Chapter 2

Probability

2.1 Axioms of Probability

2.1.1 Frequency definition

A mathematical definition of probability (called the *frequency definition*) is based upon the concept of data collection from an experiment. Suppose an experiment is set up to observe a random process, and that the experiment could (in principle) be repeated many times, each time under identical conditions. If the experiment (observation) had been performed N times, and the event A had occurred a number of times $N(A)$ then one could sensibly define the the probability of the event as:

$$P(A) = \lim_{N\to\infty} \frac{N(A)}{N} \qquad , \tag{2.1}$$

and since, $0 \leq N(A) \leq N$, we have $0 \leq P(A) \leq 1$.

This leads us to define an *impossible event*, $A = \emptyset$, as one that will never occur:

$$N(A) = N(\emptyset) = 0 \qquad \text{and} \qquad P(A) = 0 \qquad .$$

Conversely, we could define a *certain event*, A, as one that happens every time, $A = \Omega$, and therefore

$$N(A) = N, \qquad \text{we have} \qquad P(A) = 1 \qquad .$$

While this definition makes sense, it is rather impractical and mathematically dubious. From the practical viewpoint, there are very few systems that lend themselves to repeated experiments under identical conditions. Even with apparently simple systems, such as a six-sided die, each roll will be executed differently by the person rolling the die. Even if we are able to execute $N \gg 1$ experiments, by requiring 'identical conditions', we are imposing strict guidelines on our experiments.

If such a simple, controllable experiment is open to question, this does not augur well for real problems for which we have very little data (N is not infinite) or the conditions are not identical. Even if we can overcome the practical problems of designing an experiment that meets these strict conditions, there is a more serious mathematical problem, namely the existence (or not) of the limit. If the system is governed by underlying random variables that are identical and independent, for example tossing a coin repeatedly to observe the frequency of 'heads' (H), there will always be fluctuations in $N(H)$. So any 'limit' will not be in the sense of uniform convergence. Nonetheless, in *the long run* ($N \to \infty$), such limits can be defined, but not in the usual sense. This fact is addressed when we look at the *laws of large numbers* - the convergence theorems of probability. Nonetheless, we are still on shaky ground if we cannot make solid definitions.

From a purely Mathematical viewpoint, this is not a concern. We can pose axioms which define probability in a mathematical sense, without any reference to experiment. However, at some point we would like to connect the Mathematics with applications including the use of data.

2.2 Mathematical Foundations

Suppose the sets A and B are *disjoint*. Then by definition: $A \cap B = \emptyset$, and the events are mutually exclusive. Therefore, the total number of events of either A or B is given by summing:

$$N(A \cup B) = N(A) + N(B) \qquad .$$

Dividing this relation by N, we get:

$$\frac{N(A \cup B)}{N} = \frac{N(A)}{N} + \frac{N(B)}{N} \qquad ,$$

and taking the limit $N \to \infty$, we have for *mutually exclusive* events, $A \cap B = \emptyset$:

$$\boxed{P(A \cup B) = P(A) + P(B)} \tag{2.2}$$

For example, consider the roll of a single die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. The events $A = \{1, 2, 3\}$ and $B = \{6\}$ are obviously *mutually exclusive*. Then the event $C = A \cup B = \{1, 2, 3, 6\}$ has probability, using relation (2.2),

$$P(C) = P(A \cup B) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$$

2.2.1 Probability space

Suppose we have a sample space, Ω, and a σ-field, $\mathcal{F}$, of subsets of Ω, then we can define a *probability measure* (or *probability distribution*) P, that maps (converts) the σ-field to the set of real numbers in the closed interval: $[0, 1]$.

$$P : \mathcal{F} \to [0, 1] \tag{2.3}$$

That is for each event we can ascribe a number between (and including) 0 and 1 that represents the probability of the event. Then the three quantities ('the triple'): $(\sigma, \mathcal{F}, P)$ form a *probability space*.

The axioms that define P are:

(a) $P(\emptyset) = 0$, $P(\Omega) = 1$.

(b) $0 \leq P(A) \leq 1$, for any event A.

(c) If $A \cap B = \emptyset$: $\qquad P(A \cup B) = P(A) + P(B)$

Set/Probability synomyns and analogues

Mathematical notation	Set meaning	Probability meaning
Ω	Set (collection) of objects	Sample Space
$\emptyset$	Empty (null) set	Impossible event
ω	Element (member) of Set, Ω	Elementary Event, Outcome
A	Subset of Ω	Event such that the outcome is A
A^c	The complement of A	Event such that the outcome is NOT A
$A \cap B$	Intersection: A and B	Both events A and B
$A \cup B$	Union: A or B	Either event A or B or both
$A \setminus B$	Difference: set A minus set B	Event A but not event B
$A \triangle B$	Symmetric difference	Either A or B but not both
$A \subset B$	Inclusion: A is a subset of B	If event A occurs then so too does B

2.3 Null events and measure zero

There is a mathematical distinction between an *impossible event* and a *null event*. To a statistician this is splitting hairs, but to a mathematician it is a fundamental feature.

An event is called *null* if: $P(A) = 0$.

Now it is always true that, $P(\emptyset) = 0$. where $\emptyset$ is the 'impossible event'. However, given $P(A) = 0$, this does *not* imply that $A = \emptyset$. If A is an event that is possible, but *extremely* unlikely, then in effect the probability of the event is zero. In this case we say A is a set of measure/probability zero.

Similarly we have a formal difference between the *certain event* and the *almost surely event*. An event is said to occur 'almost surely' (often seen abbreviated as 'a.s.' after an equation) is this case:

$$P(A) = 1 \qquad (a.s.) \qquad .$$

An event is said to be 'almost certain' if it is equal to the sample space Ω, except for sets of 'measure zero'. That is, the event will occur, with the exception of a few almost impossible events.

Let's give an example of an *almost impossible event*. Consider a football match and we are interested in the probability of there being exactly 60 goals in the game. Now, while such an event is (theoretically) possible, it is extremely unlikely. Another example would be the probability that the first goal is scored after 6.22234 minutes (to the nearest 100,000th of a minute). Here again, the event space is so restricted and extreme a part of the sample space (one 100,000th of a minute) that such an event (while possible) is extraordinarily unlikely. So both a statistician (and even some mathematicians) would be comfortable with saying that the probability of such an event is zero.

2.4 Lemmas

The algebra of probability can be based on the three axioms stated. Indeed this course is based upon building a skyscraper of results on top of these foundations. So consider some of the first elementary results (lemmas) that follow from the axioms:

(a) $P(A^c) = 1 - P(A)$.

(b) If $A \subset B$:

$$P(B) = P(A) + P(A \setminus B) \geq P(A).$$

(c) The probability version of the *inclusion-exclusion theorem*:

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)} \qquad . \tag{2.4}$$

The *inclusion-exclusion theorem* (2.4) is a very important result. It can extended to many events $\{A_1, A_2, \ldots, A_n\}$ as follows:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j}\sum_{j} P(A_i \cap A_j) + \sum_{i<j}\sum_{j<k}\sum_{k} P(A_i \cap A_j \cap A_k) + \cdots \tag{2.5}$$

$$\cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n) \tag{2.6}$$

Proof of lemma (a): We note that A and A^c are disjoint sets (by definition). Thus, according to axiom (c), for any pair of disjoint sets:

$$P(A \cup B) = P(A) + P(B)$$

Then, letting $B = A^c$, we note that $A \cup A^c = \Omega$. Thus:

$$P(\Omega) = P(A) + P(A^c).$$

According to axiom (a), we have $P(\Omega) = 1$, thus:

$$1 = P(A) + P(A^c) \qquad \Rightarrow \qquad P(A^c) = 1 - P(A).$$

2.5 Probability as uncertainty

In practical applications it is necessary to deal with complex processes. Very few problems fall into the category of repeatable experiments under identical conditions. In a realistic problem, we may only have one opportunity to observe the event. In addition, the problem will usually have large elements of uncertainty (lack of knowledge of the determining factors) and indeed poor quality data (missing or erroneous measurements). So even if we had a large amount of data, it would be very difficult to determine what is deterministic and what is random, what is 'clean' data and what is 'dirty' data.

However, rather than abandoning all hope, suppose we are prepared to widen our scope of interest to consider such problems. That is problems which are not random at all, but purely uncertain.

2.5.1 Implied Probability

In contrast to flipping a coin, in most practical cases, it is very difficult to estimate the probability of an event happening. However this does not prevent us from making rough guesses based on our perception. Examples commonly arise in financial markets, in which the prices express the probabilities are *implied* by the market.

(a) GAMBLING MARKET

In a tennis match between player A and player B, there are only two outcomes (excluding the cancellation of the match); player A loses or player A wins.

In a gambling market the probabilities of these events are expressed by the 'odds' assigned to each event by the bookmakers. The odds are expressed as a ratio of two numbers in the form a/b (or equivalently a–b)

For example, the odds for a match were given as follows: player A wins 1/4, player B wins 11/4. In the language of gambling, this is stated as *four-to-one on* and *eleven-to-four against*. What this means is that, if A wins and you bet £1 on this event you receive a profit of £1/4= £0.25. While if you bet £1 on B and this is successful, you receive £11/4 = £2.75 profit.

Mathematically, this ratio a/b can be ascribed the equivalent *implied probability*:

$$P = \frac{b}{(a+b)} \quad .$$

If we consider the event W, as A winning, then W^c, is the event B wins. Thus

$$P(W) = \frac{4}{5} = 0.800 \qquad P(W^c) = \frac{4}{15} \approx 0.267 \quad .$$

Implied probabilities of this type do not add to 1! For the example above we see that:

$$P(W) + P(W^c) = 1.067 \quad .$$

There is a very good reason for this, it helps ensure the bookmaker makes a profit. If, however, the bookmaker set the odds such that:

$$P(W) + P(W^c) < 1 \quad ,$$

then he/she has made a serious mistake. In these circumstances we can ensure a profit (and that the bookmaker loses), whatever the outcome by betting on both outcomes (called 'arbitrage' and described in the examples).

The bookmaker's odds are simply a reflection of the quantity of money wagered by the public on the event. That is, the probabilities are based on the public perception (market) of the likelihood of each event. These probabilities are not based on the true (mathematical) probabilities. When the market (implied) probabilities are approximately equal to the true (mathematical) probabilities, we say the that market is 'efficient': that is the market is a reliable estimate of the true probabilities.

(b) INSURANCE MARKET

Suppose you buy a new phone, costing $a =$ £100 from a shop. The shop assistant mentions that 1 in 10 phone owners have had their phone lost/stolen or damaged requiring replacement in the first year, and offers you insurance at $b =$ £3 per month (called the premium).

In effect, this quote is an implied probability that you will lose or break your phone in the next month with probability $P = 3/100 = 0.03$. If you buy this insurance you are, in effect, estimating that the probability you lose you phone in the next month is 3% or more. Even more confusing is the claim made by the sales assistant that the 'probability' of losing your phone within the year is 10% (1/10). There is a third probability; the probability that you (as an individual) will lose/break your phone.

We say that the cost of insurance is a 'fair price' if the *implied probability* (in this case 0.03) is equal to the true probability. Of course, no one knows the true probability, although the insurance company might be able to estimate this based on previous claims. Now, if your chance of losing the phone is less than 0.03, then the price is unfair - in favour of the insurance company. That doesn't mean you won't buy the insurance anyway!

Like any market, the customer aims for the lowest price available. Conversely, the insurance company must avoid quoting prices at values so low that they risk going bankrupt through claims (payouts) exceeding the premiums (income). Insurance prices are dictated more by the market (the price someone is willing to sell/buy) as actuarial tables (statistical data).

This is the case in car insurance as much as it is in 'derivatives' pricing in financial capital markets. This is simply because the probability (in most cases) defies calculation. That does not mean to say that one cannot have a go at 'calculating the risk' and thus determining the true probability.

2.5.2 Degrees of belief

Following this principle of uncertainty, one can assign probabilities according to the *degree of belief* that an event will happen. This philosophy of probability as a degree of uncertainty, as opposed to the more familiar random process, is a radically different way of thinking. It is called the Bayesian approach, since these estimates of probability are not pure guesswork, but rather conditional probabilities based on information. Unearthing the connection between known information and uncertainties of events is at the heart of the Bayesian approach to data analysis.

The upside to this Bayesian method is that one can apply probability to any number of scientific problems. The downside is that these probabilities (based on incomplete knowledge) may be entirely subjective, speculative and, at worst, completely wrong. There is another 'fly in the ointment' with this approach. If for example, in your opinion, the probability of your bus arriving in the next 5 minutes is 10%, in order to be consistent then you must logically conclude that the probability that it does not arrives in the next 5 minutes is 90%. The requirement that probability estimates (whatever their veracity) should be self-consistent is called *coherence* in this context. The more probability estimates you make, the more effort is required to tidy up the consequences of these guesses.

Recall that probability is imbedded in the estimates of risk, for example the price of assets and financial instruments. If the probability estimates (asset prices) are not *coherent* then a risk-free *arbitrage* opportunity may exist. In a market of many assets and instruments, correlations between these prices will often occur. Any imbalances, or incoherence, in these prices can be exploited for risk-free trading (*arbitrage*). The trading strategy based on this technique is called *statistical arbitrage*, and is very much in vogue with 'hedge funds'.