \mid \text{fairness}\} = 2^{-n}.$$

What assignments do we use for

$$\mathbb{P}\{\text{fairness}\}, \quad \mathbb{P}\{\text{not fair}\} \quad \text{and} \quad \mathbb{P}\{\text{data} \mid \text{not fair}\}?$$

For concreteness, let us assume

$$\mathbb{P}\{\text{fairness}\} = 0.5 \quad \text{and} \quad \mathbb{P}\{\text{not fair}\} = 0.5. \tag{1}$$

This is actually a very strong assumption in favor of fairness because a coin can be not fair in many many variety of ways. So to assume that the probability of fairness is the same as the combined probability of the many variety of ways in which the coin can be non-fair seems quite strong.

Let us now come to $\mathbb{P}\{\text{data} \mid \text{not fair}\}$. If the coin is not fair, we can assume that it has a heads probability of p and that the coin tosses are still independent. We can then write

$$\mathbb{P}\{\text{data} \mid \text{not fair}\} = \int_0^1 \mathbb{P}\{\text{data} \mid \text{not fair}, p\} f_{p|\text{not fair}}(p) dp = \int_0^1 p^3(1-p)^9 f_{p|\text{not fair}}(p) dp.$$

To proceed further, we need to assign $f_{p|\text{not fair}}(p)$. One concrete assumption might be that

$$f_{p|\text{not fair}}(p) = 1 \quad \text{for every } p \in [0, 1]. \quad (2)$$

This corresponds to the assumption that, under the alternative (not fair), p has the uniform distribution on $[0, 1]$. Then (using an online integrator)

$$\mathbb{P}\{\text{data} \mid \text{not fair}\} = \int_0^1 p^3(1-p)^9 dp = \frac{1}{2860}.$$

We then get

$$\mathbb{P}\{\text{fairness} \mid \text{data}\} = \frac{2^{-12} * 0.5}{2^{-12} * 0.5 + \frac{1}{2860} * 0.5} = 0.4111558.$$

Note that this Bayesian probability calculation does not depend at all on whether the number of tosses ($n = 12$) was decided a priori or whether it was decided to toss until getting 3 heads. It is the same for both those cases.

Also note that the Bayesian approach (based on (1) and (2)) is only slightly supporting the alternative hypothesis (roughly 60% to the null 40%) while the frequentist p -values are fairly small indicating more evidence for the alternative. This discrepancy also persists when the sample size is large. Consider the following example.

Example 1.2. *In a certain city, 49581 boys and 48870 girls have been born over a certain time period (note $49581/(49581 + 48870) = 0.5036109$). Assuming that the number of male births is binomially distributed with parameters $n = 49581 + 48870 = 98451$ and p , test the hypothesis $H_0 : p = 0.5$.*

The usual frequentist p -value is:

$$\mathbb{P}\{\text{Bin}(n, 0.5) \geq 49581\} \approx 0.01163$$

which is fairly small.

On the other hand, the Bayesian method above with the priors (1) and (2) gives

$$\mathbb{P}\{p = 0.5 \mid \text{data}\} = \frac{2^{-n} * 0.5}{2^{-n} * 0.5 + B(x+1, n-x+1) * 0.5} = 0.950523.$$

Here $x = 49581$, $n = 98451$ and $B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp$ is the Beta function.

Thus the Bayesian method gives a high probability to the null while the frequentist method will reject the null hypothesis. The reason why the Bayesian method is so supportive of the null hypothesis is that the prior choice (1) gives strong support to $p = 0.5$ over nearby values of p (such as $p \in (0.49, 0.51)$).

This discrepancy between the Bayesian and Frequentist solutions in this problem is referred to as the Jeffreys-Lindley paradox (see https://en.wikipedia.org/wiki/Lindley%27s_paradox).

2 Example 5: MacKay sequence example

Here is another example of hypothesis testing or model selection in the Bayesian framework. This example comes from the book MacKay [1, Chapter 28].

Problem 2.1. *Find the next number in the sequence: $-1, 3, 7, 11$.*

Most people would look at the sequence and guess the next number as 15. In other words, they recognize that the data form an arithmetic progression. If they are willing to consider alternative models, then we can pose the question as that of model selection, and use Bayesian methods (probability) to formally answer it.

Here are two possible models:

1. **Model 1:** Arithmetic Progression i.e., $a_1 = \alpha$ and $a_{n+1} = a_n + \beta$.
2. **Model 2:** Random (will be made precise later).

A Bayesian solution to this problem will attempt to calculate

$$\mathbb{P}\{\text{Model } i \mid \text{data}\} \quad \text{for } i = 1, 2.$$

What probability assignments would we need to calculate the above? We can use the Bayes Rule to write

$$\mathbb{P}\{\text{Model } i \mid \text{data}\} = \frac{\mathbb{P}\{\text{data} \mid \text{Model } i\} \mathbb{P}\{\text{Model } i\}}{\mathbb{P}\{\text{data} \mid \text{Model } 1\} \mathbb{P}\{\text{Model } 1\} + \mathbb{P}\{\text{data} \mid \text{Model } 2\} \mathbb{P}\{\text{Model } 2\}} \quad (3)$$

To be equally fair to both models, we shall take

$$\mathbb{P}\{\text{Model } i\} = \frac{1}{2} \quad \text{for each } i = 1, 2 \quad (4)$$

We now need to calculate $\mathbb{P}\{\text{data} \mid \text{Model } i\}$ for $i = 1, 2$. For $i = 1$, we have (below α and β are the parameters in Model 1):

$$\mathbb{P}\{\text{data} \mid \text{Model } 1\} = \mathbb{P}\{\alpha = -1, \beta = 4\}$$

To calculate the above, we need to make a probability assignment for the probability with which α and β take various values. MacKay [1, Chapter 28] assumes that α and β are integer-valued that they are independently uniformly distributed over the set $\{-50, -49, \dots, 49, 50\}$ which has cardinality 101. Then

$$\mathbb{P}\{\text{data} \mid \text{Model } 1\} = \mathbb{P}\{\alpha = -1, \beta = 4\} = \mathbb{P}\{\alpha = -1\} \mathbb{P}\{\beta = 4\} = \left(\frac{1}{101}\right)^2 \approx 9.8 \times 10^{-5}. \quad (5)$$

For the second model, we need to specify what we mean by “random”. We shall take this to mean that a_1, a_2, a_3, a_4 are independently distributed according to the uniform distribution on $\{-50, -49, \dots, 49, 50\}$. Then

$$\begin{aligned} \mathbb{P}\{\text{data} \mid \text{Model } 2\} &= \mathbb{P}\{a_1 = -1, a_2 = 3, a_3 = 7, a_4 = 11\} \\ &= \mathbb{P}\{a_1 = -1\} \mathbb{P}\{a_2 = 3\} \mathbb{P}\{a_3 = 7\} \mathbb{P}\{a_4 = 11\} = \left(\frac{1}{101}\right)^4 \approx 9.6 \times 10^{-9}. \end{aligned}$$

Plugging in the above value (as well as (4) and (5)) in (3), we get

$$\mathbb{P}\{\text{Model } 1 \mid \text{data}\} \approx \frac{101^{-2} \times 0.5}{101^{-2} \times 0.5 + 101^{-4} \times 0.5} = 0.999902 \quad \text{and} \quad \mathbb{P}\{\text{Model } 2 \mid \text{data}\} = .000098$$

This analysis clearly favors Model 1 compared to Model 2. The most interesting feature about this analysis is that

$$\mathbb{P}\{\text{Model 1}|\text{data}\} \gg \mathbb{P}\{\text{Model 2}|\text{data}\}$$

even though

$$\mathbb{P}\{\text{Model 1}\} = \mathbb{P}\{\text{Model 2}\}.$$

In other words, we did not dogmatically assert that the data was generated by an arithmetic progression but we gave a fair chance to the two models to explain the observed sequence.

In this example, some people argue in favor of Model 1 on the basis that Model 1 is “simpler” than Model 2. The Bayesian analysis above (based on probability theory) does not invoke any vague notion of simplicity but does some formal calculations which in this case preferred Model 1 to Model 2. In another situation, Model 2 may well be the preferred model.

One can consider other alternative models in this problem. For example, MacKay [1, Chapter 28] considered the following cubic model:

Model 3 (Cubic): These numbers were generated by the formula: $a_1 = a$ and $a_{n+1} = ba_n^3 + ca_n^2 + d$ for an integer a and rational numbers b, c, d .

This cubic model explains the given data perfectly if and only if its four parameters a, b, c, d are chosen as $a = -1, b = -1/11, c = 9/11, d = 23/11$. As a result,

$$\mathbb{P}\{\text{data}|\text{Model 3}\} = \mathbb{P}\{a = -1, b = -1/11, c = 9/11, d = 23/11\}.$$

In order to explicitly calculate the above, we need to make probability assignments for a, b, c, d . MacKay [1] makes the following probability assignment: we assume that these four parameters are independent with a being uniform on $\{-50, -49, \dots, 49, 50\}$ and b, c, d having the distribution of x/y where $x \sim \text{Unif}\{-50, -49, \dots, 49, 50\}$ and $y \sim \text{Unif}\{1, \dots, 50\}$ are independent. Under this assignment:

$$\begin{aligned} \mathbb{P}\{a = -1, b = -1/11, c = 9/11, d = 23/11\} &= \mathbb{P}\{a = -1\}\mathbb{P}\{b = -1/11\}\mathbb{P}\{c = 9/11\}\mathbb{P}\{d = 23/11\} \\ &= \left(\frac{1}{101}\right) \left(4 \cdot \frac{1}{101} \cdot \frac{1}{50}\right) \left(4 \cdot \frac{1}{101} \cdot \frac{1}{50}\right) \left(2 \cdot \frac{1}{101} \cdot \frac{1}{50}\right) \\ &\approx 2.5 \times 10^{-12}. \end{aligned}$$

In the above, we used $\mathbb{P}(b = -1/11) = 4 \cdot (1/101) \cdot (1/50)$ because $-1/11 = -2/22 = -3/33 = -4/44$ and each of these has the probability $(1/101) \cdot (1/50)$. A similar reasoning is used for $\mathbb{P}\{c = 9/11\}$ and $\mathbb{P}\{d = 23/11\}$.

If Model 1, 2, 3 are the only three models considered, the Bayes rule (3) becomes

$$\mathbb{P}\{\text{Model } i|\text{data}\} = \frac{\mathbb{P}\{\text{data}|\text{Model } i\} \mathbb{P}\{\text{Model } i\}}{\mathbb{P}\{\text{data}\}}$$

where the denominator should be calculated as:

$$\mathbb{P}\{\text{data}\} = \sum_{i=1}^3 \mathbb{P}\{\text{data}|\text{Model } i\} \mathbb{P}\{\text{Model } i\}.$$

Under the fair assumption

$$\mathbb{P}\{\text{Model } i\} = \frac{1}{3} \quad \text{for each } i = 1, 2, 3,$$

we obtain

$$\mathbb{P}\{\text{Model 1|data}\} = \frac{101^{-2} \times (1/3)}{101^{-2} \times (1/3) + 101^{-4} \times (1/3) + 2.5 \times 10^{-12} \times (1/3)} = 0.999902$$

and

$$\mathbb{P}\{\text{Model 2|data}\} \approx 9.8 \times 10^{-5}$$

and

$$\mathbb{P}\{\text{Model 3|data}\} \approx 9.8 \times 10^{-5} \approx 2.55 \times 10^{-8}.$$

Our preference for Model 1 is still as strong as before (when we only considered the two models Model 1 and Model 2).

The analysis given here depends on the specific choices of priors used for the three models. One can of course use alternative priors but the qualitative preference for Model 1 is unlikely to change for most **reasonable** prior choices.

References

- [1] MacKay, D. J. C., (2003) *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.