

Probability Numbers

For many statisticians the concept of “the probability that an event occurs” is ultimately rooted in the interpretation of an event as an outcome of an experiment, others would interpret the concept as a subjective degree of belief that the event occurs. This difference in interpretation has been the source of great debate between Frequentists (the former group) and Bayesians (the latter group) within the statistics profession. This philosophical debate ultimately influences the way that empirical results are obtained and interpreted, but does not affect the use of probability ideas employed in describing the distribution of happiness in a society. Hence we shall proceed under the Frequentist interpretation whilst acknowledging the existing of an alternative.

At the heart of the frequentists view of probability is the notion of an experiment. A sufficiently general definition of the term “Experiment” permits an almost universal application of the probability concept. Essentially a procedure must possess two properties to be eligible as an experiment.

- 1) **It should be notionally repeatable an infinite number of times with a well defined common set of possible outcomes each time the procedure takes place.**
- 2) **There should be uncertainty as to which outcome will occur before the procedure takes place.**

Probability Theory is best understood by using ideas from set theory in its description. The appendix provides an outline of the basic set theory ideas that are used.

The set of mutually exclusive (having nothing in common) and exhaustive (a complete list of) possible **Basic Outcomes** (denoted o_i) of an experiment is usually called **The Sample Space** (denoted S). An **Event** is defined as any subset of this sample space, including the empty set and the sample space itself, generally an event is denoted by an upper case letter say A , A^c , the complement of A , then corresponds to A not happening. An event is said to have occurred when any one of the basic outcomes in its defining subset is realized. Thus, on executing the procedure, the empty or null set is the event “no outcome occurs” which is certain not to happen, similarly the sample space set (the universal set in set theoretic terms) is the event “an outcome occurs” which is certain to happen. Given outcome uncertainty there is obviously event uncertainty ranging in degree between the empty set and the sample space.

Numbers attached to events reflecting the degree of certainty with which they occur, and which in turn obey certain coherency axioms, are called “Probabilities.” The coherency axioms are simple yet remarkably powerful, they provide the basis for all probability theory regardless of its objective or subjective foundations. Denoting the occurrence of the i 'th basic outcomes as o_i and the probability of it happening as $P(o_i)$ and denoting

events by upper case letters (with S and \emptyset respectively reserved for the sample and empty sets) the coherency axioms are as follows:

- 1) **Probability numbers are non-negative.** Using our notation this may be written as:

$$P(o_i) \geq 0 \text{ for all } i$$

- 2) **The probability of an event is the sum of the probabilities of the mutually exclusive basic outcomes that the event comprises.** Using our notation this may be written as:

$$P(A) = \sum_{o_i \in A} P(o_i)$$

- 3) **The probability that something happens is 1.** using our notation this may be written as:

$$P(S) = 1$$

From these axioms many implications follow, specifically they are that:

- i. **The probability of an event occurring is one minus the probability of it not occurring** (making the probability of nothing occurring zero).

Set Theory tells us that for any event A, $A \cap A^c = \emptyset$ and $A \cup A^c = S$. Axiom 2 tells us that $P(A \cup A^c) = P(S) = P(A) + P(A^c)$, axiom 3 tells us that $P(S) = 1$ and so it follows that $P(A) = 1 - P(A^c)$. If we let $A = S$ it follows that $P(\emptyset) = 0$.

- ii. **If event A includes event B (which may be written as $B \subset A$) then $P(A) \geq P(B)$.**

Note that $A = B \cup (A \cap B^c)$ and $B \cap (A \cap B^c) = \emptyset$. Axiom 2 tells us that $P(A) = P(B) + P(A \cap B^c)$ and Axiom 1 tells us that $P(A \cap B^c) \geq 0$ so $P(A) \geq P(B)$.

- iii. **For any event B, $P(B) \leq 1$.**

This can be seen by simply letting A in ii) be the sample space S and nothing that Axiom 3 tells us that $P(S) = 1$.

- iv. **The probability of any one of a collection of mutually exclusive events occurring is the sum of their individual probabilities.**

This is just a repeated application of Axiom 2.

- v. For any two events A and B, the probability of either one or both of them occurring is the sum of their probabilities minus the probability that they both occur.

This is possibly the most complicated idea to justify. First note that $A = (A \cap B) \cup (A \cap B^c)$ and $(A \cap B) \cap (A^c \cap B) = \emptyset$ so that $P(A) = P(A \cap B) + P(A \cap B^c)$ from Axiom 2, in a similar fashion $P(B) = P(A \cap B) + P(A^c \cap B)$ can be established. Secondly note that $A \cup B = (A \cap B) \cup (A \cap B^c) \cup (B \cap A^c)$ where $(A \cap B)$, $(A \cap B^c)$ and $(B \cap A^c)$ are all mutually exclusive sets so that $P(A \cup B) = P(A \cap B) + P(A \cap B^c) + P(B \cap A^c)$ also from Axiom 2. Comparing the equations for $P(A)$, $P(B)$ and $P(A \cup B)$ it can be seen that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Conditional and Marginal Probability and the notion of Independence

Within the above set of implications the notion of jointly occurring events has already been entertained. Here it is extended so that the concepts of conditional and marginal probability and independence can be understood. We suppose the sample space is covered by a collection of mutually exclusive and exhaustive events A_i , $i=1, \dots, n$ and another collection of mutually exclusive and exhaustive events B_j , $j=1, \dots, m$. for example the sample space may be a collection of people and the events A_i refer to income categories and B_j to occupational categories or the sample space may be a regular deck of cards (excluding jokers) and the events A_i refer to the card values and the events B_j refer to the suits. A notional experiment may be the random selection of an individual or a card, the element selected will have both characteristics prompting questions as to how will they jointly relate to the probability of the element being selected.

The basic building block is the joint probability of selecting an element which jointly exhibits characteristics A_i and B_j (written $P(A_i \cap B_j)$). To obey the above axioms the following must be true:

$$P(A_i \cap B_j) \geq 0 \text{ for all } i, j.$$

$$\sum_{i=1}^n \sum_{j=1}^m P(A_i \cap B_j) = 1$$

In the context of our examples the value $P(A_i \cap B_j)$ is the answer to questions like “what is the probability of selecting someone in the poverty group who is a lawyer?” or “what is the probability of selecting the four of clubs?”

If characteristics of only one type are of concern the probability of that characteristic, known as the marginal probability, may be calculated as follows:

$$P(A_i) = \sum_{j=1}^m P(A_i \cap B_j), \quad P(B_j) = \sum_{i=1}^n P(A_i \cap B_j)$$

Intuitively, in the case of the marginal probability of “A” events, the probability of event (A_i and “all possible B events”) is calculated and in the case of the two examples answers the questions “what is the chance of selecting someone in the poverty group regardless of their occupation?” or “what is the chance of drawing an ace regardless of its suit?”. Note that marginal probabilities also obey all of the probability axioms.

Slightly more complex questions of the form “given the person selected is a lawyer, what is the chance she is in the poverty group?” or “having been reliably informed that the card drawn is an ace, what is the chance it is a diamond?” require the concept of conditional probability. What is really happening is that the sample space is being modified or constrained by the introduction of information upon which the event is now predicated or “conditioned”. As long as the conditioning event has a non-zero probability (is real information in some sense) a cogent set of conditional probability numbers can be calculated by dividing the joint probability of the two events by the marginal probability of the conditioning event. Writing the conditional probability of A_i given B_j as $P(A_i|B_j)$ the appropriate calculation is:

$$P(A_i | B_j) = \frac{P(A_i \cap B_j)}{P(B_j)}$$

Independence between two events is readily defined and interpreted in the context of conditional probability. Events A_i and B_j are independent if

$$P(A_i \cap B_j) = P(A_i)P(B_j)$$

Note that this only refers to the specific characteristics A_i being independent of specific characteristic B_j . For all characteristics A to be independent of characteristics B the independence condition must hold for all possible pairs i, j . From the definition of conditional probability this implies:

$$P(A_i|B_j) = P(A_i), \quad P(B_j|A_i) = P(B_j)$$

Which may readily be interpreted as the occurrence (or otherwise) of event B does not influence the probability of A occurring. Most importantly the independence concept can be extended to a collection or sequence of many events so that their mutual independence implies that the probability of their joint occurrence will be the product of their individual probabilities, formally:

$$P(A \cap B \cap C \cap \dots) = (P(A)P(B)P(C)\dots)$$

Bayes Rule

This rule follows directly from the relationships between conditional, marginal and joint probabilities and may be stated as:

The Bayes Rule:

$$P(A | B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Probability Functions

The mechanism for attaching probability numbers to events is the probability function. Generically there are two types of probability functions, which one is employed depends upon the nature of the sample space it relates to. Some sample spaces contain a finite or countably infinite number of basic outcomes and are covered by “Discrete Probability functions”, others contain an uncountably infinite number of basic outcomes and are covered by “Continuous Probability Functions”. In the discrete case, where ω is any basic outcome in the sample space S , any $P(\omega)$ that satisfies $P(\omega) \geq 0$ for all such ω and $P(S) = 1$ is a Discrete Probability Distribution. In the continuous case the situation is slightly more complex and is greatly facilitated by assuming that the sample space is an interval on the real line containing a typical point “ x ”. In this case the continuous probability function is defined implicitly as any function $f(x)$ that, for any event A defined on the sample space S , satisfies:

$$\begin{aligned} f(x) &\geq 0 \\ \int_S f(x)dx &= 1 \\ P(A) &= \int_A f(x)dx \end{aligned}$$

So that Discrete Probability Distributions attach probabilities to points, Continuous Probability Distributions attach probabilities to intervals.

Appendix. Basic Set Theory Concepts

A set is simply a collection or list of objects or things called elements, the elements are referred to with lower case letters and the sets are referred to with upper case letters.

$$a \in A$$

May be read as saying “the element a belongs to the set A” similarly:

$$a \notin A$$

May be read as “the element a does not belong to A”.

The set of all elements of interest is referred to as the Universal Set, here it will be denoted by the letter S since in the realm of probability theory the Universal Set is the set of all the possible outcomes of an experiment or the sample Space. The empty set or Null Set (which has no elements in it) is referred to with the character \emptyset . The complementary set of A which is denoted A^c is the set of all elements in S that are not in A.

Two basic operations are employed with sets, Set Union and Set Intersection.

Set Union

The Union of two sets A and B is written as $A \cup B$ and is the set of all elements in either A or B (Or both).

Set Intersection

The Intersection of two sets A and B is written as $A \cap B$ and is the set of all elements that are in both A and B.

Note that both union and intersection operations can be employed successively on more than two sets so that $A \cup B \cup C$ is the set of elements that are in any one of A, B or C and $A \cap B \cap C$ is the set of elements that are common to A, B, C.

Some basic Set Theory Ideas

Mutual Exclusivity

Two sets A and B are said to be mutually exclusive when they have no elements in common so that $A \cap B = \emptyset$, and similarly $A \cap A^c = \emptyset$.

Subsets

When all the elements in set A are in set B, A is said to be a subset of B, which is written as $A \subset B$. It follows that all sets are subsets of the sample space so that $A \subset S$ for any A.

Some Rules for Operations

Associative Rule

$$\begin{aligned}A \cup (B \cap C) &= (A \cup B) \cap C \\A \cap (B \cup C) &= (A \cap B) \cup C\end{aligned}$$

Distributive Rule

$$\begin{aligned}A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\A \cup (B \cap C) &= (A \cup B) \cap (A \cup C)\end{aligned}$$

Complementation Rule

$$\begin{aligned}(A \cap B)^c &= A^c \cup B^c \\(A \cup B)^c &= A^c \cap B^c\end{aligned}$$