Single Maths B: Introduction to Probability

Overview

Lecturer	Dr Jonathan Cumming
Email	j.a.cumming@durham.ac.uk
Office	CM233
Homework	None!
Webpage	`http://maths.dur.ac.uk/stats/people/jac/singleb/`

1 Introduction to Probability

1.1 Introduction

What is probability?

- Probability the mathematical study of uncertainty
- Probability is a useful concept like *mass* or *energy* and its behaviour is extremely simple. It is attached to events and satisfies some very simple rules.
- Some events can be said to be uncertain – we do not know their outcomes before they occur and we observe what happened.
- Standard mathematics deals only with the certain, so we need some new tools which will allow us to capture, manipulate and reason with this uncertainty
- We begin by quantifying this uncertainty by assigning numbers to each of the possible outcomes to give a measure of "what is likely to happen."
- Larger values will indicate a particular outcome is more likely.
  Lower numbers will indicate an outcome is less likely.

$$\mathrm{P}\left[\text{fair coin lands heads}\right] = \frac{1}{2},$$
$$\mathrm{P}\left[\text{climate change}\right] = ?$$

Why is it useful?

- *Probability can be fundamental to our understanding of the world*
  Quantum mechanics, statistical mechanics, Ising model of magnetism, genetics
- *Probability can used to build models of complex systems or phenomena*
  Epidemics, population growth, chemical interactions, financial markets, routing within networks
- Probability theory leads to the discipline of *statistics*
- *Statistics can be used to analyse data gathered from experiments, and drawing conclusion under uncertainty*
  Important to all the experimental sciences

1.2 Events

- Probability theory is used to describe any process whose outcome is not known in advance with certainty. In general, we call these situations *experiments* or *trials*.
- The set of all possible outcomes of an experiment (or random phenomenon) is the *sample space* S.
- An *event* is a subset of the outcomes in a sample space.

- We treat events as *sets*, and so have three basic operations to combine and manipulate them.

  Event operations

  - Let A, B be some events.
  - The event *not* A is A^c (the *complement*), which is the set of all outcomes in $\mathcal{S}$ and not in A.
  - The event A *or* B is $A \cup B$ (the *union*), which the set of all outcomes in A, or in B or in both.
  - The event A *and* B is $A \cap B$ (the *intersection*), which is the set of all outcomes that are both in A and in B.

Disjoint Events

- Two (or more) events are called *disjoint* (or *incompatible*, or *mutually exclusive*) if they *cannot occur at the same time.*
- The event which contains no outcomes is written $\emptyset$, and is called the *empty set.*
- So if A and B are disjoint, then we must have $A \cap B = \emptyset$

Example: Cluedo

Dr. Black has been murdered! There are four possible suspects: Colonel **M**ustard, Professor **P**lum, Miss **S**carlet, Reverend **G**reen. There are three possible murder weapons: **C**andlestick, **L**ead Piping, **R**ope. There can be only one murderer and one murder weapon.

Working with Events

The following basic set rules will be useful when working with events:

Event Rules

Commutivity:

$$A \cup B = B \cup A, \qquad A \cap B = B \cap A$$

Associativity:

$$(A \cup B) \cup C = A \cup (B \cup C), \qquad (A \cap B) \cap C = A \cap (B \cap C)$$

Distributivity:

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C), \qquad (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

DeMorgan's laws:

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$

1.3 Probability

Axioms of Probability

- We associate a probability with every outcome (and hence every event) in the sample space $\mathcal{S}$.
- For any event A (i.e. any subset of $\mathcal{S}$) we define a number $\mathrm{P}[A]$ which we call the *probability* of A.
- $\mathrm{P}[A]$ is the quantification of our uncertainty about the occurrence of the event A.
- **Note:** A is an event which is a set, $\mathrm{P}[A]$ is a probability which is a number, and $\mathrm{P}[\cdot]$ is a function which maps events to numbers.

  The Axioms of Probability (Komolgorov)

  1. $0 \leq P[A] \leq 1$ – probability is a number in the interval $[0, 1]$.
  2. $P[\mathcal{S}] = 1$ – some outcome from the sample space *must* happen; certain events have probability 1.
  3. If A and B are disjoint events, then $P[A \cup B] = P[A] + P[B]$

- We can think of these three axioms as the "Laws of Probability"

Consequences of the Axioms

- These axioms imply some additional useful properties of probabilities:

  Consequences to the Axioms of Probability

  1. $P[A^c] = 1 - P[A]$.
  2. $P[\emptyset] = 0$ – impossible events have probability zero
  3. In general for any events A and B, $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.
  4. If A and B are events, and A contains all of the outcomes in B and more, then we say that B is a *subset* of A, $B \subset A$ and $P[B] < P[A]$.

Probability Interpretations

- There are three different interpretations of probability:

  1. **Classical probability**: considers only sample spaces where every outcome is *equally likely.* If we have n outcomes in our sample space ($\#\mathcal{S} = n$), then for every outcome $s \in \mathcal{S}$ and event $A \subseteq \mathcal{S}$ we have

  $$\mathrm{P}\left[\{s\}\right] = \frac{1}{n}, \qquad \mathrm{P}\left[A\right] = \frac{\#A}{n} = \frac{\text{number of ways A can occur}}{\text{total no. outcomes}}.$$

  2. **Frequentist probability**: Suppose we repeat the trial n times, and count the number of trials where the event A occurred. The frequentist approach claims that the probability of the event A occurring is the *limit of its relative frequency* in a large number of trials:

  $$\mathrm{P}\left[A\right] = \lim_{n \to \infty} \frac{n_A}{n}$$

  3. **Subjective probability** views the probability of an event as a measure of an individual's degree of belief that that event will occur.

- Regardless of which interpretation of probability we use, all probabilities must follow the same laws and axioms to be coherent.

Examples

- The probability that student A will fail a certain examination is 0.5, for student B the probability is 0.2, and the probability that both A and B will fail the examination is 0.1. What is the probability that at least one of A and B will fail the examination?
- In the Cluedo example, suppose the probabilities of the guilty suspect are as follows:

Guilty Suspect	M	P	S	G
Probability	0.5	0.25	0.1	p

  1. Deduce the value of the missing probability p.
  2. Find the probability that both Col. Mustard and Rev. Green are innocent.

Suggested Exercises: Q1–11.

1.4 Conditional Probability

Conditional Probability and Independence

- For any two events A, B, the notation $P[A|B]$ means the *conditional* probability that event A occurs, given that the event B has already occurred.
- Conditional probabilities are obtained either directly or by using the *conditional probability rule*:

  The conditional probability rule

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}, \quad \text{for } \mathrm{P}[B] > 0.$$

- Rearranging this equation gives the *multiplication rule*, useful in simplifying probabilities: for any two events A, B,

  The multiplication rule

$$\mathrm{P}[A \cap B] = \mathrm{P}[A|B]\,\mathrm{P}[B]\,.$$

Independence

- Two events are said to be *independent* when the occurrence of one has no bearing on the occurrence of the other.
- In terms of probability, if A, B are independent then

$$\mathrm{P}[A|B] = \mathrm{P}[A]$$

  as the knowledge that B occurred is irrelevant.
- For independent events A,B, the multiplication rule can then be simplified,

$$\mathrm{P}[A \cap B] = \mathrm{P}[A]\,\mathrm{P}[B]\,.$$

- **Note:** Beware of confusing independent events with disjoint events. Independent events do not affect each other in any way, whereas disjoint events cannot occur together – disjoint events are very much dependent on each other.

Example: Two Dice

Two fair dice are rolled, what is the probability that the sum of the two numbers that appear is even?

Example: Nuclear Power Station

Suppose that a nuclear power station has three separate (and independent) devices for detecting a problem and shutting down the reactor. Suppose that each device has a probability of 0.9 of working correctly. In the event of a problem, what is the probability that the reactor will be shut down?

Partitions

- Suppose that n events $E_1, \ldots, E_n$ are *disjoint*, and suppose that exactly one must happen. Such a collection of events is called a *partition*.
- Now we can write any other event A in combination with this partition: in general,

$$\mathrm{P}[A] = \mathrm{P}[A \cap E_1] + \mathrm{P}[A \cap E_2] + \ldots + \mathrm{P}[A \cap E_n]\,,$$

- Using the multiplication rule, we can simplify this to get

  The partition theorem (or theorem of total probability)

$$\mathrm{P}[A] = \mathrm{P}[A|E_1]\,\mathrm{P}[E_1] + \mathrm{P}[A|E_2]\,\mathrm{P}[E_2] + \ldots + \mathrm{P}[A|E_n]\,\mathrm{P}[E_n]\,.$$

- Often, this is the most convenient way of getting at certain hard-to-think-about events: to associate them with a suitable partition, and then use conditional probability to simplify matters.

Bayes Theorem

- For any two events A, B, the multiplication rule gives the formula

$$\mathrm{P}[A \cap B] = \mathrm{P}[A|B]\,\mathrm{P}[B].$$

- Another equivalent formula is obviously

$$\mathrm{P}[A \cap B] = \mathrm{P}[B \cap A] = \mathrm{P}[B|A]\,\mathrm{P}[A].$$

- By equating these two formulae and rearranging, we obtain the formula known as

  Bayes theorem

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[B|A]\,\mathrm{P}[A]}{\mathrm{P}[B]}.$$

- It is useful mainly as a way of "inverting" probabilities. Often, the probability in the denominator must be calculated using the simplifying method shown in the last section; i.e. via a *partition*.

Example: Diagnosing Diseases

A clinic offers a test for a very rare and unpleasant disease which affects 1/10000 people. The test itself is 90% reliable, i.e. test results are positive 90% of the time *given you have the disease*. If you don't have the disease the test reports a false positive only 1% of the time. You decide to take the test. What is the probability that the test is positive? Your test returns a positive result. What is the probability you have the disease now?

Suggested Exercises: Q2–17.

---

2 Random Variables

- A *random variable* (rv) is a variable which takes different numerical values, according to the different possible outcomes of an experiment or random phenomenon.
- Random variables are *discrete* if they only take a finite number of values (e.g. outcome of a coin flip).
- The opposite is a *continuous* random variable with an infinite sample space (e.g. a real-valued measurement).

2.1 Discrete Random Variables

Discrete Random Variables and Probability Distributions

- A discrete random variable X is defined by a pair of two lists

Possible values:	x_1	x_2	x_3	$\ldots$
Attached probabilities:	$\mathrm{P}[X = x_1]$	$\mathrm{P}[X = x_2]$	$\mathrm{P}[X = x_3]$	$\ldots$

- This collection of all possible values with their probabilities is called the *probability distribution* of X.
- The probabilities in a probability distribution must:

  1. be non-negative – $\mathrm{P}[X = x_i] \geq 0,\ \forall i$
  2. add to one – $\sum_i \mathrm{P}[X = x_i] = 1$

Joint and Marginal Distributions

- When we have two (or more) random variables X and Y, the *joint probability distribution* is the table of every possible (x, y) value for X and Y, with the associated probabilities $\mathrm{P}[X = x, Y = y]$:

	x_1	$\dots$	x_n
y_1	$\mathrm{P}[X = x_1, Y = y_1]$	$\dots$	$\mathrm{P}[X = x_n, Y = y_1]$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_m	$\mathrm{P}[X = x_1, Y = y_m]$	$\dots$	$\mathrm{P}[X = x_n, Y = y_m]$

- Given the joint distribution for the random variables (X, Y), we can obtain the distribution of X (or Y) alone – the *marginal probability distribution* for X (or Y) – by summing across the rows or columns:
$$\mathrm{P}[X = x] = \sum_{\text{all } y} \mathrm{P}[X = x, Y = y].$$

Example: Discrete Random Variables

Let X be the random variable which takes value 3 when a fair coin lands heads up, and takes value 0 otherwise. Let Y be the value shown after rolling a fair dice. Write down the distributions of X, and Y, and the joint distribution of (X, Y). You may assume that X and Y are independent. Thus find the probability that $X > Y$

Suggested Exercises: Q18–22.

2.2 Continuous Random Variables

Continuous random variables

- Discrete rvs only make sense when our sample space is finite.
- When our experimental outcome is a measurement of some quantity, then our sample space is actually part of the real line and so is infinite.
- A random variable X which can assume every real value in an interval (bounded or unbounded) is called a *continuous random variable.*
- Since our sample space is now infinite we cannot write down a table of probabilities for every possibly outcome to describe the distribution of X.
- Instead, the probability distribution for X is described by a *probability density function* (pdf), $f(x)$, which is a function that *describes a curve over the range of possible values* taken by the random variable.

Continuous random variables

- A valid probability density function, $f(x)$, must
  1. be non-negative everywhere: $f(x) \geq 0, \forall x$,
  2. *integrate* to 1: $\int_{-\infty}^{\infty} f(x)\, dx = 1$,
- The probability for a range of values is given by *the area under the curve.*
$$\mathrm{P}[a \leq X \leq b] = \int_a^b f(x)\ dx$$
  Note: $f(x) \neq \mathrm{P}[X = x]$ — probability densities *are not* probabilities
- We can describe the probability by the function
$$F(x) \equiv \int_{-\infty}^{x} f(y)\ dy = \mathrm{P}[X \leq x]$$
  which is called the *cumulative distribution function* (cdf) of X.

- We also have the result that $f(x) = F'(x)$.

Joint and Marginal Distributions

- When we have two (or more) continuous random variables, we describe them via their joint probability density function $f_{xy}(x, y)$, which satisfies the usual conditions for pdfs
- The probability that X and Y fall into some region A of the xy-plane is then

$$\mathrm{P}\left[(X, Y) \in A\right] = \int_A \int f_{xy}(x, y)\, dxdy$$

- Given the joint pdf $f_{xy}(x, y)$, we can obtain the *marginal* pdf of x or y by integrating out the other variable

$$f_x(x) = \int_{-\infty}^{\infty} f_{xy}(x, y)\, dy$$

- When two continuous random variables x and y are *independent*, their joint pdf can be expressed as the *product* of the marginal pdfs

$$f_{xy}(x, y) = f_x(x)\, f_y(y)$$

Example: The Exponential Distribution
Let X be a continuous random variable with probability density function:

$$f(x) = \begin{cases} \beta e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases}$$

Show that $f(x)$ is a valid probability density function when $\beta > 0$. Find the cdf of X, and hence $\mathrm{P}\left[X > 3\right]$.

Suggested Exercises: Q23–26.

---

3 Expectation and Variance

Distribution Summaries

- The distribution of a random variable X contains all of the probabilistic information about X.
- However, the entire distribution of X can often be too complex to work with.
- Summaries of the distribution, such as its average or spread can be useful for conveying information about X without trying to describe it in its entirety.
- Formally, we measure the average of the distribution by calculating its expectation, and we measure the spread by its variance.

3.1 Expectation

- Suppose that X has a discrete distribution, then the expectation of X is given by

$$\mathrm{E}\left[X\right] = \sum_{\text{all } x} x \mathrm{P}\left[X = x\right]$$

- If a random variable X has a continuous distribution with a pdf $f(\cdot)$, then the expectation of X is defined as:

$$\mathrm{E}\left[X\right] = \int_{-\infty}^{\infty} x\, f(x)\; dx$$

- The value $E(X)$ is the theoretical average of the probability distribution. Because of this, it is often referred to it as the *mean* or *average* for the distribution.

Properties of Expectation

- **Expectation of a function:** If X is a random variable, then the expectation of the function $r(X)$ is given by

$$\mathrm{E}\left[r(X)\right] = \sum_{\text{all } x} r(x)\mathrm{P}\left[X = x\right], \qquad \text{or } \mathrm{E}\left[r(X)\right] = \int_{-\infty}^{\infty} r(x)\, f(x)\; dx$$

- **Linearity:** If $Y = a + b\,X$ where a and b are constants, then

$$\mathrm{E}\left[Y\right] = a + b\mathrm{E}\left[X\right].$$

- **Additivity:** If X_1, X_2, ..., X_n are any random variables then

$$\mathrm{E}\left[X_1 + X_2 + \cdots + X_n\right] = \mathrm{E}\left[X_1\right] + \mathrm{E}\left[X_2\right] + \cdots + \mathrm{E}\left[X_n\right]$$

- If X_1, X_2 are any pair of *independent* random variables then

$$\mathrm{E}\left[X_1 X_2\right] = \mathrm{E}\left[X_1\right]\mathrm{E}\left[X_2\right]$$

3.2 Variance

- Suppose that X is a random variable with mean $\mu = \mathrm{E}[X]$. The *variance* of X, denoted $\mathrm{Var}[X]$, is defined as follows:

$$\mathrm{Var}\left[X\right] = \mathrm{E}\left[(X - \mu)^2\right].$$

- **Note:** Since $\mathrm{Var}[X]$ is the expected value of a non-negative random variable $(X - \mu)^2$, it follows that $\mathrm{Var}[X] \geq 0$.
- We can re-write the variance formula in the following simpler form:

$$\mathrm{Var}\left[X\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2.$$

- The *standard deviation* of a random variable is defined as the square root of the variance: $\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$.

Properties of Variance

- For constants a and b:

$$\mathrm{Var}\left[a + bX\right] = b^2\mathrm{Var}\left[X\right], \qquad \mathrm{SD}\left[a + bX\right] = b\ \mathrm{SD}\left[X\right].$$

- If X_1, ..., X_n are *independent* random variables, then

$$\mathrm{Var}\left[X_1 + \cdots + X_n\right] = \mathrm{Var}\left[X_1\right] + \cdots + \mathrm{Var}\left[X_n\right].$$

Example: National Lottery

The National lottery has a game called 'Thunderball'. You pick 5 numbers in the range 1-34 and one number (the Thunderball number) in the range 1-14. You win a prize if you match at least two numbers, including the Thunderball number. Let X be the amount you win in a single game. The probability distribution for X is given below. Find the expectation and variance of X.

Matches	k, Prize £	$Pr(X = k)$
5 +Tb	250000	0.000000257
5	5000	0.000003337
4 +Tb	250	0.000037220
4	100	0.000483653
3 +Tb	20	0.001041124
3	10	0.013368984
2 +Tb	10	0.009293680
1 +Tb	5	0.029585799
Other	0	0.946185946
Sum		1.000000000

Suggested Exercises: Q18–26.