Probability Theory & Statistics

Conditional Probability

Introduction

All probabilities are conditional probabilities – there is always some prior knowledge (or assumption) built into every probability. Heads or tails coin flips assumes a fair coin, choosing a marble from an bag assumes the marbles are indistinguishable. Equivalently, conditional probabilities *are probabilities*, the same properties and axioms apply to conditional probabilities that apply to unconditional probabilities.[1]

Beyond prior or assumed knowledge, conditioning allows us to further update probabilities or beliefs given new evidence (or given evidence we wished we had). For example, suppose R is the event it will rain, and $P(R)$ is the probability it will rain. If we observe a change in the weather, maybe dark clouds appear in the sky, the probability it will rain is expected to increase, i.e., $P(R) < P(R|C)$.[2]

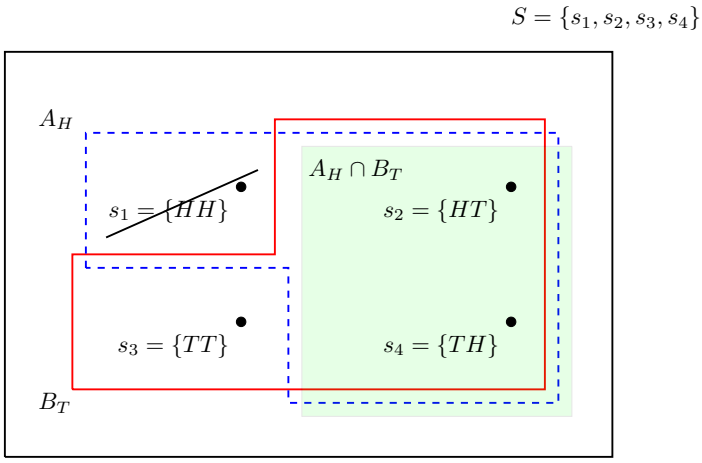

Figure 1: The 2-flip coin experiment after observing B_T, the event "at least 1 tail is observed", implies the event $\overline{B_T}$ cannot occur. This new information changes the sample space, and the probabilities assigned to each outcome must be recalculated.

[1] We won't prove this, but feel free to try!

[2] We observe something, obtain some new information, or speculate about what information we might get!

Definition (Recap)

If A and B are events with $P(B) > 0$, then the conditional probability of A given B, denoted by $P(A|B)$, is defined as[3]

$$P\left(A|B\right) := \frac{P(A \cap B)}{P(B)} \tag{1}$$

where A is an event whose uncertainty we want to update, and B is the evidence we observe (or want to treat as observed). We call $P(A)$ the *prior probability* of A, and $P(A|B)$ the *posterior probability* of A.

When we write $P(A|B)$, it does not mean that $A|B$ is an event and we are taking its probability; $A|B$ is *not an event*. Rather, $P\left(A|B\right) = Q(A)$; it is an *updated probability function* that assigns probabilities in accordance with the knowledge that B has occurred. $P(A)$ is a different probability function that assigns probabilities without regard for whether B has occurred or not, i.e., *unconditionally*.

The definition can be derived by considering subsets of finite sample spaces. The probability of the event A, conditional on B occurring, corresponds to the number of outcomes in A that occur in B, and renormalizing such that the total mass is 1 (see Figure 1).

Interpretation

The interpretation of conditional probability is often counter-intuitive. The *Linda Problem* is a famous example of the conflict of intuition and logical reasoning.

> Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which of the following two statements is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

In a survey at Stanford, 89% of undergraduates picked option 2. This is a logical fallacy because the second event is a subset of the first event. The naïve probability Linda is a bank teller corresponds to the proportion of the population that are bank tellers. The naïve probability Linda is a feminist *and* a bank teller must be lower than the naïve probability Linda is a bank

[3]The symbol := denotes "is defined as" and is used for axiomatic statements, i.e., statements that are assumed.

teller since of the proportion of the population that are bank tellers, a lesser or equal proportion are feminists.

By Bayes' theorem for $A \subseteq B$

$$P(A|B) := \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} \leq P(A) \tag{2}$$

The best way to gain intuition about conditional probability is simply through practice. The following conditional probability problems aim to build intuition.

The Two Children Problem

Presentation

Martin Gardener posed the following puzzle in his 1950s column

> *Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?*
>
> *Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys?*

Attempt to answer the question before moving to the next part.

Solution

Let G_1 be the event the younger child is a girl, and let G_2 be the event the older child is a girl. From the definition of conditional probability,

$$P\left(G_1 \cap G_2 | G_2\right) = \frac{P\left(G_1 \cap G_2\right)}{P\left(G_2\right)} = \frac{1/4}{1/2} = 1/2 \tag{3}$$

and

$$P\left(G_1 \cap G_2 | A_G\right) = \frac{P\left(G_1 \cap G_2\right)}{P\left(A_G\right)} = \frac{1/4}{3/4} = 1/3 \tag{4}$$

where A_G is the event that there is at least 1 girl.

It seems strange that it would make a difference if the child were the older or the younger of the two, indeed

$$P\left(G_1 \cap G_2 | G_2\right) = P\left(G_1 \cap G_2 | G_1\right) = 1/2 \tag{5}$$

But no such symmetry exists between the conditional probabilities $P\left(G_1 \cap G_2 | G_2\right)$ and $P\left(G_1 \cap G_2 | A_G\right)$. Saying the eldest child is a girl, *specifies* that one child must be a girl so that there is a 50% chance that the remaining child is a girl. Conditioning on the event that at least 1 child is a girl, removes the outcome $\{BB\}$ from the sample space $\{GG, GB, BG, BB\}$. (The similarity of this result with the 2-coin flip experiment conditioning example isn't by chance!)

Simulating Conditional Probability

The frequentist interpretation of probability states that the probability is...

> ... the number of favourable outcomes observed in a large number of repeated trials – *the relative frequency.*

The frequentist interpretation of conditional probability $P(A|B)$ corresponds to counting the number of occurrences in A, in the trials where B has occurred. This can be represented as a string of 0's and 1's, where B is the event that the first digit is 1, and A is the event the second digit is 1.

The number of occurrences is given by n_A, n_B, and n_{AB} for the events A, B, and $A \cap B$. The frequentist interpretation of probability is $P(A) \approx \frac{n_A}{n}$, $P(B) \approx \frac{n_B}{n}$, and $P(A \cap B) \approx \frac{n_{AB}}{n}$ for a large number of repeated experiments.

The conditional probability $P(A|B)$ is interpreted as the number of times the event A occurs in the experiments where B has occurred, i.e.,

$$P\left(A|B\right) \approx \frac{n_{AB}}{n_B} = \frac{n_{AB}/n}{n_B/n} \tag{6}$$

String	A	B	$A \cap B$
100010110001...		x	
000100100101...			
110010111101...	x	x	x
110011111100...	x	x	x
010111000100...	x		
100110101111...		x	

Table 1: String representation of conditioning on the first and second digit equal to 1.

which is equivalent to the definition of conditional probability.

We can simulate the two children problem with a frequentist interpretation of conditional probability – see supplementary script.

Corollaries

Probability of Intersection

From the definition of conditional probability

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A) \tag{7}$$

The usefulness of this theorem lies in the ability to express the *joint probability* $P(A \cap B)$ in terms of either $P(A|B)$ or $P(B|A)$. It is often difficult to directly calculate the joint probability of events, hence, rewriting the intersection of events provides a useful reformulation of the problem.

The probability of intersection can be extended indefinitely[4]

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \ldots P(A_n|A_1, \ldots, A_{n-1}) \tag{8}$$

The events on the l.h.s can be permuted without changing the joint probability. This provides further freedom in calculations when an appropriate ordering is chosen.

[4]Consider $P(C \cap E)$, where $E = A \cap B$, then by the probability of intersection, $P(A \cap B \cap C) = P(C \cap A \cap B) = P(C \cap E) = P(E)P(C|E) = P(A \cap B)P(C|A \cap B) = P(A)P(B|A)P(C|A \cap B)$.

Bayes' Rule

Bayes' Rule follows directly from the probability of intersection

$$P\left(A|B\right) = \frac{P\left(B|A\right)P(A)}{P(B)}. \tag{9}$$

Law of Total Probability

Bayes' Rule is often combined with the law of total probability (LOTP)

$$P(B) = \sum_i P\left(B|A_i\right)P(A_i), \tag{10}$$

where A_i is a disjoint partition of the sample space S.

Biased Coin Flip Problem

Presentation

Suppose you have a *fair coin* and a *biased coin*. The biased coin lands heads with a probability 3/4. You pick a coin at random and flip it three times. It lands heads all three times. Given this information, what is the probability you picked the fair coin?

Think about how to define the events and then try to use the LOTP to make the calculation easier.

Solution

Let A be the event the coin landed heads three times and define F as the event we picked the fair coin. We want to find $P(F|A)$, i.e., the probability we choose the fair coin after observing three heads.

We can calculate the probability of three heads given a fair coin with combinatorics, so Bayes' Rule is useful because it lets us express $P(F|A)$ in terms of $P(A|F)$

$$P\left(F|A\right) = \frac{P\left(A|F\right)P(F)}{P(A)} \tag{11}$$

The probability of getting three heads given a fair coin is simply $\frac{1}{2^3}$, and the probability of choosing the fair coin is $\frac{1}{2}$ (there are 2 coins, fair and biased, so it is equally likely we choose either coin). The unconditional probability of getting three heads after choosing *either coin* requires going through all of the possible combinations of choosing different coins and their outcomes.

We can use the LOTP to help with the calculation,

$$P(A) = P\left(A|F\right)P(F) + P\left(A|\overline{F}\right)P\left(\overline{F}\right) = \frac{1}{2^3}\cdot\frac{1}{2} + \left(\frac{3}{4}\right)^3\cdot\frac{1}{2} \tag{12}$$

Combining those results yields $P\left(F|A\right) \approx 0.23$, i.e., observing three heads halves our belief that the coin is fair!

Prosecutors Fallacy Problem

Presentation

In 1998, Sally Clark was tried for murder after two of her sons died shortly after birth. During the trial, an expert witness for the prosecution testified that the probability of a baby dying of sudden infant death syndrome (SIDS) was 1 in 8,500, so the probability of two deaths due to SIDS in one family is $\frac{1}{8{,}500^2} \approx \frac{1}{7.3\times10^7}$. The prosecutor argued that the probability of her innocence was 1 in 73 million, and she was sent to prison.

What issues are there with this line of reasoning?

Solution

The first issue is the assumption that $P(A_1 \cap A_2) = P(A_1)\,P(A_2)$ where A_1 is the event the first child dies of SIDS, and A_2 the second. This equality requires that A_1 and A_2 are *independent*. Ignoring the fact that there may be a genetic condition in the family.

The second issue is the claim of *innocence* given *evidence*. That is, the probability the witness is innocent is low given the evidence. The calculation in the court case was instead, the probability of the *evidence* given the witness is *innocent*.[5]

Using the LOTP

$$P(innocent|evidence) = \frac{P(evidence|innocent)\,P(innocent)}{P(evidence|innocent)\,P(innocent) + P(evidence|guilty)P(guilty)} \tag{13}$$

The final probability depends mostly on $P(evidence|innocent)$, which is very low, and $P(innocent) = 1 - P(guilty)$, which is very high (number of double infanticides in the population is very low).

Simpson's Paradox Problem

Presentation

Consider two Doctors, Dr Dave and Dr Davina, that each perform two types of surgeries. Dr Davina has a higher success rate in Brain Surgeries, $\frac{70}{90} \approx 78\%$, compared to Dr Dave's 20% success rate. Dr Davina also has a higher success rate giving Flu Jabs, 100% compared to Dr Dave's 90% success rate.

	Dr Davina		Dr Dave	
	Brain Surgery	Flu Jab	Brain Surgery	Flu Jab
Success	70	10	2	81
Failure	20	0	8	9

If we aggregate the successes though,

$$P_{dave}(success) = 83\% > P_{davina}(success) = 80\% \tag{14}$$

Dr Dave has a better overall success rate!

Use the rules of conditional probability to show that this is possible, and when this might occur.

[5]The witness did not intervene in the death and change the probability.

Solution

Let A be the event of a successful surgery, let B be the event that Dr Dave is the surgeon, and let C be the event that the surgery is Brain Surgery. We have *Simpson's paradox* if the probability of a successful surgery is lower for Dr Dave when it isn't aggregated

$$P\left(A|B,C\right) < P(A|\overline{B},C) \tag{15}$$

and

$$P\left(A|B,\overline{C}\right) < P(A|\overline{B},\overline{C}) \tag{16}$$

but

$$P\left(A|B\right) > P(A|B^c) \tag{17}$$

We can use LOTP to show that this is mathematically possible

$$P\left(A|B\right) = P\left(A|B,C\right)P\left(C|B\right) + P\left(A|B,\overline{C}\right)P(\overline{C}|B) \tag{18}$$

and

$$P\left(A|\overline{B}\right) = P\left(A|\overline{B},C\right)P\left(C|\overline{B}\right) + P\left(A|\overline{B},\overline{C}\right)P(\overline{C}|\overline{B}) \tag{19}$$

That is, the conditional probabilities $P(A|B)$ and $P(A|\overline{B})$ are *weighted averages* of $P\left(A|B,C\right)$ and $P(A|B,\overline{C})$, and $P\left(A|\overline{B},C\right)$ and $P(A|\overline{B},\overline{C})$. Because the weights are different, it's possible to have combinations of the success conditional on intersections that result in a flip of the inequality.

Numerically,

$$P\left(A|B\right) = 0.83 = \left(\frac{2}{10}\right) \times 0.1 + \left(\frac{81}{90}\right) \times 0.9 \tag{20}$$

and

$$P\left(A|B^c\right) = 0.80 = \left(\frac{70}{90}\right) \times 0.9 + \left(\frac{10}{10}\right) \times 0.1 \tag{21}$$

Putting more weight on the easier surgery, the second term in the first Equation, corresponding to the fact that Dr Dave is much less likely than Dr Davina to perform Brain Surgery, i.e.,

$$P\left(C|B\right) < P(C|\overline{B}) \tag{22}$$

and

$$P\left(\overline{C}|B\right) > P\left(\overline{C}|\overline{B}\right). \tag{23}$$

In practice this translates into, aggregating across *different types*, which presents a misleading picture. If we believe there might be *confounding variables* at play, we should examine the disaggregated data.

Other examples of when Simpson's paradox occurs:

- Gender discrimination in University applications where across all departments, Women are admitted at a lower rate, but in individual departments, are admitted at a higher rate.[6]
- Cricket batting averages, Player 1 has better in individual overs than Player 2, but has a lower average over the whole game
- Health effects of smoking in any given age group gives higher mortality than cigar smoking, but lower mortality across all age groups.

Practice

1) What are the assumptions in Gardener's problem?

2) Mr Watson has two children and you randomly run into one of the two and learn that she is a girl. What is the conditional probability that both are girls?

3) Prove the LOTP by representing the sample space as the area of a rectangle, and events, confined within the sample space.

[6] Women apply to more competitive departments, i.e., with more applicants. Equivalent to Dr Davina being more successful at Brain Surgery.