Lecture 1: Probability Fundamentals

IB Paper 7: Probability and Statistics

Carl Edward Rasmussen

Department of Engineering, University of Cambridge

January 18th, 2019

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.

The Foundations of Probability: 3 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 three axioms are sufficient. Prove these 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: $p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$.

We write the probability of the *intersection* or *joint* as $p(E_1 \cap E_2) = p(E_1, E_2)$.

Example: Medical inference (diagnosis)

Breast cancer and mammography facts:

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography – this is called the *false positive* rate.

Question: A woman get's a scan, and it is positive; what is the probability that she has breast cancer?

1. less than 1%
2. around 10%
3. around 90%
4. more than 99%

Question: Do you trust your doctor?

Medical inference, conditional probability

Define: C = breast cancer present, $\bar{C}$ = no cancer, M = pos scan and $\bar{M}$ = neg scan

- $p(C) = 1\%$.
- If there is cancer, the probability of a pos mammography is $p(M|C) = 80\%$
- If there is no cancer, we still have $p(M|\bar{C}) = 9.6\%$

The question is: what is the *conditional probability* $p(C|M)$?

Consider 10000 subjects of screening

- $p(C) = 1\%$, so 100 of them have cancer, of which
  - $p(M|C) = 80\%$, so 80 get a positive scan
  - 20 get a negative mammography
- $p(\bar{C}) = 99\%$, so 9900 do not have cancer, of which
  - $p(M|\bar{C}) = 9.6\%$, so 950 get a positive scan
  - 8950 get a a negative mammography

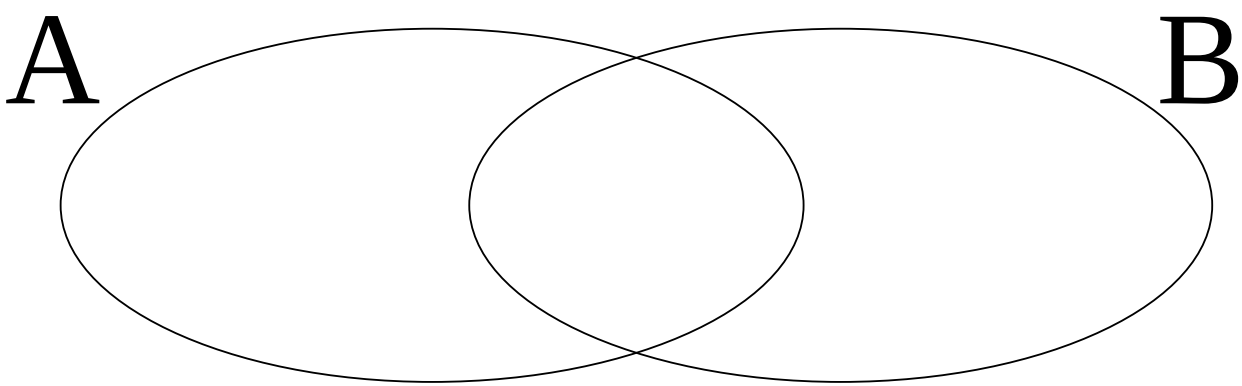
	M	$\bar{M}$
C	80	20
$\bar{C}$	950	8950

$p(C|M)$ is obtained as the proportion of all positive mammographies for which there actually is breast cancer

$$p(C|M) = \frac{p(C,M)}{p(C,M)+p(\bar{C},M)} = \frac{p(C,M)}{p(M)} = \frac{80}{80+950} \simeq 7.8\%$$

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 $p(A, B)$.

Definition: The *conditional probability* of A given that we already know B is defined as

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

This requires $p(B) \neq 0$. We cannot condition on an impossible event; we can't know that something impossible is true.

Just two rules of probability theory

Astonishingly, the rich theory of probability can be derived using just two rules:

The *sum rule* states that

$$p(A) = \sum_B p(A,B), \quad \text{or} \quad p(A) = \int_B p(A,B)dB,$$

for discrete and continuous variables. Sometimes called *marginalization*.

The *product rule* states that

$$p(A,B) = p(A|B)p(B).$$

It follows directly from the definition of conditional probability, and leads directly to Bayes' rule

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

Special case:
if A and B are *independent*, $p(A|B) = p(A)$, and thus $p(A,B) = p(A)p(B)$.

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:

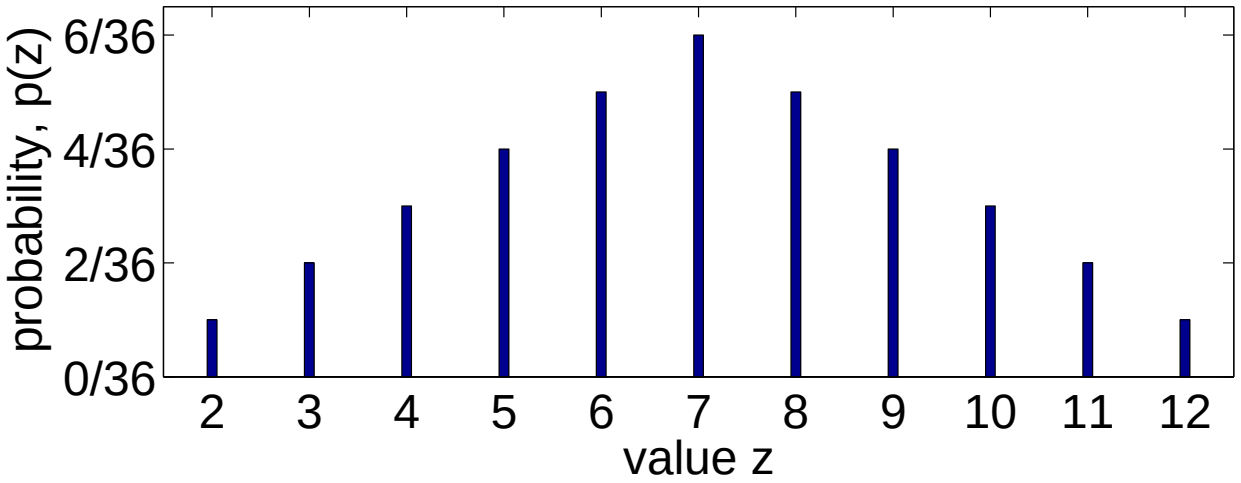
$$F(x) = p(X \leqslant x).$$

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 *average surprise* is called the *entropy* and quantifies the amount of randomness

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

measured in *bits* (binary digits) if the log is base 2, or in *nats* (natural digits) if the log is base e.

Example: The entropy associated with flipping a fair coin is $-2\frac{1}{2}\log_2(\frac{1}{2}) = 1$ bit.

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.