In probability theory and statistics, Bayes’ theorem (alternatively Bayes’ law or Bayes' rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes’ theorem, a person’s age can be used to more accurately assess the probability that they have cancer, compared to the assessment of the probability of cancer made without knowledge of the person's age.

One of the many applications of Bayes’ theorem is Bayesian inference, a particular approach to statistical inference. When applied, the probabilities involved in Bayes’ theorem may have different probability interpretations. With the Bayesian probability interpretation the theorem expresses how a subjective degree of belief should rationally change to account for availability of related evidence. Bayesian inference is fundamental to Bayesian statistics.

Bayes’ theorem is named after Rev. Thomas Bayes (/ˈbeɪz/; 1701–1761), who first provided an equation that allows new evidence to update beliefs. It was further developed by Pierre-Simon Laplace, who first published the modern formulation in his 1812 “Théorie analytique des probabilités.” Sir Harold Jeffreys put Bayes’ algorithm and Laplace's formulation on an axiomatic basis. Jeffreys wrote that Bayes’ theorem “is to the theory of probability what the Pythagorean theorem is to geometry.”[1]

Contents [hide]

1 Statement of theorem

2 History

3 Examples

3.1 Drug testing

3.2 A more complicated example

4 Interpretations

4.1 Bayesian interpretation

4.2 Frequentist interpretation

4.2.1 Example

5 Forms

5.1 Events

5.1.1 Simple form

5.1.2 Alternative form

5.1.3 Extended form

5.2 Random variables

5.2.1 Simple form

5.2.2 Extended form

5.3 Bayes’ rule

6 Derivation

6.1 For events

6.2 For random variables

7 See also

8 Notes

9 Further reading

10 External links

Statement of theorem[edit]

Visualization of Bayes' theorem.

Bayes' theorem is stated mathematically as the following equation:[2]

{\displaystyle P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B)}},} {\displaystyle P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B)}},}

where {\displaystyle A} A and {\displaystyle B} B are events and {\displaystyle P(B)\neq 0} {\displaystyle P(B)\neq 0}.

{\displaystyle P(A)} P(A) and {\displaystyle P(B)} P(B) are the probabilities of observing {\displaystyle A} A and {\displaystyle B} B without regard to each other.

{\displaystyle P(A\mid B)} P(A\mid B), a conditional probability, is the probability of observing event {\displaystyle A} A given that {\displaystyle B} B is true.

{\displaystyle P(B\mid A)} {\displaystyle P(B\mid A)} is the probability of observing event {\displaystyle B} B given that {\displaystyle A} A is true.

History[edit]

Bayes’ theorem was named after the Reverend Thomas Bayes (1701–1761), who studied how to compute a distribution for the probability parameter of a binomial distribution (in modern terminology). Bayes’ unpublished manuscript was significantly edited by Richard Price before it was posthumously read at the Royal Society. Price edited[3] Bayes’ major work “An Essay towards solving a Problem in the Doctrine of Chances” (1763), which appeared in “Philosophical Transactions,”[4] and contains Bayes’ Theorem. Price wrote an introduction to the paper which provides some of the philosophical basis of Bayesian statistics. In 1765 he was elected a Fellow of the Royal Society in recognition of his work on the legacy of Bayes.[5][6]

The French mathematician Pierre-Simon Laplace reproduced and extended Bayes’ results in 1774, apparently quite unaware of Bayes’ work.[7][8] The Bayesian interpretation of probability was developed mainly by Laplace.[9]

Stephen Stigler suggested in 1983 that Bayes’ theorem was discovered by Nicholas Saunderson, a blind English mathematician, some time before Bayes;[10][11] that interpretation, however, has been disputed.[12] Martyn Hooper[13] and Sharon McGrayne[14] have argued that Richard Price's contribution was substantial:

By modern standards, we should refer to the Bayes–Price rule. Price discovered Bayes’ work, recognized its importance, corrected it, contributed to the article, and found a use for it. The modern convention of employing Bayes’ name alone is unfair but so entrenched that anything else makes little sense.[14]

Examples[edit]

Drug testing[edit]

Tree diagram illustrating drug testing example. U, Ū, "+" and "−" are the events representing user, non-user, positive result and negative result. Percentages in parentheses are calculated.

Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual tests positive, what is the probability that he is a user?

{\displaystyle {\begin{aligned}P({\text{User}}\mid {\text{+}})&={\frac {P({\text{+}}\mid {\text{User}})P({\text{User}})}{P(+)}}\\&={\frac {P({\text{+}}\mid {\text{User}})P({\text{User}})}{P({\text{+}}\mid {\text{User}})P({\text{User}})+P({\text{+}}\mid {\text{Non-user}})P({\text{Non-user}})}}\\[8pt]&={\frac {0.99\times 0.005}{0.99\times 0.005+0.01\times 0.995}}\\[8pt]&\approx 33.2\%\end{aligned}}} {\displaystyle {\begin{aligned}P({\text{User}}\mid {\text{+}})&={\frac {P({\text{+}}\mid {\text{User}})P({\text{User}})}{P(+)}}\\&={\frac {P({\text{+}}\mid {\text{User}})P({\text{User}})}{P({\text{+}}\mid {\text{User}})P({\text{User}})+P({\text{+}}\mid {\text{Non-user}})P({\text{Non-user}})}}\\[8pt]&={\frac {0.99\times 0.005}{0.99\times 0.005+0.01\times 0.995}}\\[8pt]&\approx 33.2\%\end{aligned}}}

Despite the apparent accuracy of the test, if an individual tests positive, it is more likely that they do not use the drug than that they do. This surprising result arises because the number of non-users is very large compared to the number of users; thus the number of false positives outweighs the number of true positives. To use concrete numbers, if 1000 individuals are tested, there are expected to be 995 non-users and 5 users. From the 995 non-users, 0.01 × 995 ≃ 10 false positives are expected. From the 5 users, 0.99 × 5 ≈ 5 true positives are expected. Out of 15 positive results, only 5, about 33%, are genuine. This illustrates the importance of base rates, and how the formation of policy can be egregiously misguided if base rates are neglected.[15]

The importance of specificity in this example can be seen by calculating that even if sensitivity is raised to 100% and specificity remains at 99% then the probability of the person being a drug user only rises from 33.2% to 33.4%, but if the sensitivity is held at 99% and the specificity is increased to 99.5% then probability of the person being a drug user rises to about 49.9%.

A more complicated example[edit]

The entire output of a factory is produced on three machines. The three machines account for different amounts of the factory output, namely 20%, 30%, and 50%. The fraction of defective items produced is this: for the first machine, 5%; for the second machine, 3%; for the third machine, 1%. If an item is chosen at random from the total output and is found to be defective, what is the probability that it was produced by the third machine?

A solution is as follows. Let Ai denote the event that a randomly chosen item was made by the ith machine (for i = 1,2,3). Let B denote the event that a randomly chosen item is defective. Then, we are given the following information:

P(A1) = 0.2, P(A2) = 0.3, P(A3) = 0.5.

If the item was made by the first machine, then the probability that it is defective is 0.05; that is, P(B | A1) = 0.05. Overall, we have

P(B | A1) = 0.05, P(B | A2) = 0.03, P(B | A3) = 0.01.

To answer the original question, we first find P(B). That can be done in the following way:

P(B) = Σi P(B | Ai) P(Ai) = (0.05)(0.2) + (0.03)(0.3) + (0.01)(0.5) = 0.024.

Hence 2.4% of the total output of the factory is defective.

We are given that B has occurred, and we want to calculate the conditional probability of A3. By Bayes' theorem,

P(A3 | B) = P(B | A3) P(A3)/P(B) = (0.01)(0.50)/(0.024) = 5/24.

Given that the item is defective, the probability that it was made by the third machine is only 5/24. Although machine 3 produces half of the total output, it produces a much smaller fraction of the defective items. Hence the knowledge that the item selected was defective enables us to replace the prior probability P(A3) = 1/2 by the smaller posterior probability P(A3 | B) = 5/24.

Once again, the answer can be reached without recourse to the formula by applying the conditions to any hypothetical number of cases. For example, if 100,000 items are produced by the factory, 20,000 will be produced by Machine A, 30,000 by Machine B, and 50,000 by Machine C. Machine A will produce 1000 defective items, Machine B 900, and Machine C 500. Of the total 2400 defective items, only 500, or 5/24 were produced by Machine C.

Interpretations[edit]

A geometric visualisation of Bayes’ theorem. In the table, the values w, x, y and z give the relative weights of each corresponding condition and case. The figures denote the cells of the table involved in each metric, the probability being the fraction of each figure that is shaded. This shows that P (A |B ) · P (B ) = P (B |A ) · P (A ) i.e. P (A |B ) =

P (B |A ) · P (A )

/

P (B )

. Similar reasoning shows that P (Ā |B ) =

P (B |Ā ) · P (Ā )

/

P (B )

and so forth.

The interpretation of Bayes’ theorem depends on the interpretation of probability ascribed to the terms. The two main interpretations are described below.

Bayesian interpretation[edit]

In the Bayesian (or epistemological) interpretation, probability measures a “degree of belief.” Bayes’ theorem then links the degree of belief in a proposition before and after accounting for evidence. For example, suppose it is believed with 50% certainty that a coin is twice as likely to land heads than tails. If the coin is flipped a number of times and the outcomes observed, that degree of belief may rise, fall or remain the same depending on the results.

For proposition A and evidence B,

P (A ), the prior, is the initial degree of belief in A.

P (A | B ), the “posterior,” is the degree of belief having accounted for B.

the quotient

P(B |A )

/

P(B)

represents the support B provides for A.

For more on the application of Bayes’ theorem under the Bayesian interpretation of probability, see Bayesian inference.

Frequentist interpretation[edit]

Illustration of frequentist interpretation with tree diagrams. Bayes’ theorem connects conditional probabilities to their inverses.

In the frequentist interpretation, probability measures a “proportion of outcomes.” For example, suppose an experiment is performed many times. P(A) is the proportion of outcomes with property A, and P(B) that with property B. P(B | A ) is the proportion of outcomes with property B out of outcomes with property A, and P(A | B ) the proportion of those with A out of those with B.

The role of Bayes’ theorem is best visualized with tree diagrams, as shown to the right. The two diagrams partition the same outcomes by A and B in opposite orders, to obtain the inverse probabilities. Bayes’ theorem serves as the link between these different partitionings.

Example[edit]

Tree diagram illustrating frequentist example. R, C, P and P bar are the events representing rare, common, pattern and no pattern. Percentages in parentheses are calculated. Note that three independent values are given, so it is possible to calculate the inverse tree (see figure above).

An entomologist spots what might be a rare subspecies of beetle, due to the pattern on its back. In the rare subspecies, 98% have the pattern, or P(Pattern | Rare) = 98%. In the common subspecies, 5% have the pattern. The rare subspecies accounts for only 0.1% of the population. How likely is the beetle having the pattern to be rare, or what is P(Rare | Pattern)?

From the extended form of Bayes’ theorem (since any beetle can be only rare or common),

{\displaystyle {\begin{aligned}P({\text{Rare}}\mid {\text{Pattern}})&={\frac {P({\text{Pattern}}\mid {\text{Rare}})P({\text{Rare}})}{P({\text{Pattern}}\mid {\text{Rare}})P({\text{Rare}})+P({\text{Pattern}}\mid {\text{Common}})P({\text{Common}})}}\\[8pt]&={\frac {0.98\times 0.001}{0.98\times 0.001+0.05\times 0.999}}\\[8pt]&\approx 1.9\%\end{aligned}}} {\displaystyle {\begin{aligned}P({\text{Rare}}\mid {\text{Pattern}})&={\frac {P({\text{Pattern}}\mid {\text{Rare}})P({\text{Rare}})}{P({\text{Pattern}}\mid {\text{Rare}})P({\text{Rare}})+P({\text{Pattern}}\mid {\text{Common}})P({\text{Common}})}}\\[8pt]&={\frac {0.98\times 0.001}{0.98\times 0.001+0.05\times 0.999}}\\[8pt]&\approx 1.9\%\end{aligned}}}

Forms[edit]

Events[edit]

Simple form[edit]

For events A and B, provided that P(B) ≠ 0,

{\displaystyle P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B)}}\cdot \,} P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B)}}\cdot \,

In many applications, for instance in Bayesian inference, the event B is fixed in the discussion, and we wish to consider the impact of its having been observed on our belief in various possible events A. In such a situation the denominator of the last expression, the probability of the given evidence B, is fixed; what we want to vary is A. Bayes’ theorem then shows that the posterior probabilities are proportional to the numerator:

{\displaystyle P(A\mid B)\propto P(A)\cdot P(B\mid A)\ } P(A\mid B)\propto P(A)\cdot P(B\mid A)\ (proportionality over A for given B).

In words: posterior is proportional to prior times likelihood.[16]

If events A1, A2, ..., are mutually exclusive and exhaustive, i.e., one of them is certain to occur but no two can occur together, and we know their probabilities up to proportionality, then we can determine the proportionality constant by using the fact that their probabilities must add up to one. For instance, for a given event A, the event A itself and its complement ¬A are exclusive and exhaustive. Denoting the constant of proportionality by c we have

{\displaystyle P(A\mid B)=c\cdot P(A)\cdot P(B\mid A){\text{ and }}P(\neg A\mid B)=c\cdot P(\neg A)\cdot P(B\mid \neg A).} {\displaystyle P(A\mid B)=c\cdot P(A)\cdot P(B\mid A){\text{ and }}P(\neg A\mid B)=c\cdot P(\neg A)\cdot P(B\mid \neg A).}

Adding these two formulas we deduce that

{\displaystyle 1=c\cdot (P(B\mid A)\cdot P(A)+P(B\mid \neg A)\cdot P(\neg A)),} {\displaystyle 1=c\cdot (P(B\mid A)\cdot P(A)+P(B\mid \neg A)\cdot P(\neg A)),}

or

{\displaystyle c={\frac {1}{P(B\mid A)\cdot P(A)+P(B\mid \neg A)\cdot P(\neg A)}}={\frac {1}{P(B)}}.} {\displaystyle c={\frac {1}{P(B\mid A)\cdot P(A)+P(B\mid \neg A)\cdot P(\neg A)}}={\frac {1}{P(B)}}.}

Alternative form[edit]

Another form of Bayes’ Theorem that is generally encountered when looking at two competing statements or hypotheses is:

{\displaystyle P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B\mid A)P(A)+P(B\mid \neg A)P(\neg A)}}\cdot } P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B\mid A)P(A)+P(B\mid \neg A)P(\neg A)}}\cdot

For an epistemological interpretation:

For proposition A and evidence or background B,[17]

P(A),the prior probability, is the initial degree of belief in A.

P(−A), is the corresponding probability of the initial degree of belief against A: 1 − P(A) = P(−A)

P(B | A), the conditional probability or likelihood, is the degree of belief in B, given that the proposition A is true.

P(B | −A), the conditional probability or likelihood, is the degree of belief in B, given that the proposition A is false.

P(A | B), the posterior probability, is the probability for A after taking into account B for and against A.

Extended form[edit]

Often, for some partition {Aj} of the sample space, the event space is given or conceptualized in terms of P(Aj) and P(B | Aj). It is then useful to compute P(B) using the law of total probability:

{\displaystyle P(B)={\sum \_{j}P(B\mid A\_{j})P(A\_{j})},} P(B)={\sum \_{j}P(B\mid A\_{j})P(A\_{j})},

{\displaystyle \Rightarrow P(A\_{i}\mid B)={\frac {P(B\mid A\_{i})\,P(A\_{i})}{\sum \limits \_{j}P(B\mid A\_{j})\,P(A\_{j})}}\cdot } \Rightarrow P(A\_{i}\mid B)={\frac {P(B\mid A\_{i})\,P(A\_{i})}{\sum \limits \_{j}P(B\mid A\_{j})\,P(A\_{j})}}\cdot

In the special case where A is a binary variable:

{\displaystyle P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B\mid A)P(A)+P(B\mid \neg A)P(\neg A)}}\cdot } P(A\mid B)={\frac {P(B\mid A)\,P(A)}{P(B\mid A)P(A)+P(B\mid \neg A)P(\neg A)}}\cdot

Random variables[edit]

Diagram illustrating the meaning of Bayes' theorem as applied to an event space generated by continuous random variables X and Y. Note that there exists an instance of Bayes' theorem for each point in the domain. In practice, these instances might be parametrized by writing the specified probability densities as a function of x and y.

Consider a sample space Ω generated by two random variables X and Y. In principle, Bayes’ theorem applies to the events A = {X = x} and B = {Y = y}. However, terms become 0 at points where either variable has finite probability density. To remain useful, Bayes’ theorem may be formulated in terms of the relevant densities (see Derivation).

Simple form[edit]