UNIVERSITY OF LONDON

Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

2.1 Probability principles

Probability is very important for statistics because it provides the rules which allow us to reason about uncertainty and randomness, which is the basis of statistics and must be fully understood in order to think clearly about any statistical investigation.

The first basic concepts in probability will be the following:

- **Experiment**: For example, rolling a single die and recording the outcome.
- **Outcome** of the experiment: For example, rolling a 3.
- **Sample space** S: The *set* of all possible outcomes, here $\{1, 2, 3, 4, 5, 6\}$.
- **Event**: Any *subset* A of the sample space, for example $A = \{4, 5, 6\}$.

Probability, $P(A)$, will be defined as a *function* which assigns probabilities (real numbers) to events (sets). This uses the language and concepts of **set theory**. So we need to study the basics of set theory first.

A **set** is a collection of **elements** (also known as 'members' of the set).

Example

The following are all examples of sets, where '|' can be read as 'such that':

- $A = \{\text{Amy, Bob, Sam}\}$
- $B = \{1, 2, 3, 4, 5\}$
- $C = \{x \,|\, x \text{ is a prime number}\} = \{2, 3, 5, 7, 11, \dots\}$
- $D = \{x \,|\, x \geq 0\}$ (that is, the set of all non-negative real numbers).

We consider four basic concepts in probability.

An **experiment** is a process which produces outcomes and which can have several *different* **outcomes**. The **sample space** S is the set of all possible outcomes of the experiment. An **event** is any subset A of the sample space such that $A \subset S$, where $\subset$ denotes a subset.

Example

If the experiment is 'select a trading day at random and record the % change in the FTSE 100 index from the previous trading day', then the outcome is the % change in the FTSE 100 index.

$S = [-100, +\infty)$ for the % change in the FTSE 100 index (in principle).

An event of interest might be $A = \{x \,|\, x > 0\}$ – the event that the daily change is positive, i.e. the FTSE 100 index gains value from the previous trading day. We would then denote the probability of this event as:

$$P(A) = P(\% \text{ daily change is positive}).$$

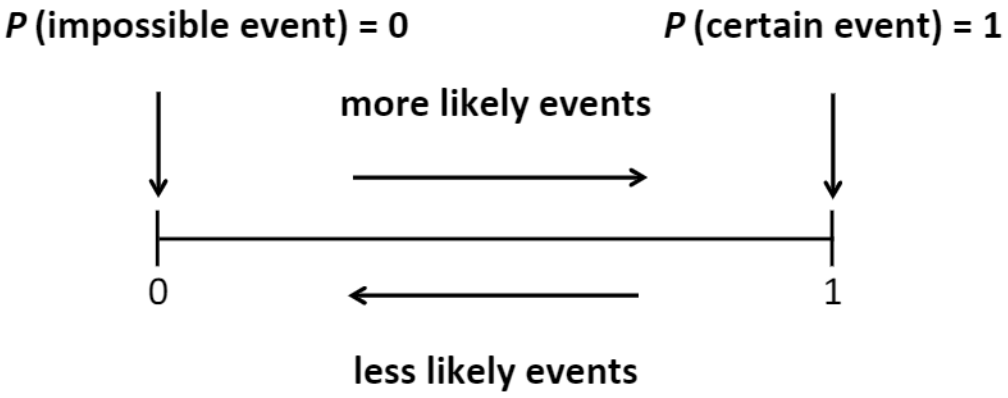
What does 'probability' mean?

Probability theory tells us how to work with the probability function and derive 'probabilities of events' from it. However, it does not tell us what 'probability' really means.

We define probabilities to span the **unit interval**, i.e. [0, 1], such that for any event A we have:

$$\mathbf{0 \leq P(A) \leq 1.}$$

At the extremes, an *impossible* event occurs with a probability of zero, and a *certain* event occurs with a probability of one, hence $P(S) = 1$ by definition of the sample space. For any event A, $P(A) \to 1$ as the event becomes more likely, and $P(A) \to 0$ as the event becomes less likely. Therefore, the probability value is a quantified measure of how likely an event is to occur.



There are several alternative interpretations of the real-world meaning of 'probability' in this sense. One of them is outlined below. The mathematical theory of probability and calculations on probabilities are the same whichever interpretation we assign to 'probability'.

Frequency interpretation of probability

This states that the probability of an outcome A of an experiment is the proportion (**relative frequency**) of trials in which A would be the outcome if the experiment was repeated a very large number of times under similar conditions.

Example

How should we interpret the following, as statements about the real world of coins and babies?

- 'The probability that a tossed coin comes up heads is 0.5.' If we tossed a coin a large number of times, and the proportion of heads out of those tosses was 0.5, the 'probability of heads' could be said to be 0.5, for that coin.
- 'The probability is 0.51 that a child born in the United Kingdom today is a boy.' If the proportion of boys among a large number of live births was 0.51, the 'probability of a boy' could be said to be 0.51.

How to find probabilities?

A key question is how to determine appropriate numerical values, $P(A)$, for the probabilities of particular events.

In practice we could determine probabilities using one of three methods:

- **subjectively**
- **by experimentation (empirically)**
- **theoretically**.

Subjective estimates are employed when it is not feasible to conduct experimentation or use theoretical tools. For example, although:

$$0 \leq P(\text{World War III starts next year}) \leq 1$$

as it is a probability (so must be between 0 and 1, inclusive), what is the correct value which should be attributed to this? Clearly, we *must* resort to subjective estimates taking into account relevant geopolitical events etc. Of course, the probabilistic evaluation of such information is highly subjective, hence different people would assess the chance of this event happening with

different probabilities. As such there is no 'right' answer! That said, you may wish to do some research on the 'Doomsday Clock', which is an attempt to determine how close humanity is to a global catastrophe – what a happy thought!

Ignoring extreme events like a world war, the determination of probabilities is usually done *empirically*, by observing actual realisations of the experiment and using them to **estimate** probabilities. In the simplest cases, this basically applies the frequency definition to observed data.

Example

- If I toss a coin 10,000 times, and 5,050 of the tosses come up heads, it seems that, approximately, $P(\text{heads}) = 0.5$, for that coin.
- Of the 7,098,667 live births in England and Wales in the period 1999 2009, 51.26% were boys. So we could assign the value of about 0.51 to the probability of a boy in that population.

The estimation of probabilities of events from observed data is an important part of statistics!