Lecture 1: Probability Fundamentals

IB Paper 7: Probability and Statistics

Carl Edward Rasmussen

Department of Engineering, University of Cambridge

January 25th, 2011

What is Probability and Statistics?

probability: *a mathematical formalisation describing uncertain events*

statistics: *the practice or science of collecting and analysing data*

Central questions:

- Why do we need this, is it useful?
  - Make inference about uncertain events
  - Form the basis of information theory
  - Test the strength of statistical evidence
- How is it possible to say something about uncertain (or *stochastic*) events?
- How can we *measure* uncertainty (or information)?

Examples

Example: In Premier League football, the probability of a home win is (roughly) 48%, draw 26% and away win 26%. This 48/26/26 rule forms a *summary* of the outcomes.

Example: Three different laboratories have measured the speed of light, with slightly differing results. What is the true speed likely to be?

Example: Two drugs are compared. Five out of nine patients responded to treatment with drug A, where as seven out of ten responded to drug B. What do you conclude?

Examples of places where uncertainty plays a role: medical diagnosis, scientific measurements, speech recognition (human or artificial), budgets, …

Probability is useful, since there is uncertainty everywhere.

Probability

Probability is used to quantify the extent to which an uncertain event is likely to occur.

Probability theory is the calculus of uncertain events.

It enables one to *infer* probabilities of interest based on assumptions and observations.

Example: The probability of getting 2 heads when tossing a pair of coins is 1/4, as probability theory tells us to multiply the probability of the (independent) individual outcomes.

Whereas probability theory is uncontroversial, the *meaning* of probability is sometimes debated.

Statistics is concerned with the analysis of collections of observations.

The Meaning of Probability

In Classical (or frequentist) statistics, the probability of an event is defined as its long run frequency in a repeatable experiment.

Example: The probability of rolling a 6 with a fair die is 1/6 because this is the relative frequency of this event as the number of experiments tends to infinity.

However, some notions of chance don't lend themselves to a frequentist interpretation:

Example: In "There is a 50% chance that the arctic polar ice-cap will have melted by the year 2100", it is not possible to define a repeatable experiment.

An interpretation of probability as a (subjective) *degree of belief* is possible here. This is known also as the Bayesian interpretation.

Both types of probability can be treated using (the same) probability theory.

What is randomness?

An event is said to be random when it is uncertain whether it is going to happen or not.

But there are several possible reasons for such uncertainty. Here are two examples:

inherent uncertainty
: as eg. whether a radioactive atom may decay within some time interval.

lack of knowledge
: I may be uncertain about the number of legs my pet centipede has (if I haven't counted them).

Another important concept is a *random sample* from a population.

Example: An opinion poll was based on telephone interviews of a *representative sample* of 994 voters.

Axioms of Probability

The foundations of probability are based on three axioms:

- The probability of an event E is a non-negative real number

$$p(E) \geqslant 0, \quad \forall E \subseteq \Omega,$$

  where Ω is the sample space.
- The certain event has unit probability

$$p(\Omega) = 1.$$

- (Countable) additivity: for disjoint events $E_1, E_2, \ldots, E_n$

$$p(E_1 \cup E_2 \cup \ldots E_n) = \sum_{i=1}^{n} p(E_i).$$

Remarkably, these axioms are sufficient.

Some consequences

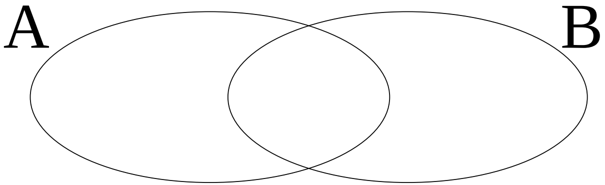
- Complement rule: $p(\Omega - E) \; = \; p(\bar{E}) = 1 - p(E)$.
- Impossible event: $p(\emptyset) \; = \; 0$.
- If $E_1 \subseteq E_2$ then $p(E_1) \leqslant p(E_2)$.
- General addition rule: $p(E_1 \cup E_2) \; = \; p(E_1) + p(E_2) - p(E_1 \cap E_2)$.

As an exercise, prove these!

Generally, we write the probability of the intersection event as $p(E_1 \cap E_2) = p(E_1, E_2)$.

The Venn Diagram and Conditional Probability

Events can sometimes usefully be visualised in a Venn diagram



The intersection of A and B corresponds to both events happening $p(A, B)$.

Assuming that $p(B) \neq 0$, the *conditional* probability of A given that we already know B is defined as

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(A, B)}{p(A, B) + p(\bar{A}, B)}.$$

Example: In the Premier League (using the 48/26/26 rule), the probability of winning is $p(\text{win}) = (1 - 0.26)/2 = 0.37$. The *conditional* probability of winning at home is $p(\text{win}|\text{home game}) = 0.48$.

Example: Inference in Medical Diagnosis

A rare disease occurs with probability $p(D) = 0.01$.

A screening test exists, which detects the disease with probability $p(T|D) = 0.99$. However, the test has a false positive rate of $p(T|\bar{D}) = 0.05$.

A patient takes the test, and the result is positive. What is the probability that she has the disease?

We are looking for $p(D|T)$:

$$
\begin{aligned}
p(D|T) &= \frac{p(D,T)}{p(T)} = \frac{p(T|D)p(D)}{p(T)} = \frac{p(T|D)p(D)}{p(T,D) + p(T,\bar{D})} \\
&= \frac{p(T|D)p(D)}{p(T|D)p(D) + p(T|\bar{D})p(\bar{D})} \simeq \frac{1}{6}.
\end{aligned}
$$

Despite the positive test result, it is still more likely that she does not have the disease.

The first and third terms of the equation above is known as *Bayes rule*.

Random Variables

A random variable is an abstraction of the intuitive concept of chance into the theoretical domains of mathematics, forming the foundations of probability theory and mathematical statistics [wikipedia].

Throughout, we'll use intuitive notions of random variables, and won't even bother defining them precisely.

Sloppy definition: a random variable associates a numerical value with the outcome of a random experiment (measurement).

Example: a random variable X takes values form $\{1, \ldots, 6\}$ reflecting the number of eyes showing when rolling a die.

Example: a random variable Y takes the values in $\mathbb{R}_+$ reflecting measured car velocity in a radar speed detector.

Probability distributions

The *probability function* specifies that the random variable X takes on the value x with a certain probability, written $p(X = x)$.

Example: X represents the number of eyes on a fair die. The probability of rolling a 5 is 1/6, written

$$p(X = 5) = 1/6.$$

The notation $p(X = x)$ is precise but a bit pedantic. Sometimes, we use the shorthand $p(x)$, when it is clear from the context which random variable we are talking about.

The *cumulative probability function*, $F(x)$ is related to the probability function through:

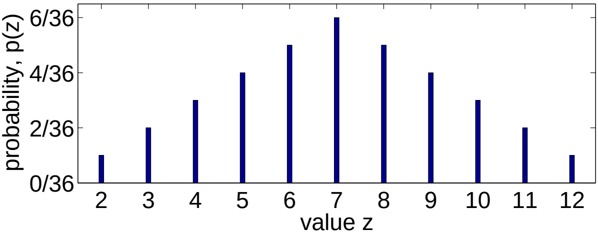
$$F(x) = p(X \leqslant x).$$

Example

Let Z be the sum of the values of two fair dice.

Altogether, there are 36 possible outcomes for the two dice.

For each value of the random variable, the probability is the number of outcomes which agrees with this value of Z divided by the total number of outcomes.



For example, you can get $Z = 4$ in 3 possible ways (1,3), (2,2) and (3,1).

Mean = Average = Expectation

Example: The (arithmetic) average, or mean of a set of numbers: $3, 4, 6, 9$ is

$$\text{mean} \;=\; \frac{1}{4}(3+4+6+9) \;=\; 5.5.$$

In the *mean of a random variable*, the values are *weighted* by their probability

$$\mathbb{E}[X] \;=\; \sum_i p(x_i)x_i \;=\; \langle x \rangle_{p(x)}.$$

The mean is also called the *average* or *expectation*.

Example: The expected number of points from a Premier League home game is:

$$\mathbb{E}[X] \;=\; 0.48 \times 3 + 0.26 \times 1 + 0.26 \times 0 \;=\; 1.7.$$

The mean provides a very basic but useful characterisation of a probability distribution. Examples: average income, life expectancy, radioactive half-life etc.

Example: You can take expectations of *functions* of random variables:

$$\mathbb{E}[f(X)] \;=\; \sum_i p(x_i)f(x_i).$$

Randomness, Surprise and Entropy

How random is something? The *surprise* of an event is given by

$$\text{surprise}(x_i) \;=\; -\log(p(x_i)).$$

The lower the probability, the more surprised we are when the event occurs.

The mean or average surprise is called the entropy and quantifies the amount of randomness

$$\text{entropy}(p) \;=\; \mathbb{E}[-\log(p)] \;=\; -\sum_i p(x_i)\log(p(x_i)).$$

Example: The entropy associated with flipping a fair coin is $-2\frac{1}{2}\log(\frac{1}{2}) = 1$ bit (when the log is taken to base 2).

Example: The entropy of a fair die is 2.6 bits.

Example: A fair coin has more entropy than an unfair one, why?

Note the strong indication that *information* and *probability* are intricately linked.

Something to think about until next week

12 balls look identical, but one is either lighter or heavier than the others.

Use old-fashioned scales (where the outcome is one of: 'right side heavier', 'equal' or 'left side heavier') to find which is the odd ball *and* whether it is heavier or lighter then the others.

What is the minimum number of weighings necessary?

Use the notion of entropy.

Hints are available from the course web page.