

STAT 238 - Bayesian Statistics

Lecture Six

Spring 2026, UC Berkeley

Aditya Guntuboyina

02 Feb 2026

1 Interpretation of Probability

Because Bayesian statistics is simply probability theory applied to inference, understanding the meaning and interpretation of probability is essential.

There are broadly two ways of understanding probability: frequentist and Bayesian.

2 Frequentist/Objective Understanding of Probability

From the frequentist viewpoint, probability is applicable only in the context of “random experiments” (such as tossing coins and rolling dice). The probability $\mathbb{P}(A)$ of an event A is defined as the relative frequency that A occurs in N repeated trials of the experiment in the limit as $N \rightarrow \infty$:

$$\mathbb{P}(A) := \lim_{N \rightarrow \infty} \frac{N_A}{N}$$

where N_A is the number of trials out of N where A occurs. This definition of probability is the basis of frequentist statistics.

Here are some examples:

1. The statement $\mathbb{P}(H) = 0.5$ means that the proportion of heads in a large number of tosses of the coin approaches 0.5.
2. The statement

$$\epsilon_1, \dots, \epsilon_n \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

means that if the experiment generating $\epsilon_1, \dots, \epsilon_n$ is repeated a large number of times, the proportion of times the values of $(\epsilon_1, \dots, \epsilon_n)$ lie in a set A approaches

$$\int_A \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_i^2}{2\sigma^2}\right) dx_1 \dots dx_n$$

and this should be true for all subsets A of \mathbb{R}^n .

The following are some obvious problems with the frequentist definition:

1. It is very restrictive and hardly ever applicable. In many simple situations where we would like to use probability, the frequency definition is simply does not apply:

- a) Is the suspect X guilty?
- b) What is the chance of rain in Berkeley today?
- c) What is the chance that Y is cancer positive given that they tested positive?

For an interesting anecdote about how this restrictive notion does not simply make sense in some important problems, see deGroot [4, pages 43-44].

2. Even in situations where the frequency definition is seemingly applicable, closer thought might reveal some issues. For example, the frequentist probability that a coin comes up heads is 0.6 means that 60% of a large number of tosses of the coin should result in 0.6. But the mechanics of no two tosses are really identical and if two tosses are done exactly identically, then we would expect the same outcome by the laws of physics. So the term “identical and independent repetitions of an experiment” is ambiguous.

In the frequentist definition, probability is considered an intrinsic property of the object under investigation which is only accessible by an experiment generating samples of infinite size. Thus frequentist probability is also referred to as “objective probability”. The implication is that we cannot assign it arbitrarily because any probability assignment that does not agree with the frequency in infinite trials is wrong. Unfortunately, the actual frequentist probability is seldom known because one cannot generally observe a large number of repetitions of an experiment and so almost all probability assignments are wrong from the frequentist point of view. This is one way of understanding the statistics aphorism: “All models are wrong” (usually attributed to George Box; see https://en.wikipedia.org/wiki/All_models_are_wrong).

Here are some quotes by famous statisticians/probabilists illustrating how widespread frequentist thinking in probability is:

The numbers p_r should, in fact, be regarded as physical constants of the particular die that we are using, and the question as to their numerical values cannot be answered by the axioms of probability, any more than the size and the weight of the die are determined by the geometrical and mechanical axioms. However, experience shows that in a well-made die the frequency of any event r in any long series of throws usually approaches $1/6$, and accordingly we shall often assume that all the p_r are equal to $1/6$... – Cramér.

Here is Jaynes’s response to the above quote (from page 317 of his book): *To a physicist, this statement seems to show utter contempt for the known laws of mechanics. The results of tossing a die many times do **not** tell us any definite number characteristic only of the die. They tell us also something about how the die was tossed. If you toss ‘loaded’ dice in different ways, you can easily alter the relative frequencies of the faces. With only slightly more difficulty, you can still do this if your dice are perfectly ‘honest’.*

Here is a quote by Feller (see page 322 of the Jaynes book) illustrating the thinking that bridge hands possess physical probabilities and that the uniform probability assignment is a convention whose correctness can only be verified by observed frequencies in a random experiment : *The number of possible distributions of cards in bridge is almost 10^{30} . Usually we agree to consider them as equally probable. For a check of this convention more than 10^{30} experiments would be required – a billion of billion of years if every living person played one game every second, day and night. – Feller.*

In spite of these objections, one positive aspect of the frequentist meaning of probability is that the Rules of Probability follow easily from this definition. Recall that the rules of probability are:

1. $\mathbb{P}(A)$ always lies between 0 and 1. The probability of an impossible event is 0 and the probability of a certain event is 1.
2. Product rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B)$.
3. Sum rule: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for disjoint events A and B .

3 Subjective or Bayesian Understanding of Probability

In Bayesian statistics, probability is considered a general method of reasoning under uncertainty. It is applicable to all situations involving uncertainty, and is *not* restricted to situations involving “random experiments”.

Further, probability is assumed to have nothing to do with frequency. This means there is no right or wrong probability model. Different analysts are welcome to use different models, and they can be assessed in terms of performance. One can also assess different models in a Bayesian model selection framework.

Meaning can be assigned to probability statements without relying on any connection to long-run frequencies. For example, suppose a doctor assigns a probability of 0.02 to a patient having cancer based on some background information. This can be quantified as (see Lindley [5, Chapter 3]): *The doctor’s degree of belief in the uncertain event that the patient has cancer is the same as the degree of belief of the uncertain event of drawing a red ball from an urn containing 2 red balls and 98 green balls.*

Lindley goes on to clearly say that there is no frequency interpretation here. *There is no repetition in this definition. The ball is to be taken once, and once only, and the long-run frequency of red balls in repeated drawings is irrelevant. After its withdrawal, the urn and its contents can go up in smoke for all that it matters.*

For a concrete example, consider the probability assignment $\mathbb{P}\{H \mid I\} = 0.5$ where H denotes heads (in the context of tossing a coin) and I denotes some background information. In the Bayesian context, this is not an informative statement on some long-run frequency of heads while tossing the coin. Instead, it is an assignment of probability by some individual for the next coin toss, based on some information. Since $\mathbb{P}\{H \mid I\} = \mathbb{P}\{T \mid I\}$, it implies that the background information is symmetric between H and T. The actual information itself might vary, for example, consider the following two kinds of background information both of which can justify this assignment $\mathbb{P}\{H \mid I\} = 0.5$:

1. **Information I_1 :** We don’t know anything at all about the coin. We don’t even know if it really has two sides H and T or both of its sides are of only one kind (either H or T). In addition, we don’t know how exactly it will be tossed.
2. **Information I_2 :** We know that it is a “regular” coin and that it has two sides H and T , and that it will be tossed in the “usual” way.

In the first case above, it might very well happen that the coin has both sides H, in which case repeated tossing of the coin will lead to HHHHHH.. so the long run frequency of heads will be 1 (and not 0.5). But this does not invalidate the assignment $\mathbb{P}(H \mid I_1) = 0.5$ because it is not a statement on long run frequency of heads, but it claims something about the next toss.

Now consider the third kind of information: We know that this coin has been tossed a large number of times in the past and it landed heads 70% of the time. Now is the assumption

$\mathbb{P}\{H \mid I\} = 0.5$ justifiable? In this case, the probability we need to calculate is:

$$\mathbb{P}\left\{X_{n+1} = 1 \mid \frac{X_1 + \dots + X_n}{n} = 0.7\right\} \quad (1)$$

where X_1, \dots, X_n represent the historical tosses and X_{n+1} denotes the outcome of the next toss (1 here represents H and 0 represents T). The probability (1) cannot be assigned arbitrarily but instead it should be calculated based on some joint model for X_1, \dots, X_{n+1} . Consider the following two models:

$$X_1, \dots, X_{n+1} \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.5) \quad (2)$$

and

$$X_1, \dots, X_{n+1} \mid \theta \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta) \quad \text{and} \quad \theta \sim \text{uniform}(0, 1). \quad (3)$$

For the first model, X_{n+1} is independent of X_1, \dots, X_n and we indeed get $\mathbb{P}(X_{n+1} = 1) = 0.5$. For the second model, the situation is more interesting, and we will get an answer to (1) that is close to 0.7 when n is large. So here, probability still does not have anything to do with frequency, but in this case, the right kind of model will lead to the frequency assignment.

This flexibility with modeling is a positive feature of the Bayesian framework. However, there remains the question of justification of the rules of probability. If analysts are allowed to come up arbitrary probability models, then why do they have to carry out calculations in accordance with the rules of probability. What is the justification for using the rules of probability? This question was raised by many people including R. A. Fisher, who in [2] writes:

Keynes establishes the laws of addition and multiplication of probabilities, by stating these laws in the form of definitions of the processes of addition and multiplication. The important step of showing that, when these probabilities have numerical values, “addition” and “multiplication” are so defined, are equivalent to the arithmetical processes ordinarily known by these names, is omitted. The omission is an interesting one, since it shows the difficulty of establishing the laws of mathematical probability, without basing the notion of probability on the concept of frequency, for which these laws are really true, and from which they were originally derived.

It turns out that the rules of probability can be justified without using any connections between probability and frequency. The following arguments are due to the physicist R. T. Cox and are described in Chapters 1 and Chapter 2 of Jaynes [3]. I will give a sketch of the argument skipping some important technical details. For the full argument, please read Jaynes [3, Chapter 1 and 2].

4 Justification of the Rules of Probability without using any connections to Long-run Frequencies

As just mentioned, the following argument is due to the physicist R. T. Cox, and can be read in the book Cox [1]. I will follow the treatment given in Jaynes [3, Chapter 1 and 2].

Let us first remove all restrictions on probabilities and even allow them to take values outside the interval $[0, 1]$. To avoid confusion, let us use the term “plausibilities”. We are assigning plausibilities of various events (or propositions) conditional on other events. Let us denote the plausibility of event A conditional on event B by $(A|B)$. Let us first make the assumption that plausibilities take values in the set of real numbers (no restriction now to be in the interval $[0, 1]$) and that a higher value of plausibility represents a greater belief.

4.1 Product Rule

Let us first investigate why the product rule should be true. The product rule in terms of probabilities states that

$$\mathbb{P}(AB|C) = \mathbb{P}(B|C)\mathbb{P}(A|BC)$$

Here AB denotes the event $A \cap B$. Should our plausibilities satisfy a similar inequality? Let us first assume that the plausibility $(AB|C)$ should really be determined by the two plausibilities $(B|C)$ and $(A|BC)$. This is basically because the process of deciding that AB is true can be broken down into first deciding whether B is true and then, having accepted B as true, deciding whether A is true. We shall therefore assume that there should be a function F such that

$$(AB|C) = F((B|C), (A|BC)).$$

We also assume that we should use the same function F for all possible events A, B, C (i.e., we are not using one function F for some A, B, C while calculating $(AB|C)$ from $(B|C)$ and $(A|BC)$ and using another function F for different A, B, C). This means in particular that

$$(AB|C) = F((A|C), (B|AC)).$$

It is also reasonable to assume that $F(x, y)$ is monotone increasing in each of its arguments and that it is continuous. If it is not continuous, then a small change in $(B|C)$ (or $(A|BC)$) might lead to a large change in $(AB|C)$ which is undesirable.

Now if we have four events A, B, C, D , we can write

$$(ABC|D) = F((BC|D), (A|BCD)) = F(F((C|D), (B|CD)), (A|BCD)).$$

We can also write

$$(ABC|D) = F((C|D), (AB|CD)) = F((C|D), F((B|CD), (A|BCD))).$$

We shall now make the following important consistency assumption: **If a plausibility can be calculated via two different methods, then both methods should give the same answer.** Clearly if this assumption were violated, then our answer to a plausibility calculation would depend on the specific method chosen to calculate and this would be highly undesirable. This assumption immediately implies that

$$F(F((C|D), (B|CD)), (A|BCD)) = F((C|D), F((B|CD), (A|BCD))).$$

for all A, B, C, D . If the individual plausibilities are arbitrary, we would get the following condition that the function F should satisfy

$$F(F(x, y), z) = F(x, F(y, z)) \text{ for all real numbers } x, y, z.$$

It now turns out the only functions F which satisfy the above equation are of the form

$$F(x, y) = w^{-1}(w(x)w(y))$$

for a positive continuous increasing function w . I will skip this derivation (see Section 2.1, Chapter 2 of Jaynes [3]). We thus have

$$(AB|C) = F((B|C), (A|BC)) = w^{-1}(w(B|C)w(A|BC)). \quad (4)$$

This is equivalent to

$$w(AB|C) = w(B|C)w(A|BC).$$

Now if we take $B = A$, we get

$$w(A|C) = w(A|C)w(A|AC)$$

The event $A|AC$ can be seen as certainty so we get

$$w(A|C) = w(A|C)w(\text{certainty})$$

for all A and C (note here that $w(\text{certainty})$ refers to the function w applied to the plausibility of the certain proposition). This can happen only if

$$w(\text{certainty}) = 1. \tag{5}$$

Also if we take $B = A^c$ in (4), we get

$$w(AA^c|C) = w(A^c|C)w(A|A^cC)$$

$AA^c|C$ and $A|A^cC$ can both be taken to represent impossibility so we get

$$w(\text{impossible}) = w(A^c|C)w(\text{impossible})$$

for all A and C which gives

$$w(\text{impossible}) = 0. \tag{6}$$

(5) and (6), along with the monotonicity of w , imply

$$0 \leq w(A|B) \leq 1 \quad \text{for all } A \text{ and } B. \tag{7}$$

We have thus proved that $w(A|B)$ lies always between 0 and 1 (is 0 for impossibility and 1 for certainty) and it satisfies the product rule of probability:

$$w(AB|C) = w(A|C)w(B|AC) = w(B|C)w(A|BC). \tag{8}$$

In other words, if we apply this this function w to our plausibilities, then the resulting assignments satisfy the first two rules of probability.

4.2 Sum Rule

Next the goal is to derive the sum rule. We will first derive the sum rule in the simplified form: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, and then prove it in the general case. We shall discuss this argument in the next lecture.

References

- [1] Cox, R. T. (2001) *Algebra of Probable Inference*. John Hopkins University Press, 2001.
- [2] R. A. Fisher, "Probability, likelihood and quantity of information in the logic of uncertain inference," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 146, no. 856, pp. 1–8, 1934.
- [3] Jaynes, E. T, (2003) *Probability theory: the logic of science*. Cambridge University Press, 2003.
- [4] DeGroot, M. H. (1986) A conversation with David Blackwell. *Statistical Science*, **1**, 40–53.
- [5] Lindley, D. V. (2013) *Understanding Uncertainty*. John Wiley and sons, 2013.