Bayesian Statistics

Kolmogorov Axioms and Conditional Probabilities

We denote events by A, B, C, $\ldots$, and use the following notation:

1. $A \cap B = A \text{ and } B$, the event that A and B both occur.

2. $A^c = \text{not } A$, the event that A does not occur.

3. E, the event which always occurs.

4. $A \cup B = A \text{ or } B$, the event that A or B, including both occur. Equivalently, at least one of the event A or B occurs.

The probability of an event A is denoted by $P(A)$. One can interpret the operations $A \text{ and } B$, A^c, and $A \text{ or } B$ mathematically as either a Boolean algebra of events or as a field of sets. The two interpretations are equivalent, since every Boolean algebra is isomorphic to a suitable field of sets. We follow Kolmogorov and interpret all events as sets of elementary events. According to this interpretation, the events are subsets of a set E, the set of all elementary events which are considered possible in a given situation, called sample space, $A \cap B$ is the intersection, A^c the complement of A, and $A \cup B$ the union. Events A and B are said mutually exclusive if $A \cap B = \emptyset$, where $\emptyset$ is the empty set.

According to Kolmogorov we can construct a theory of probability from the following axioms:

1. If A and B are events, then A^c, $A \cap B$, and $A \cup B$ are also events.

2. To each event there corresponds a real number $P(A) \geq 0$. E is an event with $P(E) = 1$.

3. If A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$.

4. If A_1, A_2, $\ldots$ are events, which never occur all together, then

$$\lim_{n\to\infty} P(A_1 \cap A_2 \cap \ldots \cap A_n) = 0 \ .$$

From axioms 2 and 3 it follows that

$$P(A^c) = 1 - P(A)\,,$$

and that the highest value of $P(A)$ is 1:

$$0 \leq P(A) \leq 1\,.$$

In particular $P(\emptyset) = 0$ as $E^c = \emptyset$. If A_1, A_2, $\ldots$ are mutually exclusive events, the addition rule

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ \ldots$$

holds.

The **conditional probability** $P(B|A)$ of B under the assumption that A has occured is defined by

$$P(B \cap A) = P(B|A)\, P(A)\ .$$

Event B is independent of A if $P(B|A) = P(B)$, which implies

$$P(B \cap A) = P(B)\, P(A)\ .$$

In practice $P(B|A)$ is rarely calculated from the knowledge of $P(B \cap A)$ and $P(A)$, but instead $P(B \cap A)$ from $P(B|A)$ and $P(A)$. Also, as $B \cap A = A \cap B$, we have

$$P(A|B)\, P(B) = P(B|A)\, P(A) \qquad (1)$$

for the conditional probability of A under the assumption that B has occured.

Example: Consider the sample space of a die, $E = \{1, 2, 3, 4, 5, 6\}$, and the events $A = \{3, 4, 5, 6\}$ and $B = \{2, 3\}$. The die is supposed to deliver a uniform distribution, so that $P(A) = 4/6$ and $P(B) = 2/6$. Assume that A has occured, then the conditional probability is $P(B|A) = 1/4$ (the likelihood for the $\{3\}$ in A). As $A \cap B = \{3\}$, $P(A \cap B) = 1/6 = P(B|A)\, P(A)$.

Total probability:

Events $A_1, A_2, \ldots, A_n$ form a partition of the sample space E if

1. They are mutually exclusive, $i.e.$, $A_i \cap A_j = \emptyset$ for $i \neq j$, and
2. Their union is the sample space, $E = \cup_{i=1}^{n} A_i$.

Let the event of interest be B and assume that the conditional probabilites $P(B|A_i)$ are known. Then $P(B)$ can be calculated by the total probability formula

$$P(B) = \sum_{i=1}^{n} P(B|A_i)\, P(A_i)\ . \tag{2}$$

Example: Two-Headed Coin. Assume that out of N coins in a box, one has heads on both sides. A coin is selected at random from the box and (without inspecting it) flipped k times. Denote the event that the randomly selected coin lands heads up k times by B_k. What is the probability that the coin is two-headed?

Answer: Let us denote the event that the coin is two-headed by A_1 and that the coin is fair by A_2. Obviously, we have the *a-priori* probabilities $P(A_1) = 1/N$ and $P(A_2) = (N-1)/N$. The conditional probabilities for B_k are $P(B_k|A_1) = 1$ and $P(B_k|A_2) = 2^{-k}$. Hence, by the total probability formula

$$P(B_k) = \frac{1}{N} + \frac{N-1}{2^k N} = \frac{2^k + N - 1}{2^k N} \,. \tag{3}$$

Therefore, the probability that we picked the two-headed coin is

$$P(A_1|B_k) = \frac{2^k}{2^k + N - 1} \tag{4}$$

as follows from

$$P(A_1|B_k)\,P(B_k) = P(B_k|A_1)\,P(A_1) = \frac{1}{N} \tag{5}$$

and expression (3) for $P(B_k)$.

Another example:

Consider the following game: In one of three boxes A, B, C, an award of \$ 900 is waiting. For a fee of \$ 300 the contestant will pick one of the three boxes. Then, one of the remaining boxes is opened and found empty. Afterwards the contestant is allowed to change his choice for an additional fee of \$ 100. Should she or he do that?

Bayes Theorem

The interchange identity (1) for conditional probabilities together with the total probability formula (2) incorperate in essence **Bayes Theorem:**

Let the event B happen under any of the possibilities A_i, $i = 1, \dots, n$, $A_1 \cup A_2 \cup \dots \cup A_n = E$, $A_i \cap A_j = \emptyset$ for $i \neq j$, with known conditional probabilities $P(B|A_i)$. Assume, in addition, that the probabilities $P(A_i)$ are known. Then the conditional probability of A_i, given that B has happened, is

$$P(A_i|B) \;=\; \frac{P(B|A_i)\,P(A_i)}{P(B)} \;=\; \frac{P(B|A_i)\,P(A_i)}{\sum_{i=1}^{n} P(B|A_i)}\;.$$

In the context of this equation the events A_i are often called **hypotheses**, the $P(A_i)$ are the **prior** and the $P(A_i|B)$ are the **posteriori probabilities.**

Example (was homework): A Medical Test.

The following question was posed to students at the Harvard Medical School: A test for a disease whose prevalence is $1/1\,000$ fails never when the person is infected, but has a false positive rate of 5%. A person is picked randomly from the population at large and tests positive. What is the probability that this person actually has the disease?

Let A be the event that the patient has the disease and B the event that the test result is positive. The conditional probability is

$$\begin{aligned} P(A|B) &= \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P(B|A)\,P(A^c)} \qquad (6) \\ &= \frac{1 \times 0.001}{1 \times 0.001 + 0.05 \times 0.999} \approx 0.02\ . \end{aligned}$$

A physicists shortcut to that: Let us test 1001 people. Then, 50 test positive by chance and one by infection. So, the likelihood to be infected is $1/51 \approx 0.02$.

Assume the false positive rates are purely due to chance ($i.e.$, no dependence on conditions of the persons tested). Under the assumptions stated above, how often needs the test to be repeated, so that we are sure with at least 99.9% probability that a person is infected? What is then the expected number of tests needed for a sample of $1\,001$ persons?

First question: The failure rate decreases like $(1/20)^n$, where n is the number of tests. Let B_n be the event that a person tests n times positive. We want

$$0.999 < \frac{0.001}{0.001 + (1/20)^n \times 0.999}\ .$$

Solving for the smallest n gives $n_{\rm min} = 5$.

Second question (physicist's answer): After the intial $1\,001$ tests we are (in average) down to 51 persons. After the second series of tests down to $51/20 \approx 3$. So, altogether about $1\,001 + 51 + 3 = 1\,055$ tests are needed. Add two more tests for each person still positive after three tests. At the end of story the above test is quite powerful, because it can show with certainty that a person is not infected.

Example: Prosecutor's Fallacy.

Let A be the event "innocent" and B the evidence. The prosecutor's fallacy is a subtle exchange of $P(A|B)$ (the probability to be innocent, given the evidence) by $P(B|A)$ (the probability of the evidence, given to be innocent). Assume that the probability to match some evidence (e.g., fingerprints) by chance is 10^{-6} and that an actual match is found in an archive of two million candidates. A zealous prosecutor may then argue that the probability of the accused being innocent is one in a million, while the correct probability is obtained by applying (6). Here $P(B|A) = 10^{-6}$, $P(A^c) = 0.5 \times 10^{-6}$ (in lack of other evidence this is an upper bound for the prior probability, as the guilty person may or may not be found in the archive), $P(A) = 1 - P(A^c) \approx 1$, and $P(B|A^c) = 1$. Therefore, we have

$$P(A|B) = \frac{10^{-6} \times 1}{10^{-6} \times 1 + 1 \times 0.5 \times 10^{-6}} = \frac{2}{3},$$

which is quite distinct from 10^{-6}.

As a real-life example the convinction of Sally Clark in the U.K. is sometimes cited. She was accused of having killed her first child at the age of 11 weeks and her second child at the age of 8 weeks. The prosecution had an expert witness testify that the probability of two children dying from sudden infant death syndrome is about 1 in 73 million. Sally Clark was convicted in 1999 and finally aquitted four years later.