Probability

About these notes. Many people have written excellent notes for introductory courses in probability. Mine draw freely on material prepared by others in presenting this course to students at Cambridge. I wish to acknowledge especially Geoffrey Grimmett, Frank Kelly and Doug Kennedy.

The order I follow is a bit different to that listed in the Schedules. Most of the material can be found in the recommended books by Grimmett & Welsh, and Ross. Many of the examples are classics and mandatory in any sensible introductory course on probability. The book by Grinstead & Snell is easy reading and I know students have enjoyed it.

There are also some very good Wikipedia articles on many of the topics we will consider.

In these notes I attempt a 'Goldilocks path' by being neither too detailed or too brief.

- Each lecture has a title and focuses upon just one or two ideas.
- My notes for each lecture are limited to 4 pages.

I also include some entertaining, but nonexaminable topics, some of which are unusual for a course at this level (such as random permutations, entropy, reflection principle, Benford and Zipf distributions, Erdős's probabilistic method, value at risk, eigenvalues of random matrices, Kelly criterion, Chernoff bound).

You should enjoy the book of Grimmett & Welsh, and the notes notes of Kennedy.

Printed notes, good or bad? I have wondered whether it is helpful or not to publish full course notes. On balance, I think that it is. It is helpful in that we can dispense with some tedious copying-out, and you are guaranteed an accurate account. But there are also benefits to hearing and writing down things yourself during a lecture, and so I recommend that you still do some of that.

I will say things in every lecture that are not in the notes. I will sometimes tell you when it would be good to make an extra note. In learning mathematics repeated exposure to ideas is essential. I hope that by doing all of reading, listening, writing and (most importantly) solving problems you will master and enjoy this course.

I recommend Tom Körner's treatise on how to listen to a maths lecture.

Contents

About these notes i

Table of Contents ii

Schedules vi

Learning outcomes vii

1 Classical probability 1
- 1.1 Diverse notions of 'probability' . . . 1
- 1.2 Classical probability . . . 1
- 1.3 Sample space and events . . . 2
- 1.4 Equalizations in random walk . . . 3

2 Combinatorial analysis 6
- 2.1 Counting . . . 6
- 2.2 Sampling with or without replacement . . . 6
- 2.3 Sampling with or without regard to ordering . . . 8
- 2.4 Four cases of enumerative combinatorics . . . 8

3 Stirling's formula 10
- 3.1 Multinomial coefficient . . . 10
- 3.2 Stirling's formula . . . 11
- 3.3 Improved Stirling's formula . . . 13

4 Axiomatic approach 14
- 4.1 Axioms of probability . . . 14
- 4.2 Boole's inequality . . . 15
- 4.3 Inclusion-exclusion formula . . . 17

5 Independence 18
- 5.1 Bonferroni's inequalities . . . 18
- 5.2 Independence of two events . . . 19
- 5.3 Independence of multiple events . . . 20
- 5.4 Important distributions . . . 20
- 5.5 Poisson approximation to the binomial . . . 21

6 Conditional probability 22
- 6.1 Conditional probability . . . 22
- 6.2 Properties of conditional probability . . . 22
- 6.3 Law of total probability . . . 23
- 6.4 Bayes' formula . . . 23
- 6.5 Simpson's paradox . . . 24

7 Discrete random variables **26**
7.1 Continuity of P . . . 26
7.2 Discrete random variables . . . 27
7.3 Expectation . . . 27
7.4 Function of a random variable . . . 29
7.5 Properties of expectation . . . 29

8 Further functions of random variables **30**
8.1 Expectation of sum is sum of expectations . . . 30
8.2 Variance . . . 30
8.3 Indicator random variables . . . 32
8.4 Reproof of inclusion-exclusion formula . . . 33
8.5 Zipf's law . . . 33

9 Independent random variables **34**
9.1 Independent random variables . . . 34
9.2 Variance of a sum . . . 35
9.3 Efron's dice . . . 36
9.4 Cycle lengths in a random permutation . . . 37

10 Inequalities **38**
10.1 Jensen's inequality . . . 38
10.2 AM–GM inequality . . . 39
10.3 Cauchy-Schwarz inequality . . . 39
10.4 Covariance and correlation . . . 40
10.5 Information entropy . . . 41

11 Weak law of large numbers **42**
11.1 Markov inequality . . . 42
11.2 Chebyshev inequality . . . 42
11.3 Weak law of large numbers . . . 43
11.4 Probabilistic proof of Weierstrass approximation theorem . . . 44
11.5 Benford's law . . . 45

12 Probability generating functions **46**
12.1 Probability generating function . . . 46
12.2 Combinatorial applications . . . 48

13 Conditional expectation **50**
13.1 Conditional distribution and expectation . . . 50
13.2 Properties of conditional expectation . . . 51
13.3 Sums with a random number of terms . . . 51
13.4 Aggregate loss distribution and VaR . . . 52
13.5 Conditional entropy . . . 53

14 Branching processes **54**
14.1 Branching processes 54
14.2 Generating function of a branching process 54
14.3 Probability of extinction 56

15 Random walk and gambler's ruin **58**
15.1 Random walks 58
15.2 Gambler's ruin 58
15.3 Duration of the game 60
15.4 Use of generating functions in random walk 61

16 Continuous random variables **62**
16.1 Continuous random variables 62
16.2 Uniform distribution 64
16.3 Exponential distribution 64
16.4 Hazard rate 65
16.5 Relationships among probability distributions 65

17 Functions of a continuous random variable **66**
17.1 Distribution of a function of a random variable 66
17.2 Expectation 67
17.3 Stochastic ordering of random variables 68
17.4 Variance 68
17.5 Inspection paradox 69

18 Jointly distributed random variables **70**
18.1 Jointly distributed random variables 70
18.2 Independence of continuous random variables 71
18.3 Geometric probability 71
18.4 Bertrand's paradox 72
18.5 Buffon's needle 73

19 Normal distribution **74**
19.1 Normal distribution 74
19.2 Calculations with the normal distribution 75
19.3 Mode, median and sample mean 76
19.4 Distribution of order statistics 76
19.5 Stochastic bin packing 77

20 Transformations of random variables **78**
20.1 Transformation of random variables 78
20.2 Convolution 80
20.3 Cauchy distribution 81

21 Moment generating functions **82**

21.1 What happens if the mapping is not 1–1? . . . 82
21.2 Minimum of exponentials is exponential . . . 82
21.3 Moment generating functions . . . 83
21.4 Gamma distribution . . . 84
21.5 Beta distribution . . . 85

22 Multivariate normal distribution **86**
22.1 Moment generating function of normal distribution . . . 86
22.2 Functions of normal random variables . . . 86
22.3 Bounds on tail probability of a normal distribution . . . 87
22.4 Multivariate normal distribution . . . 87
22.5 Bivariate normal . . . 88
22.6 Multivariate moment generating function . . . 89

23 Central limit theorem **90**
23.1 Central limit theorem . . . 90
23.2 Normal approximation to the binomial . . . 91
23.3 Estimating π with Buffon's needle . . . 93

24 Continuing studies in probability **94**
24.1 Large deviations . . . 94
24.2 Chernoff bound . . . 94
24.3 Random matrices . . . 96
24.4 Concluding remarks . . . 97

A Problem solving strategies **98**

B Fast Fourier transform and p.g.fs **100**

C The Jacobian **101**

D Beta distribution **103**

E Kelly criterion **104**

F Ballot theorem **105**

G Allais paradox **106**

H IB courses in applicable mathematics **107**

Index **107**

Richard Weber, Lent Term 2014

This is reproduced from the Faculty handbook.

Schedules

All this material will be covered in lectures, but in a slightly different order.

Basic concepts: Classical probability, equally likely outcomes. Combinatorial analysis, permutations and combinations. Stirling's formula (asymptotics for $\log n!$ proved). [3]

Axiomatic approach: Axioms (countable case). Probability spaces. Inclusion-exclusion formula. Continuity and subadditivity of probability measures. Independence. Binomial, Poisson and geometric distributions. Relation between Poisson and binomial distributions. Conditional probability, Bayes' formula. Examples, including Simpson's paradox. [5]

Discrete random variables: Expectation. Functions of a random variable, indicator function, variance, standard deviation. Covariance, independence of random variables. Generating functions: sums of independent random variables, random sum formula, moments. Conditional expectation. Random walks: gambler's ruin, recurrence relations. Difference equations and their solution. Mean time to absorption. Branching processes: generating functions and extinction probability. Combinatorial applications of generating functions. [7]

Continuous random variables: Distributions and density functions. Expectations; expectation of a function of a random variable. Uniform, normal and exponential random variables. Memoryless property of exponential distribution. Joint distributions: transformation of random variables (including Jacobians), examples. Simulation: generating continuous random variables, independent normal random variables. Geometrical probability: Bertrand's paradox, Buffon's needle. Correlation coefficient, bivariate normal random variables. [6]

Inequalities and limits: Markov's inequality, Chebyshev's inequality. Weak law of large numbers. Convexity: Jensens inequality for general random variables, AM/GM inequality. Moment generating functions and statement (no proof) of continuity theorem. Statement of central limit theorem and sketch of proof. Examples, including sampling. [3]

This is reproduced from the Faculty handbook.

Learning outcomes

From its origin in games of chance and the analysis of experimental data, probability theory has developed into an area of mathematics with many varied applications in physics, biology and business.

The course introduces the basic ideas of probability and should be accessible to students who have no previous experience of probability or statistics. While developing the underlying theory, the course should strengthen students' general mathematical background and manipulative skills by its use of the axiomatic approach. There are links with other courses, in particular Vectors and Matrices, the elementary combinatorics of Numbers and Sets, the difference equations of Differential Equations and calculus of Vector Calculus and Analysis. Students should be left with a sense of the power of mathematics in relation to a variety of application areas. After a discussion of basic concepts (including conditional probability, Bayes' formula, the binomial and Poisson distributions, and expectation), the course studies random walks, branching processes, geometric probability, simulation, sampling and the central limit theorem. Random walks can be used, for example, to represent the movement of a molecule of gas or the fluctuations of a share price; branching processes have applications in the modelling of chain reactions and epidemics. Through its treatment of discrete and continuous random variables, the course lays the foundation for the later study of statistical inference. By the end of this course, you should:

- understand the basic concepts of probability theory, including independence, conditional probability, Bayes' formula, expectation, variance and generating functions;
- be familiar with the properties of commonly-used distribution functions for discrete and continuous random variables;
- understand and be able to apply the central limit theorem.
- be able to apply the above theory to 'real world' problems, including random walks and branching processes.

1 Classical probability

Classical probability. Sample spaces. Equally likely outcomes. *Equalizations of heads and tails*. *Arcsine law*.

1.1 Diverse notions of 'probability'

Consider some uses of the word 'probability'.

1. The probability that a fair coin will land heads is 1/2.
2. The probability that a selection of 6 numbers wins the National Lottery Lotto jackpot is 1 in $\binom{49}{6}$ =13,983,816, or 7.15112×10^{-8}.
3. The probability that a drawing pin will land 'point up' is 0.62.
4. The probability that a large earthquake will occur on the San Andreas Fault in the next 30 years is about 21%.
5. The probability that humanity will be extinct by 2100 is about 50%.

Clearly, these are quite different notions of probability (known as classical[1,2], frequentist[3] and subjective[4,5] probability).

Probability theory is useful in the biological, physical, actuarial, management and computer sciences, in economics, engineering, and operations research. It helps in modeling complex systems and in decision-making when there is uncertainty. It can be used to prove theorems in other mathematical fields (such as analysis, number theory, game theory, graph theory, quantum theory and communications theory).

Mathematical probability began its development in Renaissance Europe when mathematicians such as Pascal and Fermat started to take an interest in understanding games of chance. Indeed, one can develop much of the subject simply by questioning what happens in games that are played by tossing a fair coin. We begin with the classical approach (lectures 1 and 2), and then shortly come to an axiomatic approach (lecture 4) which makes 'probability' a well-defined mathematical subject.

1.2 Classical probability

Classical probability applies in situations in which there are just a finite number of equally likely possible outcomes. For example, tossing a fair coin or an unloaded die, or picking a card from a standard well-shuffled pack.

Example 1.1 [*Problem of points considered by Pascal, Fermat 1654*]. Equally skilled players A and B play a series of games. The winner of each game gets a point. The winner is the first to reach 10 points. They are forced to stop early, when A has 8 points and B has 7 points, How should they divide the stake?

Consider the next 4 games. Exactly one player must reach 10 points. There are 16 equally likely outcomes:

A wins			B wins	
AAAA	AAAB	AABB	ABBB	BBBB
	AABA	ABBA	BABB	
	ABAA	ABAB	BBAB	
	BAAA	BABA	BBBA	
		BAAB		
		BBAA		

Player A wins if she wins 4, 3 or 2 of the next 4 games (and loses if she wins only 1 or 0 games). She can win 4, 3 or 2 games in 1, 4 and 6 ways, respectively. There are 16 ($= 2 \times 2 \times 2 \times 2$) possible results for the next 4 games. So $P(\text{A wins}) = 11/16$. It would seem fair that she should receive 11/16 of the stake.

1.3 Sample space and events

Let's generalise the above example. Consider an experiment which has a random outcome. The set of all possible outcomes is called the **sample space**. If the number of possible outcomes is countable we might list them, as $\omega_1, \omega_2, \ldots$, and then the sample space is $\Omega = \{\omega_1, \omega_2, \ldots\}$. Choosing a particular point $\omega \in \Omega$ provides an **observation**.

Remark. A sample space need not be countable. For example, an infinite sequence of coin tosses like TTHTHHT... is in 1–1 relation to binary fractions like 0.0010110... and the number of these is uncountable.

Certain set theory notions have special meaning and terminology when used in the context of probability.

1. A subset A of Ω is called an **event**.
2. For any events $A, B \in \Omega$,
   - The **complement event** $A^c = \Omega \setminus A$ is the event that A does not occur, or 'not A'. This is also sometimes written as $\bar{A}$ or A'.
   - $A \cup B$ is 'A or B'.
   - $A \cap B$ is 'A and B'.
   - $A \subseteq B$: occurrence of A implies occurrence of B.
   - $A \cap B = \varnothing$ is 'A and B are **mutually exclusive** or **disjoint events**'.

As already mentioned, in classical probability the sample space consists of a finite number of equally likely outcomes, $\Omega = \{\omega_1, \ldots, \omega_N\}$. For $A \subseteq \Omega$,

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } \Omega} = \frac{|A|}{N}.$$

Thus (as Laplace put it) $P(A)$ is the quotient of 'number of favourable outcomes' (when A occurs) divided by 'number of possible outcomes'.

Example 1.2. Suppose r digits are chosen from a table of random numbers. Find the probability that, for $0 \leq k \leq 9$, (i) no digit exceeds k, and (ii) k is the greatest digit drawn.

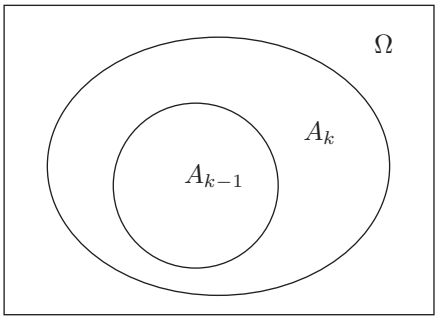
Take

$$\Omega = \{(a_1, \ldots, a_r) : 0 \leq a_i \leq 9, \quad i = 1, \ldots, r\}.$$

Let $A_k = [\text{no digit exceeds } k]$, or as a subset of Ω

$$A_k = \{(a_1, \ldots, a_r) : 0 \leq a_i \leq k, \quad i = 1, \ldots, r\}.$$

Thus $|\Omega| = 10^r$ and $|A_k| = (k+1)^r$. So (i) $P(A_k) = (k+1)^r/10^r$.



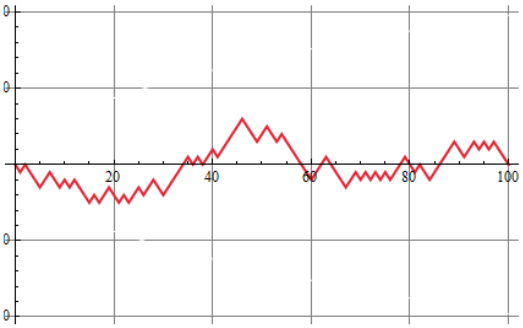
(ii) The event that k is the greatest digit drawn is $B_k = A_k \setminus A_{k-1}$. So $|B_k| = |A_k| - |A_{k-1}|$ and

$$P(B_k) = \frac{(k+1)^r - k^r}{10^r}.$$

1.4 Equalizations in random walk

In later lectures we will study random walks. Many interesting questions can be asked about the random path produced by tosses of a fair coin (+1 for a head, −1 for a tail).

By the following example I hope to convince you that probability theory contains beautiful and surprising results.



What is the probability that after an odd number of steps the walk is on the positive side of the x-axis? (Answer: obviously 1/2.)
How many times on average does a walk of length n cross the x-axis?
When does the first (or last) crossing of the x-axis typically occur?
What is the distribution of terminal point? The walk at the left returned to the x-axis after 100 steps. How likely is this?

Example 1.3. Suppose we toss a fair coin $2n$ times. We say that an *equalization* occurs at the $(2k)$th toss if there have been k heads and k tails. Let u_n be the probability that equalization occurs at the $(2n)$th toss (so there have been n heads and n tails).

Here are two rare things that might happen when we toss a coin $2n$ times.

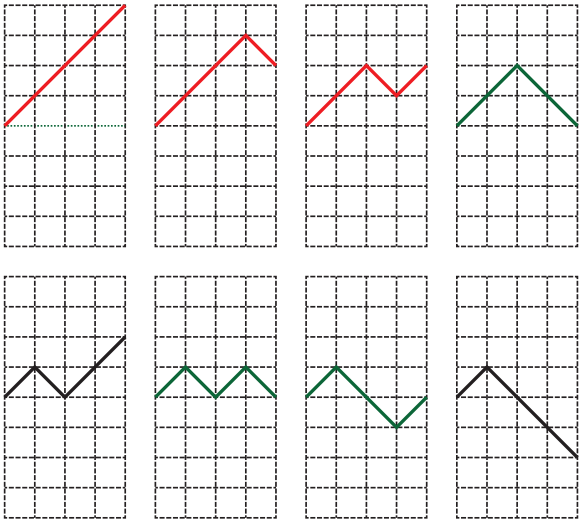
- No equalization ever takes place (except at the start).
- An equalization takes place at the end (exactly n heads and n tails).

Which do you think is more likely?

Let $\alpha_k = P(\text{there is no equalization after any of } 2, 4, \ldots, 2k \text{ tosses})$.

Let $u_k = P(\text{there is equalization after } 2k \text{ tosses})$.

The pictures at the right show results of tossing a fair coin 4 times, when the first toss is a head.

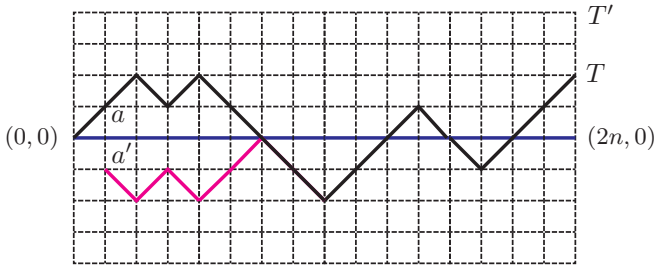


Notice that of these 8 equally likely outcomes there are 3 that have no equalization except at the start, and 3 that have an equalization at the end.

So $\alpha_2 = u_2 = 3/8 = \dfrac{1}{2^4}\binom{4}{2}$.

We will prove that $\alpha_n = u_n$.

Proof. We count the paths from the origin to a point T above the axis that do not have any equalization (except at the start). Suppose the first step is to $a = (1, 1)$. Now we must count all paths from a to T, *minus* those that go from a to T but at some point make an equalization, such as the path shown in black below:



But notice that every such path that has an equalization is in 1–1 correspondence with a path from $a' = (1, -1)$ to T. This is the path obtained by reflecting around the axis the part of the path that takes place before the first equalization.

The number of paths from a' to $T = (2n, k)$ equals the number from a to $T' = (2n, k+2)$. So the number of paths from a to some $T > 0$ that have no equalization is

$$\sum_{k=2,4,\ldots,2n} \Big(\#[a \to (2n,k)] - \#[a \to (2n,k+2)] \Big) = \#[a \to (2n,2)] = \#[a \to (2n,0)].$$

We want twice this number (since the first step might have been to a'), which gives $\#[(0,0) \to (2n,0)] = \binom{2n}{n}$. So as claimed

$$\alpha_n = u_n = \frac{1}{2^{2n}} \binom{2n}{n}. \qquad \square$$

Arcsine law. The probability that the last equalization occurs at $2k$ is therefore $u_k \alpha_{n-k}$ (since we must equalize at $2k$ and then not equalize at any of the $2n - 2k$ subsequent steps). But we have just proved that $u_k \alpha_{n-k} = u_k u_{n-k}$. Notice that therefore the last equalization occurs at $2n - 2k$ with the same probability.

We will see in Lecture 3 that u_k is approximately $1/\sqrt{\pi k}$, so the last equalization is at time $2k$ with probability proportional to $1/\sqrt{k(n-k)}$.

The probability that the last equalization occurs before the $2k$th toss is approximately

$$\int_0^{\frac{2k}{2n}} \frac{1}{\pi} \frac{1}{\sqrt{x(1-x)}} dx = (2/\pi) \sin^{-1} \sqrt{k/n}.$$

For instance, $(2/\pi) \sin^{-1} \sqrt{0.15} = 0.2532$. So the probability that the last equalization occurs during either the first or last 15% of the $2n$ coin tosses is about 0.5064 $(> 1/2)$.

This is a nontrivial result that would be hard to have guessed!

2 Combinatorial analysis

Combinatorial analysis. Fundamental rules. Sampling with and without replacement, with and without regard to ordering. Permutations and combinations. Birthday problem. Binomial coefficient.

2.1 Counting

Example 2.1. A menu with 6 starters, 7 mains and 6 desserts has $6 \times 7 \times 6 = 252$ meal choices.

Fundamental rule of counting: Suppose r multiple choices are to be made in sequence: there are m_1 possibilities for the first choice; then m_2 possibilities for the second choice; then m_3 possibilities for the third choice, and so on until after making the first $r-1$ choices there are m_r possibilities for the rth choice. Then the total number of different possibilities for the set of choices is

$$m_1 \times m_2 \times \cdots \times m_r.$$

Example 2.2. How many ways can the integers $1, 2, \ldots, n$ be ordered?

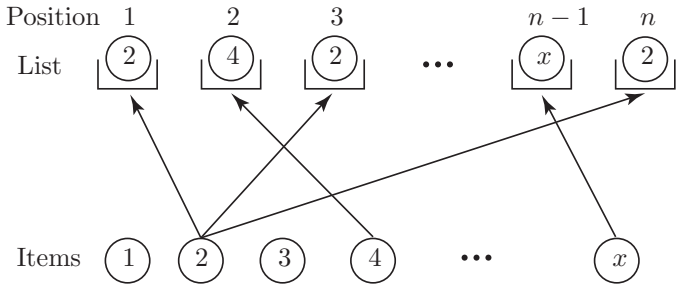
The first integer can be chosen in n ways, then the second in $n-1$ ways, etc., giving $n! = n(n-1)\cdots 1$ ways ('factorial n').

2.2 Sampling with or without replacement

Many standard calculations arising in classical probability involve counting numbers of equally likely outcomes. This can be tricky!

Often such counts can be viewed as counting the number of lists of length n that can be constructed from a set of x items $X = \{1, \ldots, x\}$.

Let $N = \{1, \ldots, n\}$ be the set of list positions. Consider the function $f : N \to X$. This gives the ordered list $(f(1), f(2), \ldots, f(n))$. We might construct this list by drawing a sample of size n from the elements of X. We start by drawing an item for list position 1, then an item for list position 2, etc.



1. **Sampling with replacement**. After choosing an item we put it back so it can be chosen again. E.g. list $(2, 4, 2, \ldots, x, 2)$ is possible, as shown above.
2. **Sampling without replacement**. After choosing an item we set it aside. We end up with an ordered list of n distinct items (requires $x \geq n$).
3. Sampling with replacement, but requiring each item is chosen at least once (requires $n \geq x$).

These three cases correspond to 'any f', 'injective f' and 'surjective f', respectively

Example 2.3. Suppose $N = \{a, b, c\}$, $X = \{p, q, r, s\}$. How many different injective functions are there mapping N to X?

Solution: Choosing the values of $f(a)$, $f(b)$, $f(c)$ in sequence without replacement, we find the number of different injective $f : N \to X$ is $4 \times 3 \times 2 = 24$.

Example 2.4. I have n keys in my pocket. I select one at random and try it in a lock. If it fails I replace it and try again (sampling with replacement).

$$P(\text{success at } r\text{th trial}) = \frac{(n-1)^{r-1} \times 1}{n^r}.$$

If keys are not replaced (sampling without replacement)

$$P(\text{success at } r\text{th trial}) = \frac{(n-1)!}{n!} = \frac{1}{n},$$

or alternatively

$$= \frac{n-1}{n} \times \frac{n-2}{n-1} \times \cdots \times \frac{n-r+1}{n-r+2} \times \frac{1}{n-r+1} = \frac{1}{n}.$$

Example 2.5 [*Birthday problem*]. How many people are needed in a room for it to be a favourable bet (probability of success greater than 1/2) that two people in the room will have the same birthday?

Since there are 365 possible birthdays, it is tempting to guess that we would need about 1/2 this number, or 183. In fact, the number required for a favourable bet is only 23. To see this, we find the probability that, in a room with r people, there is no duplication of birthdays; the bet is favourable if this probability is less than one half.

Let $f(r)$ be probability that amongst r people there is a match. Then

$$P(\text{no match}) = 1 - f(r) = \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdot \cdots \cdot \frac{366-r}{365}.$$

So $f(22) = 0.475695$ and $f(23) = 0.507297$. Also $f(47) = 0.955$.

Notice that with 23 people there are $\binom{23}{2} = 253$ pairs and each pair has a probability 1/365 of sharing a birthday.

Remarks. A slightly different question is this: interrogating an audience one by one, how long will it take on average until we find a first birthday match? Answer: 23.62 (with standard deviation of 12.91).

The probability of finding a triple with the same birthday exceeds 0.5 for $n \geq 88$. (How do you think I computed that answer?)

2.3 Sampling with or without regard to ordering

When counting numbers of possible $f : N \to X$, we might decide that the labels that are given to elements of N and X do or do not matter.

So having constructed the set of possible lists $(f(1), \ldots, f(n))$ we might

(i) leave lists alone (order matters);

(ii) sort them ascending: so (2,5,4) and (4,2,5) both become (2,4,5).
(labels of the positions in the list do not matter.)

(iii) renumber each item in the list by the number of the draw on which it was first seen: so (2,5,2) and (5,4,5) both become (1,2,1).
(labels of the items do not matter.)

(iv) do both (ii) then (iii), so (2,5,2) and (8,5,5) both become (1,1,2).
(no labels matter.)

For example, in case (ii) we are saying that $(g(1), \ldots, g(n))$ is the same as $(f(1), \ldots, f(n))$ if there is permutation of π of $1, \ldots, n$, such that $g(i) = f(\pi(i))$.

2.4 Four cases of enumerative combinatorics

Combinations of 1,2,3 (top of page 7) and (i)–(iv) above produce a 'twelvefold way of enumerative combinatorics', but involve the partition function and **Bell numbers**. Let's consider just the four possibilities obtained from combinations of 1,2 and (i),(ii).

1(i) **Sampling with replacement and with ordering**. Each location in the list can be filled in x ways, so this can be done in x^n ways.

2(i) **Sampling without replacement and with ordering**. Applying the fundamental rule, this can be done in $x_{(n)} = x(x-1)\cdots(x-n+1)$ ways. Another notation for this falling sequential product is $x^{\underline{n}}$ (read as 'x to the n falling').

In the special case $n = x$ this is $x!$ (the number of permutations of $1, 2, \ldots, x$).

2(ii) **Sampling without replacement and without ordering**. Now we care only which items are selected. (The positions in the list are indistinguishable.) This can be done in $x_{(n)}/n! = \binom{x}{n}$ ways, i.e. the answer above divided by $n!$.

This is of course the **binomial coefficient**, equal to the number of distinguishable sets of n items that can be chosen from a set of x items.

Recall that $\binom{x}{n}$ is the coefficient of t^n in $(1+t)^x$.

$$\underbrace{(1+t)(1+t)\cdots(1+t)}_{x \text{ times}} = \sum_{n=0}^{x} \binom{x}{n} t^n.$$

1(ii) **Sampling with replacement and without ordering**. Now we care only how many times each item is selected. (The list positions are indistinguishable; we care only how many items of each type are selected.) The number of distinct f is the number of nonnegative integer solutions to

$$n_1 + n_2 + \cdots + n_x = n.$$

Consider $n = 7$ and $x = 5$. Think of marking off 5 bins with 4 dividers: $|$, and then placing 7 $*$s. One outcome is

$$\underbrace{***}_{n_1}|\underbrace{*}_{n_2}|\underbrace{\ }_{n_3}|\underbrace{***}_{n_4}|\underbrace{\ }_{n_5}$$

which corresponds to $n_1 = 3, n_2 = 1, n_3 = 0, n_4 = 3, n_5 = 0$.

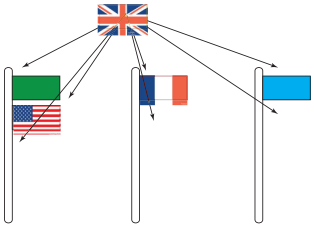
In general, there are $x + n - 1$ symbols and we are choosing n of them to be $*$. So the number of possibilities is $\binom{x+n-1}{n}$.

Above we have attempted a systematic description of different types of counting problem. However it is often best to just think from scratch, using the fundamental rule.

Example 2.6. How may ways can k different flags be flown on m flag poles in a row if ≥ 2 flags may be on the same pole, and order from the top to bottom is important?

There are m choices for the first flag, then $m+1$ for the second. Each flag added creates one more distinct place that the next flag might be added. So

$$m(m+1)\cdots(m+k-1) = \frac{(m+k-1)!}{(m-1)!}.$$



Remark. Suppose we have a diamond, an emerald, and a ruby. How many ways can we store these gems in identical small velvet bags? This is case of 1(iii). Think gems $\equiv$ list positions; bags $\equiv$ items. Take each gem, in sequence, and choose a bag to receive it. There are 5 ways: $(1,1,1)$, $(1,1,2)$, $(1,2,1)$, $(1,2,2)$, $(1,2,3)$. The 1,2,3 are the first, second and third bag to receive a gem. Here we have $B(3) = 5$ (the **Bell numbers**).

3 Stirling's formula

Multinomial coefficient. Stirling's formula *and proof*. Examples of application. *Improved Stirling's formula*.

3.1 Multinomial coefficient

Suppose we fill successive locations in a list of length n by sampling with replacement from $\{1, \dots, x\}$. How may ways can this be done so that the numbers of times that each of $1, \dots, x$ appears in the list is $n_1, \dots, n_x$, respectively where $\sum_i n_i = n$?

To compute this: we choose the n_1 places in which '1' appears in $\binom{n}{n_1}$ ways, then choose the n_2 places in which '2' appears in $\binom{n-n_1}{n_2}$ ways, etc.

The answer is the **multinomial coefficient**

$$\binom{n}{n_1, \dots, n_x} := \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_x}$$
$$= \frac{n!}{n_1! n_2! \cdots n_x!},$$

with the convention $0! = 1$.

Fact:

$$(y_1 + \cdots + y_x)^n = \sum \binom{n}{n_1, \dots, n_x} y_1^{n_1} \cdots y_x^{n_x},$$

where the sum is over all $n_1, \dots, n_x$ such that $n_1 + \cdots + n_x = n$. [Remark. The number of terms in this sum is $\binom{n+x-1}{x-1}$, as found in §2.4, 1(ii).]

Example 3.1. How may ways can a pack of 52 cards be dealt into bridge hands of 13 cards for each of 4 (distinguishable) players?

$$\text{Answer:} \quad \binom{52}{13, 13, 13, 13} = \binom{52}{13}\binom{39}{13}\binom{26}{13} = \frac{52!}{(13!)^4}.$$

This is 53644737765488792839237440000$= 5.36447 \times 10^{28}$. This is $\dfrac{(4n)!}{n!^4}$ evaluated at $n = 13$. How might we estimate it for greater n?

$$\text{Answer:} \quad \frac{(4n)!}{n!^4} \approx \frac{2^{8n}}{\sqrt{2}(\pi n)^{3/2}} \qquad (= 5.49496 \times 10^{28} \text{ when } n = 13).$$

Interesting facts:

(i) If we include situations that might occur part way through a bridge game then there are 2.05×10^{33} possible 'positions'.

(ii) The 'Shannon number' is the number of possible board positions in chess. It is roughly 10^{43} (according to Claude Shannon, 1950).

The age of the universe is thought to be about 4×10^{17} seconds.

3.2 Stirling's formula

Theorem 3.2 (Stirling's formula)**.** *As $n \to \infty$,*

$$\log\left(\frac{n!e^n}{n^{n+1/2}}\right) = \log(\sqrt{2\pi}) + O(1/n).$$

The most common statement of Stirling's formula is given as a corollary.

Corollary 3.3. *As $n \to \infty$, $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$.*

In this context, $\sim$ indicates that the ratio of the two sides tends to 1.

This is good even for small n. It is always a slight underestimate.

n	$n!$	Approximation	Ratio
1	1	.922	1.084
2	2	1.919	1.042
3	6	5.836	1.028
4	24	23.506	1.021
5	120	118.019	1.016
6	720	710.078	1.013
7	5040	4980.396	1.011
8	40320	39902.395	1.010
9	362880	359536.873	1.009
10	3628800	3598696.619	1.008

Notice that from the Taylor expansion of $e^n = 1 + n + \cdots + n^n/n! + \cdots$ we have $1 \le n^n/n! \le e^n$.

We first prove the weak form of Stirling's formula, namely that $\log(n!) \sim n \log n$.

Proof. (examinable) $\log n! = \sum_1^n \log k$. Now

$$\int_1^n \log x \, dx \le \sum_1^n \log k \le \int_1^{n+1} \log x \, dx,$$

and $\int_1^z \log x \, dx = z \log z - z + 1$, and so

$$n \log n - n + 1 \le \log n! \le (n+1)\log(n+1) - n.$$

Divide by $n \log n$ and let $n \to \infty$ to sandwich $\frac{\log n!}{n \log n}$ between terms that tend to 1. Therefore $\log n! \sim n \log n$. □

Now we prove the strong form.

Proof. (not examinable) Some steps in this proof are like 'pulling-a-rabbit-out-of-a-hat'. Let

$$d_n = \log\left(\frac{n!e^n}{n^{n+1/2}}\right) = \log n! - \left(n+\tfrac{1}{2}\right)\log(n) + n.$$

Then with $t = \frac{1}{2n+1}$,

$$d_n - d_{n+1} = \left(n+\tfrac{1}{2}\right)\log\left(\frac{n+1}{n}\right) - 1 = \frac{1}{2t}\log\left(\frac{1+t}{1-t}\right) - 1.$$

Now for $0 < t < 1$, if we subtract the second of the following expressions from the first:

$$\log(1+t) - t = -\tfrac{1}{2}t^2 + \tfrac{1}{3}t^3 - \tfrac{1}{4}t^4 + \cdots$$
$$\log(1-t) + t = -\tfrac{1}{2}t^2 - \tfrac{1}{3}t^3 - \tfrac{1}{4}t^4 + \cdots$$

and divide by $2t$, we get

$$\begin{aligned} d_n - d_{n+1} &= \tfrac{1}{3}t^2 + \tfrac{1}{5}t^4 + \tfrac{1}{7}t^6 + \cdots \\ &\le \tfrac{1}{3}t^2 + \tfrac{1}{3}t^4 + \tfrac{1}{3}t^6 + \cdots \\ &= \tfrac{1}{3}\frac{t^2}{1-t^2} \\ &= \tfrac{1}{3}\frac{1}{(2n+1)^2 - 1} = \tfrac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right). \end{aligned}$$

This shows that d_n is decreasing and $d_1 - d_n < \frac{1}{12}(1 - \frac{1}{n})$. So we may conclude $d_n > d_1 - \frac{1}{12} = \frac{11}{12}$. By convergence of monotone bounded sequences, d_n tends to a limit, say $d_n \to A$.

For $m > n$, $d_n - d_m < -\frac{2}{15}(\frac{1}{2n+1})^4 + \frac{1}{12}(\frac{1}{n} - \frac{1}{m})$, so we also have $A < d_n < A + \frac{1}{12n}$.

It only remains to find A.

Defining I_n, and then using integration by parts we have

$$\begin{aligned} I_n &:= \int_0^{\pi/2} \sin^n\theta\, d\theta = -\cos\theta\sin^{n-1}\theta\,\Big|_0^{\pi/2} + \int_0^{\pi/2}(n-1)\cos^2\theta\sin^{n-2}\theta\, d\theta \\ &= (n-1)(I_{n-2} - I_n). \end{aligned}$$

So $I_n = \frac{n-1}{n} I_{n-2}$, with $I_0 = \pi/2$ and $I_1 = 1$. Therefore

$$I_{2n} = \frac{1}{2}\cdot\frac{3}{4}\cdots\frac{2n-1}{2n}\cdot\frac{\pi}{2} = \frac{(2n)!}{(2^n n!)^2}\frac{\pi}{2}$$
$$I_{2n+1} = \frac{2}{3}\cdot\frac{4}{5}\cdots\frac{2n}{2n+1} = \frac{(2^n n!)^2}{(2n+1)!}.$$

For $\theta \in (0, \pi/2)$, $\sin^n \theta$ is decreasing in n, so I_n is also decreasing in n. Thus

$$1 \leq \frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} = 1 + \frac{1}{2n} \to 1.$$

By using $n! \sim n^{n+1/2}e^{-n+A}$ to evaluate the term in square brackets below,

$$\frac{I_{2n}}{I_{2n+1}} = \pi(2n+1)\left[\frac{((2n)!)^2}{2^{4n+1}(n!)^4}\right] \sim \pi(2n+1)\frac{1}{ne^{2A}} \to \frac{2\pi}{e^{2A}},$$

which is to equal 1. Therefore $A = \log(\sqrt{2\pi})$ as required. □

Notice we have actually shown that

$$n! = \left(\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}\right)e^{\epsilon(n)} = S(n)e^{\epsilon(n)}$$

where $0 < \epsilon(n) < \frac{1}{12n}$. For example, $10!/S(10) = 1.008365$ and $e^{1/120} = 1.008368$.

Example 3.4. Suppose we toss a fair coin $2n$ times. The probability of equal number of heads and tails is

$$\frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{[2^n(n!)]^2}. \approx \frac{\sqrt{2\pi}(2n)^{2n+\frac{1}{2}}e^{-2n}}{\left[2^n\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}\right]^2} = \frac{1}{\sqrt{\pi n}}.$$

For $n = 13$ this is 0.156478. The exact answer is 0.154981.

Compare this to the probability of extracting 26 cards from a shuffled deck and obtaining 13 red and 13 black. That is

$$\frac{\binom{26}{13}\binom{26}{13}}{\binom{52}{26}} = 0.2181.$$

Do you understand why this probability is greater?

3.3 Improved Stirling's formula

In fact, (see Robbins, A Remark on Stirling's Formula, 1955), we have

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12n+1}} < n! < \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12n}}.$$

We have already proved the right hand part, $d_n < A + \frac{1}{12n}$. The left hand part of this follows from

$$d_n - d_{n+1} > \tfrac{1}{3}t^2 + \tfrac{1}{3^2}t^4 + \tfrac{1}{3^3}t^6 + \cdots = \frac{t^2}{3-t^2} \geq \tfrac{1}{12}\left(\frac{1}{n+\frac{1}{12}} - \frac{1}{n+1+\frac{1}{12}}\right),$$

where one can check the final inequality using Mathematica. It implies $d_n - A > \frac{1}{12n+1}$.

For $n = 10$, $1.008300 < n!/S(n) < 1.008368$.

4 Axiomatic approach

Probability axioms. Properties of P. Boole's inequality. Probabilistic method in combinatorics. Inclusion-exclusion formula. Coincidence of derangements.

4.1 Axioms of probability

A **probability space** is a triple $(\Omega, \mathscr{F}, P)$, in which Ω is the sample space, $\mathscr{F}$ is a collection of subsets of Ω, and P is a **probability measure** $P : \mathscr{F} \to [0, 1]$.

To obtain a consistent theory we must place requirements on $\mathscr{F}$:

F$_1$: $\varnothing \in \mathscr{F}$ and $\Omega \in \mathscr{F}$.

F$_2$: $A \in \mathscr{F} \implies A^c \in \mathscr{F}$.

F$_3$: $A_1, A_2, \ldots \in \mathscr{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

Each $A \in \mathscr{F}$ is a possible event. If Ω is finite then we can take $\mathscr{F}$ to be the set of all subsets of Ω. But sometimes we need to be more careful in choosing $\mathscr{F}$, such as when Ω is the set of all real numbers.

We also place requirements on P: it is to be a real-valued function defined on $\mathscr{F}$ which satisfies three axioms (known as the **Kolmogorov axioms**):

I. $0 \le P(A) \le 1$, for all $A \in \mathscr{F}$.

II. $P(\Omega) = 1$.

III. For any countable set of events, $A_1, A_2, \ldots$, which are disjoint (i.e. $A_i \cap A_j = \varnothing$, $i \ne j$), we have
$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

$P(A)$ is called the **probability of the event** A.

Note. The event 'two heads' is typically written as {two heads} or [two heads].
One sees written $P\{\text{two heads}\}$, $P(\{\text{two heads}\})$, $P(\text{two heads})$, and $\mathbb{P}$ for P.

Example 4.1. Consider an arbitrary countable set $\Omega = \{\omega_1, \omega_2, \cdots\}$ and an arbitrary collection $(p_i, p_2, \ldots)$ of nonnegative numbers with sum $p_1 + p_2 + \cdots = 1$. Put
$$P(A) = \sum_{i:\omega_i \in A} p_i.$$

Then P satisfies the axioms.

The numbers $(p_1, p_2, \ldots)$ are called a **probability distribution**..

Remark. As mentioned above, if Ω is not finite then it may not be possible to let $\mathscr{F}$ be all subsets of Ω. For example, it can be shown that it is impossible to define a P for all possible subsets of the interval $[0,1]$ that will satisfy the axioms. Instead we define P for special subsets, namely the intervals $[a,b]$, with the natural choice of $P([a,b]) = b-a$. We then use F_1, F_2, F_3 to construct $\mathscr{F}$ as the collection of sets that can be formed from countable unions and intersections of such intervals, and deduce their probabilities from the axioms.

Theorem 4.2 (Properties of P)**.** *Axioms I–III imply the following further properties:*

(i) $P(\varnothing) = 0$. *(probability of the empty set)*

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

(iii) If $A \subseteq B$ *then* $P(A) \leq P(B)$. *(monotonicity)*

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(v) If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ *then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P(A_n).$$

Property (v) says that $P(\cdot)$ *is a continuous function.*

Proof. From II and III: $P(\Omega) = P(A \cup A^c) = P(A) + P(A^c) = 1$.

This gives (ii). Setting $A = \Omega$ gives (i).

For (iii) let $B = A \cup (B \cap A^c)$ so $P(B) = P(A) + P(B \cap A^c) \geq P(A)$.

For (iv) use $P(A \cup B) = P(A) + P(B \cap A^c)$ and $P(B) = P(A \cap B) + P(B \cap A^c)$.

Proof of (v) is deferred to §7.1. □

Remark. As a consequence of Theorem 4.2 (iv) we say that P is a **subadditive set function**, as it is one for which

$$P(A \cup B) \leq P(A) + P(B),$$

for all A, B. It is a also a *submodular function*, since

$$P(A \cup B) + P(A \cap B) \leq P(A) + P(B),$$

for all A, B. It is also a *supermodular function*, since the reverse inequality is true.

4.2 Boole's inequality

Theorem 4.3 (Boole's inequality)**.** *For any* $A_1, A_2, \ldots,$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i) \quad \left[\textit{special case is } P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)\right].$$

Proof. Let $B_1 = A_1$ and $B_i = A_i \setminus \bigcup_{k=1}^{i-1} A_k$.

Then $B_1, B_2, \ldots$ are disjoint and $\bigcup_k A_k = \bigcup_k B_k$. As $B_i \subseteq A_i$,

$$P(\bigcup_i A_i) = P(\bigcup_i B_i) = \sum_i P(B_i) \leq \sum_i P(A_i). \qquad \square$$

Example 4.4. Consider a sequence of tosses of biased coins. Let A_k be the event that the kth toss is a head. Suppose $P(A_k) = p_k$. The probability that an infinite number of heads occurs is

$$\begin{aligned} P\left(\bigcap_{i=1}^{\infty}\bigcup_{k=i}^{\infty} A_k\right) &\leq P\left(\bigcup_{k=i}^{\infty} A_k\right) \\ &\leq p_i + p_{i+1} + \cdots \quad \text{(by Boole's inequality).} \end{aligned}$$

Hence if $\sum_{i=1}^{\infty} p_i < \infty$ the right hand side can be made arbitrarily close to 0.

This proves that the probability of seeing an infinite number of heads is 0.

The reverse is also true: if $\sum_{i=1}^{\infty} p_i = \infty$ then $P(\text{number of heads is infinite}) = 1$.

Example 4.5. The following result is due to Erdős (1947) and is an example of the so-called **probabilistic method** in combinatorics.

Consider the complete graph on n vertices. Suppose for an integer k,

$$\binom{n}{k} 2^{1-\binom{k}{2}} < 1.$$

Then it is possible to color the edges red and blue so that no subgraph of k vertices has edges of just one colour. E.g. $n = 200, k = 12$.

Proof. Colour the edges at random, each as red or blue. In any subgraph of k vertices the probability that every edge is red is $2^{-\binom{k}{2}}$. There are $\binom{n}{k}$ subgraphs of k vertices. Let A_i be the event that the ith such subgraph has monochrome edges.

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i) = \binom{n}{k} \cdot 2 \cdot 2^{-\binom{k}{2}} < 1.$$

So there must be at least one way of colouring the edges so that no subgraph of k vertices has only monochrome edges. $\square$

Note. If n, k satisfy the above inequality then $n+1$ is a lower bound on the answer to the 'Party problem', i.e. what is the minimum number of guests needed to guarantee there will be either k who all know one another, or k who are all strangers to one another? The answer is the Ramsey number, $R(k,k)$. E.g. $R(3,3) = 6$, $R(4,4) = 18$.

4.3 Inclusion-exclusion formula

Theorem 4.6 (Inclusion-exclusion)**.** *For any events $A_1, \ldots, A_n$,*

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i_1<i_2}^n P(A_{i_1} \cap A_{i_2}) + \sum_{i_1<i_2<i_3}^n P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n-1} P(A_1 \cap \cdots \cap A_n). \tag{4.1}$$

Proof. The proof is by induction. It is clearly true for $n = 2$. Now use

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + P(A_2 \cup \cdots \cup A_n) - P(\textstyle\bigcup_{i=2}^n (A_1 \cap A_i))$$

and then apply the inductive hypothesis for $n - 1$. □

Example 4.7 [*Probability of derangement*]. Two packs of cards are shuffled and placed on the table. One by one, two cards are simultaneously turned over from the top of the packs. What is the probability that at some point the two revealed cards are identical?

This is a question about random permutations. A permutation of $1, \ldots, n$ is called a **derangement** if no integer appears in its natural position.

Suppose one of the $n!$ permutations is picked at random. Let A_k be the event that k is in its natural position. By the inclusion-exclusion formula

$$\begin{aligned} P\left(\bigcup_k A_k\right) &= \sum_k P(A_k) - \sum_{k_1<k_2} P(A_{k_1} \cap A_{k_2}) + \cdots + (-1)^{n-1} P(A_1 \cap \cdots \cap A_n) \\ &= n\frac{1}{n} - \binom{n}{2}\frac{1}{n}\frac{1}{n-1} + \binom{n}{3}\frac{1}{n}\frac{1}{n-1}\frac{1}{n-2} - \cdots + (-1)^{n-1}\frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!} \\ &\approx 1 - e^{-1}. \end{aligned}$$

So the probability of at least one match is about 0.632. The probability that a randomly chosen permutation is a derangement is $P(\bigcap_k A_k^c) \approx e^{-1} = 0.368$.

Example 4.8. The formula can also be used to answer a question like "what is the number of surjections from a set of A of n elements to a set B of $m \leq n$ elements?"

Answer. Let A_i be the set of those functions that do not have $i \in B$ in their image. The number of functions that miss out any given set of k elements of B is $(m-k)^n$. Hence the number of surjections is

$$S_{n,m} = m^n - \left|\bigcup_i A_i\right| = m^n - \sum_{k=1}^{m-1} (-1)^{k-1} \binom{m}{k} (m-k)^n.$$

5 Independence

Bonferroni inequalities. Independence. Important discrete distributions (Bernoulli, binomial, Poisson, geometric and hypergeometic). Poisson approximation to binomial.

5.1 Bonferroni's inequalities

Notation. We sometimes write $P(A, B, C)$ to mean the same as $P(A \cap B \cap C)$.

Bonferroni's inequalities say that if we truncate the sum on the right hand side of the inclusion-exclusion formula (4.1) so as to end with a positive (negative) term then we have an over- (under-) estimate of $P(\bigcup_i A_i)$. For example,

$$P(A_1 \cup A_2 \cup A_3) \geq P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_2A_3) - P(A_1A_3).$$

Corollary 5.1 (Bonferroni's inequalities)**.** *For any events $A_1, \ldots, A_n$, and for any r, $1 \leq r \leq n$,*

$$P\left(\bigcup_{i=1}^{n} A_i\right) \begin{matrix} \leq \\ or \\ \geq \end{matrix} \sum_{i=1}^{n} P(A_i) - \sum_{i_1<i_2}^{n} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1<i_2<i_3}^{n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{r-1} \sum_{i_1<\cdots<i_r} P(A_{i_1} \cap \cdots \cap A_{i_r})$$

as r is odd or even.

Proof. Again, we use induction on n. For $n = 2$ we have $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$. You should be able to complete the proof using the fact that

$$P\left(A_1 \cup \bigcup_{i=2}^{n} A_i\right) = P(A_1) + P\left(\bigcup_{i=2}^{n} A_i\right) - P\left(\bigcup_{i=2}^{n} (A_i \cap A_1)\right). \qquad \square$$

Example 5.2. Consider $\Omega = \{1, \ldots, m\}$, with all m outcomes equally likely. Suppose $x_k \in \{1, \ldots, m\}$ and let $A_k = \{1, 2, \ldots, x_k\}$. So $P(A_k) = x_k/m$ and $P(A_j \cap A_k) = \min\{x_j, x_k\}/m$. By applying Bonferroni inequalities we can prove results like

$$\max\{x_1, \ldots, x_n\} \geq \sum_i x_i - \sum_{i<j} \min\{x_i, x_j\},$$
$$\max\{x_1, \ldots, x_n\} \leq \sum_i x_i - \sum_{i<j} \min\{x_i, x_j\} + \sum_{i<j<k} \min\{x_i, x_j, x_k\}.$$

5.2 Independence of two events

Two events A and B are said to be **independent** if

$$P(A \cap B) = P(A)P(B).$$

Otherwise they are said to be **dependent**.

Notice that if A and B are independent then

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c),$$

so A and B^c are independent. Reapplying this result we see also that A^c and B^c are independent, and that A^c and B are independent.

Example 5.3. Two fair dice are thrown. Let A_1 (A_2) be the event that the first (second) die shows an odd number. Let A_3 be the event that the sum of the two numbers is odd. Are A_1 and A_2 independent? Are A_1 and A_3 independent?

Solution. We first calculate the probabilities of various events.

Event	Probability
A_1	$\frac{18}{36} = \frac{1}{2}$
A_2	as above, $\frac{1}{2}$
A_3	$\frac{6\times3}{36} = \frac{1}{2}$

Event	Probability
$A_1 \cap A_2$	$\frac{3\times3}{36} = \frac{1}{4}$
$A_1 \cap A_3$	$\frac{3\times3}{36} = \frac{1}{4}$
$A_1 \cap A_2 \cap A_3$	0

Thus by a series of multiplications, we can see that A_1 and A_2 are independent, A_1 and A_3 are independent (also A_2 and A_3). □

Independent experiments. The idea of 2 independent events models that of '2 independent experiments'. Consider $\Omega_1 = \{\alpha_1, \dots\}$ and $\Omega_2 = \{\beta_1, \dots\}$ with associated probability distributions $\{p_1, \dots\}$ and $\{q_1, \dots\}$. Then, by '2 independent experiments', we mean the sample space $\Omega_1 \times \Omega_2$ with probability distribution $P((\alpha_i, \beta_j)) = p_i q_j$.

Now, suppose $A \subset \Omega_1$ and $B \subset \Omega_2$. The event A can be interpreted as an event in $\Omega_1 \times \Omega_2$, namely $A \times \Omega_2$, and similarly for B. Then

$$P(A \cap B) = \sum_{\substack{\alpha_i \in A \\ \beta_j \in B}} p_i q_j = \sum_{\alpha_i \in A} p_i \sum_{\beta_j \in B} q_j = P(A)\, P(B),$$

which is why they are called 'independent' experiments. The obvious generalisation to n experiments can be made, but for an infinite sequence of experiments we mean a sample space $\Omega_1 \times \Omega_2 \times \dots$ satisfying the appropriate formula for all $n \in \mathbb{N}$.

5.3 Independence of multiple events

Events $A_1, A_2, \dots$ are said to be independent (or if we wish to emphasise '**mutually independent**') if for all $i_1 < i_2 < \cdots < i_r$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_r}).$$

Events can be pairwise independent without being (mutually) independent.

In Example 5.3, $P(A_1) = P(A_2) = P(A_3) = 1/2$ but $P(A_1 \cap A_2 \cap A_3) = 0$. So A_1, A_2 and A_3 are *not* independent. Here is another such example:

Example 5.4. Roll three dice. Let A_{ij} be the event that dice i and j show the same. $P(A_{12} \cap A_{13}) = 1/36 = P(A_{12})P(A_{13})$. But $P(A_{12} \cap A_{13} \cap A_{23}) = 1/36 \neq P(A_{12})P(A_{13})P(A_{23})$.

5.4 Important distributions

As in Example 4.1, consider a sample space $\Omega = \{\omega_1, \omega_2, \dots\}$ (which may be finite or countable). For each $\omega_i \in \Omega$ let $p_i = P(\{\omega_i\})$. Then

$$p_i \geq 0, \text{ for all } i, \text{ and } \sum_i p_i = 1. \tag{5.1}$$

A sequence $\{p_i\}_{i=1,2,\dots}$ satisfying (5.1) is called a **probability distribution**..

Example 5.5. Consider tossing a coin once, with possible outcomes $\Omega = \{H, T\}$. For $p \in [0, 1]$, the **Bernoulli distribution**, denoted $B(1, p)$, is

$$P(H) = p, \quad P(T) = 1 - p.$$

Example 5.6. By tossing the above coin n times we obtain a sequence of **Bernoulli trials**. The number of heads obtained is an outcome in the set $\Omega = \{0, 1, 2, \dots, n\}$. The probability of $HHT \cdots T$ is $ppq \cdots q$. There are $\binom{n}{k}$ ways in which k heads occur, each with probability $p^k q^{n-k}$. So

$$P(k \text{ heads}) = p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

This is the **binomial distribution**, denoted $B(n, p)$.

Example 5.7. Suppose n balls are tossed independently into k boxes such that the probability that a given ball goes in box i is p_i. The probability that there will be $n_1, \dots, n_k$ balls in boxes $1, \dots, k$, respectively, is

$$\frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}, \quad n_1 + \cdots + n_k = n.$$

This is the **multinomial distribution**.

Example 5.8. Consider again an infinite sequence of Bernoulli trials, with $P(\text{success}) = 1 - P(\text{failure}) = p$. The probability that the first success occurs after exactly k failures is $p_k = p(1-p)^k$, $k = 0, 1, \ldots$. This is the **geometric distribution** with parameter p. Since $\sum_0^\infty p_r = 1$, the probability that every trial is a failure is zero.

[You may sometimes see 'geometric distribution' used to mean the distribution of the trial on which the first success occurs. Then $p_k = p(1-p)^{k-1}$, $k = 1, 2, \ldots$.]

The geometric distribution has the **memoryless property** (but we leave discussion of this until we meet the exponential distribution in §16.3).

Example 5.9. Consider an urn with n_1 red balls and n_2 black balls. Suppose n balls are drawn without replacement, $n \le n_1 + n_2$. The probability of drawing exactly k red balls is given by the **hypergeometric distribution**

$$p_k = \frac{\binom{n_1}{k}\binom{n_2}{n-k}}{\binom{n_1+n_2}{n}}, \quad \max(0, n-n_2) \le k \le \min(n, n_1).$$

5.5 Poisson approximation to the binomial

Example 5.10. The **Poisson distribution** is often used to model the number of occurrences of some event in a specified time, such as the number of insurance claims suffered by an insurance company in a year. Denoted $P(\lambda)$, the Poisson distribution with parameter $\lambda > 0$ is

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots .$$

Theorem 5.11 (Poisson approximation to the binomial)**.** *Suppose that $n \to \infty$ and $p \to 0$ such that $np \to \lambda$. Then*

$$\binom{n}{k} p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots .$$

Proof. Recall that $(1 - \frac{a}{n})^n \to e^{-a}$ as $n \to \infty$. For convenience we write p rather than $p(n)$. The probability that exactly k events occur is

$$\begin{aligned} q_k &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{1}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} (np)^k \left(1 - \frac{np}{n}\right)^{n-k} \\ &\to \frac{1}{k!} \lambda^k e^{-\lambda}, \quad k = 0, 1, \ldots \end{aligned}$$

since $p = p(n)$ is such that as $np(n) \to \lambda$. □

Remark. Each of the distributions above is called a **discrete distribution** because it is a probability distribution over an Ω which is finite or countable.

6 Conditional probability

Conditional probability, Law of total probability, Bayes's formula. Screening test. Simpson's paradox.

6.1 Conditional probability

Suppose B is an event with $P(B) > 0$. For any event $A \subseteq \Omega$, the **conditional probability of A given B** is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

i.e. the probability that A has occurred if we know that B has occurred. Note also that

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A).$$

If A and B are independent then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Also $P(A \mid B^c) = P(A)$. So knowing whether or not B occurs does not affect probability that A occurs.

Example 6.1. Notice that $P(A \mid B) > P(A) \iff P(B \mid A) > P(B)$. We might say that A and B are 'attractive'. The reason some card games are fun is because '*good hands attract*'. In games like poker and bridge, 'good hands' tend to be those that have more than usual homogeneity, like '4 aces' or 'a flush' (5 cards of the same suit). If I have a good hand, then the remainder of the cards are more homogeneous, and so it is more likely that other players will also have good hands.

For example, in poker the probability of a royal flush is 1.539×10^{-6}. The probability the player on my right has a royal flush, given that I have looked at my cards and seen a royal flush is 1.959×10^{-6}, i.e. 1.27 times greater than before I looked at my cards.

6.2 Properties of conditional probability

Theorem 6.2.

1. $P(A \cap B) = P(A \mid B)\, P(B)$,
2. $P(A \cap B \cap C) = P(A \mid B \cap C)\, P(B \mid C)\, P(C)$,
3. $P(A \mid B \cap C) = \frac{P(A \cap B \mid C)}{P(B \mid C)}$,

4. the function $P(\circ \mid B)$ restricted to subsets of B is a probability function on B.

Proof. Results 1 to 3 are immediate from the definition of conditional probability. For result 4, note that $A \cap B \subset B$, so $P(A \cap B) \leq P(B)$ and thus $P(A \mid B) \leq 1$. $P(B \mid B) = 1$ (obviously), so it just remains to show the Axiom III. For $A_1, A_2, \ldots$ which are disjoint events and subsets of B, we have

$$P\left(\bigcup_i A_i \,\middle|\, B\right) = \frac{P(\bigcup_i A_i \cap B)}{P(B)} = \frac{P(\bigcup_i A_i)}{P(B)} = \frac{\sum_i P(A_i)}{P(B)}$$

$$= \frac{\sum_i P(A_i \cap B)}{P(B)} = \sum_i P(A_i \mid B). \qquad \square$$

6.3 Law of total probability

A (finite or countable) collection $\{B_i\}_i$ of disjoint events such that $\bigcup_i B_i = \Omega$ is said to be a **partition of the sample space** Ω. For any event A,

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A \mid B_i) P(B_i)$$

where the second summation extends only over B_i for which $P(B_i) > 0$.

Example 6.3 [*Gambler's ruin*]. A fair coin is tossed repeatedly. At each toss the gambler wins £1 for heads and loses £1 for tails. He continues playing until he reaches £a or goes broke.

Let p_x be the probability he goes broke before reaching a. Using the law of total probability:

$$p_x = \tfrac{1}{2} p_{x-1} + \tfrac{1}{2} p_{x+1},$$

with $p_0 = 1$, $p_a = 0$. Solution is $p_x = 1 - x/a$.

6.4 Bayes' formula

Theorem 6.4 (Bayes' formula)**.** *Suppose $\{B_i\}_i$ is a partition of the sample space and A is an event for which $P(A) > 0$. Then for any event B_j in the partition for which $P(B_j) > 0$,*

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_i P(A \mid B_i) P(B_i)}$$

where the summation in the denominator extends only over B_i for which $P(B_i) > 0$.

Example 6.5 [*Screening test*]. A screening test is 98% effective in detecting a certain disease when a person has the disease. However, the test yields a false positive rate of 1% of the healthy persons tested. If 0.1% of the population have the disease, what is the probability that a person who tests positive has the disease?

$$P(+ \mid D) = 0.98, \quad P(+ \mid D^c) = 0.01, \quad P(D) = 0.001.$$

$$\begin{aligned} P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\ &= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.01 \times 0.999} \qquad \approx 0.09. \end{aligned}$$

Thus of persons who test positive only about 9% have the disease.

Example 6.6 [*Paradox of the two children*].

(i) I have two children one of whom is a boy.

(ii) I have two children one of whom is a boy born on a Thursday.

Find in each case the probability that both are boys.

In case (i)

$$P(BB \mid BB \cup BG) = \frac{P(BB)}{P(BB \cup BG)} = \frac{\frac{1}{4}}{\frac{1}{4} + 2\frac{1}{2}\frac{1}{2}} = \frac{1}{3}.$$

In case (ii), a child can be a girl (G), a boy born on Thursday (B^*) or a boy not born on a Thursday (B).

$$\begin{aligned} P(B^*B^* \cup BB^* \mid B^*B^* \cup B^*B \cup B^*G) &= \frac{P(B^*B^* \cup BB^*)}{P(B^*B^* \cup B^*B \cup B^*G)} \\ &= \frac{\frac{1}{14}\frac{1}{14} + 2\frac{1}{14}\frac{6}{14}}{\frac{1}{14}\frac{1}{14} + 2\frac{1}{14}\frac{6}{14} + 2\frac{1}{14}\frac{1}{2}} = \frac{13}{27}. \end{aligned}$$

6.5 Simpson's paradox

Example 6.7 [*Simpson's paradox*]. One example of conditional probability that appears counter-intuitive when first encountered is the following situation. In practice, it arises frequently. Consider one individual chosen at random from 50 men and 50 women applicants to a particular College. Figures on the 100 applicants are given in the following table indicating whether they were educated at a state school or at an independent school and whether they were admitted or rejected.

All applicants	Admitted	Rejected	% Admitted
State	25	25	50%
Independent	28	22	56%

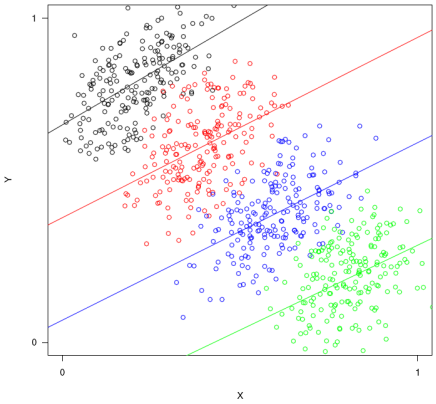
Note that overall the probability that an applicant is admitted is 0.53, but conditional on the candidate being from an independent school the probability is 0.56 while conditional on being from a state school the probability is lower at 0.50. Suppose that when we break down the figures for men and women we have the following figures.

Men applicants	Admitted	Rejected	% Admitted
State	15	22	41%
Independent	5	8	38%

Women applicants	Admitted	Rejected	% Admitted
State	10	3	77%
Independent	23	14	62%

It may now be seen that now for both men and women the conditional probability of being admitted is higher for state school applicants, at 0.41 and 0.77, respectively.

Simpson's paradox is not really a paradox, since we can explain it. Here is a graphical representation.



Scatterplot of correlation between two continuous variables X and Y, grouped by a nominal variable Z. Different colors represent different levels of Z.

It can also be understood from the fact that

$$\frac{A}{B} > \frac{a}{b} \text{ and } \frac{C}{D} > \frac{c}{d} \text{ does not imply } \frac{A+C}{B+D} > \frac{a+c}{b+d}.$$

E.g. $\{a, b, c, d, A, B, C, D\} = \{10, 10, 80, 10, 10, 5, 11, 1\}$.

Remark. It is appropriate for Cambridge students to know that this phenomenon was actually first recorded by Udny Yule (a fellow of St John's College) in 1903. It is sometimes called the **Yule-Simpson effect**.

7 Discrete random variables

Probability is a continuous set function. Definition of a discrete random variable. Distributions. Expectation. Expectation of binomial and Poisson. Function of a random variable. Properties of expectation.

7.1 Continuity of P

A sequence of events $A_1, A_2, \ldots$ is increasing (or decreasing) if

$$A_1 \subset A_2 \subset \cdots \quad (\text{or } A_1 \supset A_2 \supset \cdots).$$

We can define a limiting event

$$\lim_{n\to\infty} A_n = \bigcup_1^\infty A_n \quad \left(\text{or } = \bigcap_1^\infty A_n\right).$$

Theorem 7.1. *If $A_1, A_2, \ldots$ is an increasing or decreasing sequence of events then*

$$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n).$$

Proof. Suppose $A_1, A_2, \ldots$ is an increasing sequence. Define B_n for $n \geq 1$

$$B_1 = A_1$$
$$B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right) = A_n \cap A_{n-1}^c.$$

$(B_n,\ n \geq 1)$ are disjoint events and

$$\bigcup_{i=1}^\infty A_i = \bigcup_{i=1}^\infty B_i, \quad \bigcup_{i=1}^n A_i = \bigcup_{i=1}^n B_i$$

$$\begin{aligned} P\left(\bigcup_{i=1}^\infty A_i\right) &= P\left(\bigcup_{i=1}^\infty B_i\right) \\ &= \sum_1^\infty P(B_i) \quad \text{(axiom III)} \\ &= \lim_{n\to\infty} \sum_1^n P(B_i) \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^n A_i\right) \quad \text{(axiom III)} \\ &= \lim_{n\to\infty} P(A_n) \end{aligned}$$

Thus

$$P\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} P\left(A_n\right).$$

If $A_1, A_2, \ldots$ is a decreasing sequence then $A_1^c, A_2^c, \ldots$ is an increasing sequence. Hence

$$P\left(\lim_{n\to\infty} A_n^c\right) = \lim_{n\to\infty} P\left(A_n^c\right).$$

Use $\lim_{n\to\infty} A_n^c = (\lim_{n\to\infty} A_n)^c$. Thus *probability is a continuous set function.* □

7.2 Discrete random variables

A **random variable** (r.v.) X, taking values in a set Ω_X, is a function $X : \Omega \to \Omega_X$.

Typically $X(\omega)$ is a real number, but it might be a member of a set, like $\Omega_X = \{H, T\}$.

A r.v. is said to be a **discrete random variable** if Ω_X is finite or countable.

For any $T \subseteq \Omega_X$ we let $P(X \in T) = P(\{\omega : X(\omega) \in T\})$.

In particular, for each $x \in \Omega_X$, $P\left(X = x\right) = \sum_{\omega: X(\omega)=x} p_\omega$.

The **distribution** or **probability mass function** (p.m.f.) of the r.v. X is

$$\left(P\left(X = x\right), x \in \Omega_X\right).$$

It is a probability distribution over Ω_X. For example, if X is the number shown by the roll of a fair die, its distribution is $(P(X = i) = 1/6,\ i = 1, \ldots, 6)$. We call this the **discrete uniform distribution** over $\{1, \ldots, 6\}$.

Rolling a die twice, so $\Omega = \{(i, j), 1 \le i, j \le 6\}$, we might then define random variables X and Y by $X(i, j) = i + j$ and $Y(i, j) = \max\{i, j\}$. Here $\Omega_X = \{i, 2 \le i \le 12\}$.

Remark. It can be useful to put X as a subscript on p, as a reminder of the variable whose distribution this is; we write $p_X(x) = P(X = x)$. Also, we use the notation $X \sim B(n, p)$, for example, to indicate that X has the $B(n, p)$ distribution.

Remark. The terminology 'random variable' is somewhat inaccurate, since a random variable is neither random nor a variable. The word 'random' is appropriate because the domain of X is Ω, and we have a probability measure on subsets of Ω. Thereby we can compute $P(X \in T) = P(\{\omega : X(\omega) \in T\})$ for any T such that $\{\omega : X(\omega) \in T\} \in \mathscr{F}$.

7.3 Expectation

The **expectation** (or **mean**) of a real-valued random variable X exists, and is equal to the number

$$E\left[X\right] = \sum_{\omega\in\Omega} p_w X(\omega),$$

provided that this sum is absolutely convergent.

In practice it is calculated by summing over $x \in \Omega_X$, as follows.

$$E[X] = \sum_{\omega\in\Omega} p_w X(\omega) = \sum_{x\in\Omega_X} \sum_{\omega:X(\omega)=x} p_\omega X(\omega) = \sum_{x\in\Omega_X} x \sum_{\omega:X(\omega)=x} p_\omega$$
$$= \sum_{x\in\Omega_X} xP(X=x).$$

Absolute convergence allows the sum to be taken in any order. But if

$$\sum_{\substack{x\in\Omega_X \\ x\geq 0}} xP(X=x) = \infty \text{ and } \sum_{\substack{x\in\Omega_X \\ x<0}} xP(X=x) = -\infty$$

then $E[X]$ is undefined. When defined, $E[X]$ is always a constant.

If X is a positive random variable and if $\sum_{\omega\in\Omega} p_\omega X(\omega) = \infty$ we write $E[X] = \infty$.

Example 7.2. We calculate the expectation of some standard distributions.

Poisson. If $p_X(r) = P(X=r) = (\lambda^r/r!)e^{-\lambda}$, $r = 0, 1, \ldots$, then $E[X] = \lambda$.

$$E[X] = \sum_{r=0}^{\infty} r\frac{\lambda^r}{r!}e^{-\lambda} = \lambda e^{-\lambda}\sum_{r=1}^{\infty}\frac{\lambda^{r-1}}{(r-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

Binomial. If $p_X(r) = P(X=r) = \binom{n}{r}p^r(1-p)^{n-r}$, $r = 0, \ldots, n$, then $E[X] = np$.

$$\begin{aligned}
E[X] &= \sum_{r=0}^{n} rp^r(1-p)^{n-r}\binom{n}{r} \\
&= \sum_{r=0}^{n} r\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} \\
&= np\sum_{r=1}^{n}\frac{(n-1)!}{(r-1)!(n-r)!}p^{r-1}(1-p)^{n-r} \\
&= np\sum_{r=0}^{n-1}\frac{(n-1)!}{r!(n-1-r)!}p^r(1-p)^{n-1-r} \\
&= np\sum_{r=0}^{n-1}\binom{n-1}{r}p^r(1-p)^{n-1-r} \\
&= np.
\end{aligned}$$

7.4 Function of a random variable

Composition of $f : \mathbb{R} \to \mathbb{R}$ and X defines a new random variable $f(X)$ given by

$$f(X)(\omega) = f(X(\omega)).$$

Example 7.3. If a, b and c are constants, then $a + bX$ and $(X - c)^2$ are random variables defined by

$$\begin{aligned}(a + bX)(\omega) &= a + bX(\omega) \qquad \text{and} \\ (X - c)^2(\omega) &= (X(\omega) - c)^2.\end{aligned}$$

7.5 Properties of expectation

Theorem 7.4.

1. *If $X \geq 0$ then $E[X] \geq 0$.*
2. *If $X \geq 0$ and $E[X] = 0$ then $P(X = 0) = 1$.*
3. *If a and b are constants then $E[a + bX] = a + bE[X]$.*
4. *For any random variables X, Y then $E[X + Y] = E[X] + E[Y]$.*
   Properties 3 and 4 show that E is a linear operator.
5. *$E[X]$ is the constant which minimizes $E\left[(X - c)^2\right]$.*

Proof. 1. $X \geq 0$ means $X(\omega) \geq 0$ for all $\omega \in \Omega$. So $E[X] = \sum_{\omega \in \Omega} p_\omega X(\omega) \geq 0$.

2. If there exists $\omega \in \Omega$ with $p_\omega > 0$ and $X(\omega) > 0$ then $E[X] > 0$, therefore $P(X = 0) = 1$.

3. $E[a + bX] = \sum_{\omega \in \Omega} (a + bX(\omega))\, p_\omega = a \sum_{\omega \in \Omega} p_\omega + b \sum_{\omega \in \Omega} p_\omega X(\omega) = a + bE[X]$.

4. $\sum_\omega p(\omega)[X(\omega) + Y(\omega)] = \sum_\omega p(\omega)X(\omega) + \sum_\omega p(\omega)Y(\omega)]$.

5. $$\begin{aligned}E\left[(X - c)^2\right] &= E\Big[\underbrace{(X - E[X]}+\underbrace{E[X] - c)^2}\Big] \\ &= E\Big[(X - E[X])^2 + 2(X - E[X])(E[X] - c) + (E[X] - c)^2\Big] \\ &= E\big[(X - E[X])^2\big] + 2E\big[X - E[X]\big](E[X] - c) + (E[X] - c)^2 \\ &= E\big[(X - E[X])^2\big] + (E[X] - c)^2.\end{aligned}$$

   This is clearly minimized when $c = E[X]$. □

8 Further functions of random variables

Expectation of sum is sum of expectations. Variance. Variance of binomial, Poisson and geometric random variables. Indicator random variable. Reproof of inclusion-exclusion formula using indicator functions. *Zipf's law*.

8.1 Expectation of sum is sum of expectations

Henceforth, random variables are assumed to be real-valued whenever the context makes clear that this is required.

It is worth repeating Theorem 7.4, 4. This fact is very useful.

Theorem 8.1. *For any random variables $X_1, X_2, \ldots, X_n$, for which all the following expectations exist,*

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right].$$

Proof.

$$\sum_{\omega} p(\omega)\Big[X_1(\omega) + \cdots + X_n(\omega)\Big] = \sum_{\omega} p(\omega)X_1(\omega) + \cdots + \sum_{\omega} p(\omega)X_n(\omega). \qquad \square$$

8.2 Variance

The **variance** of a random variable X is defined as

$$\operatorname{Var} X = E\Big[(X - E[X])^2\Big],$$

(which we below show $= E\left[X^2\right] - E\left[X\right]^2$). The **standard deviation** is $\sqrt{\operatorname{Var} X}$.

Theorem 8.2 (Properties of variance)**.**

(i) $\operatorname{Var} X \geq 0$. If $\operatorname{Var} X = 0$, then $P\left(X = E\left[X\right]\right) = 1$.

Proof. From Theorem 7.4, properties 1 and 2. $\square$

(ii) If a, b are constants, $\operatorname{Var}\left(a + bX\right) = b^2 \operatorname{Var} X$.

Proof.

$$\operatorname{Var}(a + bX) = E\left[(a + bX - a - bE\left[X\right])^2\right] = b^2 E\left[(X - E\left[X\right])^2\right] = b^2 \operatorname{Var} X. \qquad \square$$

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

Proof.

$$\begin{aligned} E\Big[(X-E[X])^2\Big] &= E\left[X^2 - 2XE[X] + (E[X])^2\right] \\ &= E\left[X^2\right] - 2E[X]\,E[X] + E[X]^2 \\ &= E\left[X^2\right] - E[X]^2 \end{aligned}$$

□

Binomial. If $X \sim B(n,p)$ then $\mathrm{Var}(X) = np(1-p)$.

$$\begin{aligned} E[X(X-1)] &= \sum_{r=0}^{n} r(r-1)\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} \\ &= n(n-1)p^2\sum_{r=2}^{n}\binom{n-2}{r-2}p^{r-2}(1-p)^{(n-2)-(r-2)} \qquad = n(n-1)p^2. \end{aligned}$$

Hence $\mathrm{Var}(X) = n(n-1)p^2 + np - (np)^2 = np(1-p)$.

Poisson. If $X \sim P(\lambda)$ then $\mathrm{Var}(X) = \lambda$ (from the binomial, by letting $p \to 0$, $np \to \lambda$.) See also proof in Lecture 12.

Geometric. If X has the geometric distribution $P(X=r) = pq^r$ with $r = 0, 1, \cdots$ and $p+q=1$, then $E[X] = q/p$ and $\mathrm{Var}\,X = q/p^2$.

$$\begin{aligned} E[X] &= \sum_{r=0}^{\infty} rpq^r = pq\sum_{r=0}^{\infty} rq^{r-1} \\ &= pq\sum_{r=0}^{\infty}\frac{d}{dq}(q^r) = pq\frac{d}{dq}\Big(\frac{1}{1-q}\Big) \\ &= pq(1-q)^{-2} = \frac{q}{p}. \end{aligned}$$

The r.v. $Y = X + 1$ with the 'shifted-geometric distribution' has $E[Y] = 1/p$.

$$\begin{aligned} E\left[X^2\right] &= \sum_{r=0}^{\infty} r^2pq^r \\ &= pq\left(\sum_{r=1}^{\infty} r(r+1)q^{r-1} - \sum_{r=1}^{\infty} rq^{r-1}\right) \\ &= pq\left(\frac{2}{(1-q)^3} - \frac{1}{(1-q)^2}\right) = \frac{2q}{p^2} - \frac{q}{p} \\ \mathrm{Var}\,X &= E\left[X^2\right] - E[X]^2 = \frac{2q}{p^2} - \frac{q}{p} - \frac{q^2}{p^2} = \frac{q}{p^2}. \end{aligned}$$

Also, $\mathrm{Var}\,Y = q/p^2$, since adding a constant does not change the variance.

8.3 Indicator random variables

The **indicator function** $I[A]$ of an event $A \subset \Omega$ is the function

$$I[A](w) = \begin{cases} 1, & \text{if } \omega \in A; \\ 0, & \text{if } \omega \notin A. \end{cases} \tag{8.1}$$

$I[A]$ is a random variable. It may also be written I_A. It has the following properties.

1. $E\left[I[A]\right] = \sum_{\omega\in\Omega} p_\omega I[A](w) = P\left(A\right)$.
2. $I[A^c] = 1 - I[A]$.
3. $I[A \cap B] = I[A]I[B]$.
4. $I[A \cup B] = I[A] + I[B] - I[A]I[B]$.
   Proof.

$$I[A \cup B](\omega) = 1 \text{ if } \omega \in A \text{ or } \omega \in B$$
$$I[A \cup B](\omega) = I[A](\omega) + I[B](\omega) - I[A]I[B](\omega) \qquad \square$$

Example 8.3. Suppose $n \geq 2$ couples are seated at random around a table with men and women alternating. Let N be the number of husbands seated next to their wives. Calculate $E\left[N\right]$ and the $\mathrm{Var}(N)$.

Let $A_i = \{\text{couple } i \text{ are together}\}$.

$$N = \sum_{i=1}^{n} I[A_i]$$
$$E\left[N\right] = E\left[\sum_{i=1}^{n} I[A_i]\right] = \sum_{i=1}^{n} E\big[I[A_i]\big] = \sum_{i=1}^{n} \frac{2}{n} = n\frac{2}{n} = 2$$

$$\begin{aligned} E\left[N^2\right] &= E\left[\left(\sum_{i=1}^{n} I[A_i]\right)^2\right] = E\left[\sum_{i=1}^{n} I[A_i]^2 + 2\sum_{i<j} I[A_i]I[A_j]\right] \\ &= nE[I[A_i]^2] + n(n-1)E\left(I[A_1]I[A_2]\right) \\ E\left[I[A_i]^2\right] &= E\left[I[A_i]\right] = \frac{2}{n} \\ E\left[(I[A_1]I[A_2])\right] &= E\left[I[A_1 \cap A_2]\right] = P\left(A_1 \cap A_2\right) = P\left(A_1\right)P\left(A_2 \mid A_1\right) \\ &= \frac{2}{n}\left(\frac{1}{n-1}\frac{1}{n-1} + \frac{n-2}{n-1}\frac{2}{n-1}\right) \\ \mathrm{Var}\, N &= E\left[N^2\right] - E\left[N\right]^2 = n\frac{2}{n} + n(n-1)\frac{2}{n}\frac{2n-3}{(n-1)^2} - 2^2 = \frac{2(n-2)}{n-1}. \end{aligned}$$

8.4 Reproof of inclusion-exclusion formula

Proof. Let I_j be an indicator variable for the event A_j. Let

$$S_r = \sum_{i_1<i_2<\cdots<i_r} I_{i_1} I_{i_2} \cdots I_{i_r}$$

$$s_r = ES_r = \sum_{i_1<i_2<\cdots<i_r} P(A_{i_1} \cap \cdots \cap A_{i_r}).$$

Then

$$1 - \textstyle\prod_{j=1}^{n}(1 - I_j) = S_1 - S_2 + \cdots + (-1)^{n-1} S_n$$

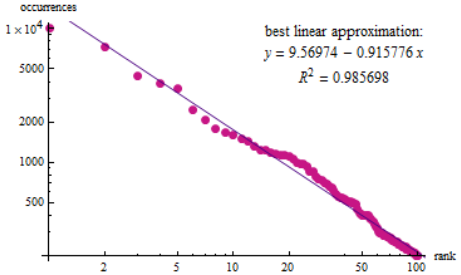
$$P\left(\textstyle\bigcup_{j=1}^{n} A_j\right) = E\left[1 - \textstyle\prod_{j=1}^{n}(1 - I_j)\right] = s_1 - s_2 + \cdots + (-1)^{n-1} s_n. \qquad \square$$

8.5 Zipf's law

(Not examinable.) The most common word in English is *the*, which occurs about one-tenth of the time in a typical text; the next most common word is *of*, which occurs about one-twentieth of the time; and so forth. It appears that words occur in frequencies proportional to their ranks. The following table is from Darwin's Origin of Species.

term	rank	occ.
the	1	10 141
of	2	7269
and	3	4364
in	4	3905
to	5	3560
a	6	2460
that	7	2081
have	8	1761
be	9	1653
as	10	1587

total number of terms: 151205
unique terms: 6915



This rule, called Zipf's Law, has also been found to apply in such widely varying places as the wealth of individuals, the size of cities, and the amount of traffic on webservers.

Suppose we have a social network of n people and the incremental value that a person obtains from other people being part of a network varies as Zipf's Law predicts. So the total value that one person obtains is proportional to $1 + 1/2 + \cdots + 1/(n-1) \approx \log n$. Since there are n people, the total value of the social network is $n \log n$.

This is empirically a better estimate of network value than Metcalfe's Law, which posits that the value of the network grows as n^2 because each of n people can connect with $n - 1$ others. It has been suggested that the misapplication of Metcalfe's Law was a contributor to the inflated pricing of Facebook shares.

9 Independent random variables

Independence of random variables and properties. Variance of a sum. Efron's dice. *Cycle lengths in a random permutation*. *Names in boxes problem*.

9.1 Independent random variables

Discrete random variables $X_1, \ldots, X_n$ are **independent** if and only if for any $x_1, \ldots, x_n$

$$P\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) = \prod_{i=1}^{n} P\left(X_i = x_i\right).$$

Theorem 9.1 (Preservation of independence)**.** *If $X_1, \ldots, X_n$ are independent random variables and $f_1, f_2 \ldots, f_n$ are functions $\mathbb{R} \to \mathbb{R}$ then $f_1(X_1), \ldots, f_n(X_n)$ are independent random variables.*

Proof.

$$P(f_1(X_1) = y_1, \ldots, f_n(X_n) = y_n) = \sum_{\substack{x_1 : f_1(x_1) = y_1 \\ \vdots \\ x_n : f_n(x_n) = y_n}} P\left(X_1 = x_1, \ldots, X_n = x_n\right)$$

$$= \prod_{i=1}^{n} \sum_{x_i : f_i(x_i) = y_i} P\left(X_i = x_i\right) = \prod_{i=1}^{n} P\left(f_i(X_i) = y_i\right). \qquad \square$$

Theorem 9.2 (Expectation of a product)**.** *If $X_1, \ldots, X_n$ are independent random variables all of whose expectations exist then:*

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E\left[X_i\right].$$

Proof. Write R_i for R_{X_i} (or Ω_{X_i}), the range of X_i.

$$E\left[\prod_{i=1}^{n} X_i\right] = \sum_{x_1 \in R_1} \cdots \sum_{x_n \in R_n} x_1 \cdots x_n P\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right)$$

$$= \prod_{i=1}^{n} \left(\sum_{x_i \in R_i} x_i P\left(X_i = x_i\right) \right) = \prod_{i=1}^{n} E\left[X_i\right]. \qquad \square$$

Notes.

(i) In Theorem 8.1 we had $E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} EX_i$ without requiring independence.

(ii) In general, Theorems 8.1 and 9.2 are not true if n is replaced by ∞.

Theorem 9.3. *If $X_1, \ldots, X_n$ are independent random variables, $f_1, \ldots, f_n$ are functions $\mathbb{R} \to \mathbb{R}$, and $\{E[f_i(X_i)]\}_i$ all exist, then:*

$$E\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} E\left[f_i(X_i)\right].$$

Proof. This follow from the previous two theorems. □

9.2 Variance of a sum

Theorem 9.4. *If $X_1, \ldots, X_n$ are independent random variables then:*

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var} X_i.$$

Proof. In fact, we only need pairwise independence.

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(E\sum_{i=1}^{n} X_i\right)^2 \\
&= E\left[\sum_i X_i^2 + \sum_{i\neq j} X_i X_j\right] - \left(\sum_i E[X_i]\right)^2 \\
&= \sum_i E\left[X_i^2\right] + \sum_{i\neq j} E\left[X_i X_j\right] - \sum_i E\left[X_i\right]^2 - \sum_{i\neq j} E\left[X_i\right] E\left[X_j\right] \\
&= \sum_i \left(E\left[X_i^2\right] - E\left[X_i\right]^2\right) \\
&= \sum_{i=1}^{n} \operatorname{Var} X_i.
\end{aligned}$$

□

Corollary 9.5. *If $X_1, \ldots, X_n$ are independent identically distributed random variables then*

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \operatorname{Var} X_i.$$

Proof.

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \operatorname{Var} \sum X_i = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var} X_i = \frac{1}{n} \operatorname{Var} X_i.$$

□

Example 9.6. If $X_1, \ldots, X_n$ are independent, identically distributed (i.i.d.) Bernoulli random variables, $\sim B(1,p)$, then $Y = X_1 + \cdots + X_n$ is a binomial random variable, $\sim B(n,p)$.

Since $\operatorname{Var}(X_i) = EX_i^2 - (EX_i)^2 = p - p^2 = p(1-p)$, we have $\operatorname{Var}(Y) = np(1-p)$.

Example 9.7 [*Experimental Design*]. Two rods of unknown lengths a, b. A rule can measure the length but with error having 0 mean (unbiased) and variance σ^2. Errors are independent from measurement to measurement. To estimate a, b we could take separate measurements A, B of each rod.

$$E[A] = a \qquad \operatorname{Var} A = \sigma^2, \qquad E[B] = b \qquad \operatorname{Var} B = \sigma^2$$

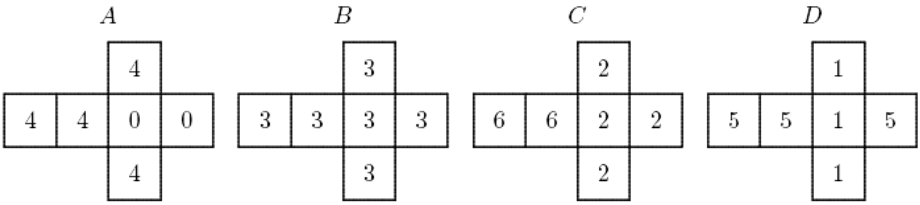
Can we do better using two measurements? Yes! Measure $a+b$ as X and $a-b$ as Y

$$E[X] = a+b, \qquad \operatorname{Var} X = \sigma^2 \qquad E[Y] = a-b, \qquad \operatorname{Var} Y = \sigma^2$$
$$E\left[\frac{X+Y}{2}\right] = a, \quad \operatorname{Var}\left(\frac{X+Y}{2}\right) = \frac{1}{2}\sigma^2$$
$$E\left[\frac{X-Y}{2}\right] = b, \quad \operatorname{Var}\left(\frac{X-Y}{2}\right) = \frac{1}{2}\sigma^2$$

So this is better.

9.3 Efron's dice

Example 9.8 [*Efron's dice*]. Consider nonstandard dice:



If each of the dice is rolled with respective outcomes A, B, C and D then

$$P(A > B) = P(B > C) = P(C > D) = P(D > A) = \frac{2}{3}.$$

It is good to appreciate that such non-transitivity can happen.

Of course we can define other ordering relations between random variables that are transitive. The ordering defined by $X \geq_E Y$ iff $EX \geq EY$, is called **expectation ordering**. The ordering defined by $X \geq \text{st}\, Y$ iff $P(X \geq t) \geq P(Y \geq t)$ for all t is called **stochastic ordering**. We say more about this in §17.3.

9.4 Cycle lengths in a random permutation

Any permutation of $1, 2, \ldots, n$ can be decomposed into cycles. For example, if (1,2,3,4) is permuted to (3,2,1,4) this is decomposed as (3,1) (2) and (4). It is the composition of one 2-cycle and two 1-cycles.

- What is the probability that a given element lies in a cycle of length m (an m-cycle)?

  Answer: $\frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdots \frac{n-m+1}{n-m+2} \cdot \frac{1}{n-m+1} = \frac{1}{n}$.

- What is the expected number of m-cycles?

  Let I_i be an indicator for the event that i is in an m-cycle.

  Answer: $\frac{1}{m} E \sum_{i=1}^{n} I_i = \frac{1}{m} n \frac{1}{n} = \frac{1}{m}$.

- Suppose $m > n/2$. Let p_m be the probability that an m-cycle exists. Since there can be at most one cycle of size $m > n/2$,

  $$p_m \cdot 1 + (1 - p_m) \cdot 0 = E(\text{number of } m\text{-cycles}) = \frac{1}{m} \implies p_m = \frac{1}{m}.$$

  Hence the probability of some large cycle of size $m > n/2$ is

  $$\sum_{m:m>n/2}^{n} p_m \leq \frac{1}{\lceil n/2 \rceil} + \cdots + \frac{1}{n} \approx \log 2 = 0.6931.$$

Names in boxes problem. Names of 100 prisoners are placed in 100 wooden boxes, one name to a box, and the boxes are lined up on a table in a room. One by one, the prisoners enter the room; each may look in at most 50 boxes, but must leave the room exactly as he found it and is permitted no further communication with the others.

The prisoners may plot their strategy in advance, and they are going to need it, because unless every prisoner finds his own name all will subsequently be executed. Find a strategy with which their probability of success exceeds 0.30.

Answer: The prisoners should use the following strategy. Prisoner i should start by looking in box i. If he finds the name of prisoner i_1 he should next look in box i_1. He continues in this manner looking through a sequence of boxes $i, i_1, i_2, \ldots, i_{49}$. His own name is contained in the box which points to the box where he started, namely i, so he will find his own name iff (in the random permutation of names in boxes) his name lies in a cycle of length ≤ 50. Every prisoners will find his name in a cycle of length ≤ 50 provided there is no large cycle. This happens with probability of $1 - 0.6931 > 0.30$.

10 Inequalities

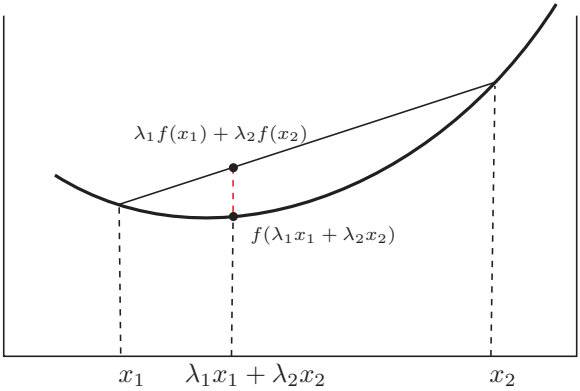
Jensen's, AM–GM and Cauchy-Schwarz inequalities. Covariance. X, Y independent $\implies \mathrm{Cov}(X,Y)=0$, but not conversely. *Information entropy*.

10.1 Jensen's inequality

A function $f:(a,b)\to\mathbb{R}$ is **convex** if for all $x_1,x_2\in(a,b)$ and $\lambda_1\geq 0$, $\lambda_2\geq 0$ with $\lambda_1+\lambda_2=1$,

$$\lambda_1 f(x_1)+\lambda_2 f(x_2)\geq f(\lambda_1 x_1+\lambda_2 x_2).$$

It is **strictly convex** if strict inequality holds when $x_1\neq x_2$ and $0<\lambda_1<1$.



chord lies above the function.

A function f is **concave** (strictly concave) if $-f$ is convex (strictly convex).

Fact. If f is a twice differentiable function and $f''(x)\geq 0$ for all $x\in(a,b)$ then f is convex [exercise in Analysis I]. It is strictly convex if $f''(x)>0$ for all $x\in(a,b)$.

Theorem 10.1 (Jensen's inequality)**.** *Let $f:(a,b)\to\mathbb{R}$ be a convex function. Then*

$$\sum_{i=1}^{n} p_i f(x_i)\geq f\left(\sum_{i=1}^{n} p_i x_i\right)$$

for all $x_1,\ldots,x_n\in(a,b)$ and $p_1,\ldots,p_n\in(0,1)$ such that $\sum_i p_i=1$. Furthermore if f is strictly convex then equality holds iff all the x_i are equal.

Jensen's inequality is saying that if X takes finitely many values then

$$E[f(X)]\geq f(E[X]).$$

Proof. Use induction. The case $n = 2$ is the definition of convexity. Suppose that the theorem is true for $n - 1$. Let $p = (p_1, \ldots, p_n)$ be a distribution (i.e. $p_i \geq 0$ for all i and $\sum_i p_i = 1$). The inductive step that proves the theorem is true for n is

$$\begin{aligned} f(p_1x_1 + \cdots + p_nx_n) &= f\left(p_1x_1 + (p_2 + \cdots + p_n)\frac{p_2x_2 + \cdots + p_nx_n}{p_2 + \cdots + p_n}\right) \\ &\leq p_1 f(x_1) + (p_2 + \cdots + p_n) f\left(\frac{p_2x_2 + \cdots + p_nx_n}{p_2 + \cdots + p_n}\right) \\ &\leq p_1 f(x_1) + (p_2 + \cdots + p_n) \sum_{i=2}^{n} \frac{p_i}{p_2 + \cdots + p_n} f(x_i) \\ &= \sum_{i=1}^{n} p_i f(x_i). \end{aligned}$$

□

10.2 AM–GM inequality

Corollary 10.2 (AM–GM inequality)**.** *Given positive real numbers $x_1, \ldots, x_n$,*

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{10.1}$$

Proof. The function $f(x) = -\log x$ is convex. Consider a random variable X such that $P(X = x_i) = 1/n$, $i = 1, \ldots, n$. By using Jensen's inequality, (10.1) follows because

$$Ef(X) \geq f(EX) \implies \frac{1}{n} \sum_i -\log x_i \geq -\log\left(\frac{1}{n} \sum_i x_i\right).$$

□

10.3 Cauchy-Schwarz inequality

Theorem 10.3. *For any random variables X and Y,*

$$E[XY]^2 \leq E[X^2]E[Y^2].$$

Proof. Suppose $EY^2 > 0$ (else $Y = 0$). Let $W = X - YE[XY]/E[Y^2]$.

$$E[W^2] = E[X^2] - 2\frac{E[XY]^2}{E[Y^2]} + \frac{E[XY]^2}{E[Y^2]} \geq 0,$$

from which the Cauchy-Schwarz inequality follows. Equality occurs only if $W = 0$. □

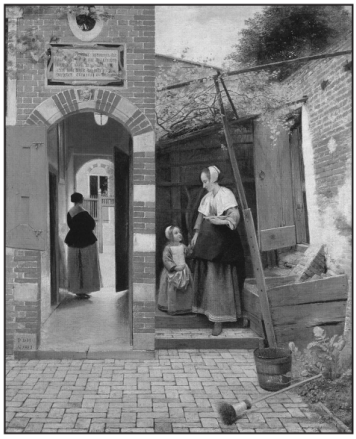


Figure 8a. *The Courtyard of a House in Delft* by Pieter de Hooch (1629–1684). Courtesty of the National Gallery, London.

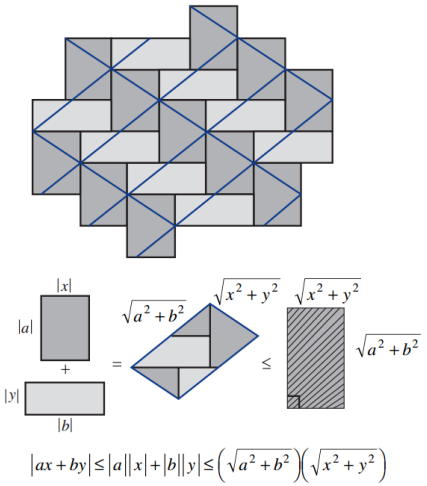


Figure 8b. The Cauchy-Schwarz inequality.

From Paintings, Plane Tilings, & Proofs, R. B. Nelsen Lewis

10.4 Covariance and correlation

For two random variable X and Y, we define the **covariance** between X and Y as

$$\mathrm{Cov}(X,Y) = E[(X - EX)(Y - EY)].$$

Properties of covariance (easy to prove, so proofs omitted) are:

- If c is a constant,
  - $\mathrm{Cov}(X, c) = 0$,
  - $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- $\mathrm{Cov}(X, Y) = EXY - EXEY$.
- $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y)$.
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.
- If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$.

  However, as the following example shows, the converse is not true.

Example 10.4. Suppose that (X, Y) is equally likely to take three possible values $(2, 0)$, $(-1, 1)$, $(-1, -1)$ Then $EX = EY = 0$ and $EXY = 0$, so $\mathrm{Cov}(X, Y) = 0$. But $X = 2 \iff Y = 0$, so X and Y are not independent.

The **correlation coefficient** (or just the correlation) between random variables X and Y with $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$ is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Corollary 10.5. $|\mathrm{Corr}(X, Y)| \leq 1$.

Proof. Apply Cauchy-Schwarz to $X - EX$ and $Y - EY$. □

10.5 Information entropy

Suppose an event A occurs with probability $P(A) = p$. How surprising is it? Let's try to invent a 'surprise function', say $S(p)$. What properties should this have?

Since a certain event is unsurprising we would like $S(1) = 0$. We should also like $S(p)$ to be decreasing and continuous in p. If A and B are independent events then we should like $S(P(A \cap B)) = S(P(A)) + S(P(B))$.

It turns out that the only function with these properties is one of the form

$$S(p) = -c \log_a p,$$

with $c > 0$. Take $c = 1$, $a = 2$. If X is a random variable that takes values $1, \ldots, n$ with probabilities $p_1, \ldots, p_n$ then on average the surprise obtained on learning X is

$$H(X) = ES(p_X) = -\sum_i p_i \log_2 p_i.$$

This is the **information entropy** of X. It is an important quantity in information theory. The 'log' can be taken to any base, but using base 2, $nH(X)$ is roughly the expected number of binary bits required to report the result of n experiments in which $X_1, \ldots, X_n$ are i.i.d. observations from distribution $(p_i, 1 \leq i \leq n)$ and we encode our reporting of the results of experiments in the most efficient way.

Let's use Jensen's inequality to prove the entropy is maximized by $p_1 = \cdots = p_n = 1/n$.

Consider $f(x) = -\log x$, which is a convex function. We may assume $p_i > 0$ for all i. Let X be a r.v. such that $X_i = 1/p_i$ with probability p_i. Then

$$-\sum_{i=1}^{n} p_i \log p_i = -Ef(X) \leq -f(EX) = -f(n) = \log n = -\sum_{i=1}^{n} \tfrac{1}{n} \log \tfrac{1}{n}.$$

11 Weak law of large numbers

Markov and Chebyshev inequalities. Weak law of large numbers. *Weierstrass approximation theorem*. *Benford's law*.

11.1 Markov inequality

Theorem 11.1. *If X is a random variable with $E|X| < \infty$ and $a > 0$, then*

$$P(|X| \geq a) \leq \frac{E|X|}{a}.$$

Proof. $I[\{|X| \geq a\}] \leq |X|/a$ (as the left-hand side is 0 or 1, and if 1 then the right-hand side is at least 1). So

$$P(|X| \geq a) = E\Big[I[\{|X| \geq a\}]\Big] \leq E\Big[|X|/a\Big] = \frac{E|X|}{a}. \qquad \square$$

11.2 Chebyshev inequality

Theorem 11.2. *If X is a random variable with $EX^2 < \infty$ and $\epsilon > 0$, then*

$$P(|X| \geq \epsilon) \leq \frac{E[X^2]}{\epsilon^2}.$$

Proof. Similarly to the proof of the Markov inequality,

$$I[\{|X| \geq \epsilon\}] \leq \frac{X^2}{\epsilon^2}.$$

Take expected value. $\square$

1. The result is "distribution free" because no assumption need be made about the distribution of X (other than $EX^2 < \infty$).
2. It is the "best possible" inequality, in the sense that for some X the inequality becomes an equality. Take $X = -\epsilon$, 0, and ϵ, with probabilities $c/(2\epsilon^2)$, $1 - c/\epsilon^2$ and $c/(2\epsilon^2)$, respectively. Then

$$EX^2 = c$$
$$P(|X| \geq \epsilon) = \frac{c}{\epsilon^2} = \frac{EX^2}{\epsilon^2}.$$

3. If $\mu = EX$ then applying the inequality to $X - \mu$ gives

$$P(|X - \mu| \geq \epsilon) \leq \frac{\operatorname{Var} X}{\epsilon^2}.$$

11.3 Weak law of large numbers

Theorem 11.3 (WLLN). *Let $X_1, X_2, \dots$ be a sequence of independent identically distributed (i.i.d.) random variables with mean μ and variance $\sigma^2 < \infty$. Let*

$$S_n = \sum_{i=1}^{n} X_i.$$

Then

$$\text{For all } \epsilon > 0,\ P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \to 0 \text{ as } n \to \infty.$$

We write this as

$$\frac{S_n}{n} \to^p \mu,$$

which reads as 'S_n/n *tends in probability to* μ'.

Proof. By Chebyshev's inequality

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) &\leq \frac{E(\frac{S_n}{n} - \mu)^2}{\epsilon^2} \\ &= \frac{E(S_n - n\mu)^2}{n^2\epsilon^2} \quad \text{(properties of expectation)} \\ &= \frac{\operatorname{Var} S_n}{n^2\epsilon^2} \quad (\text{since } ES_n = n\mu) \\ &= \frac{n\sigma^2}{n^2\epsilon^2} \quad (\text{since } \operatorname{Var} S_n = n\sigma^2) \\ &= \frac{\sigma^2}{n\epsilon^2} \to 0. \end{aligned}$$

□

Remark. We cannot relax the requirement that $X_1, X_2, \dots$ be independent. For example, we could not take $X_1 = X_2 = \cdots$, where X_1 is equally likely to be 0 or 1.

Example 11.4. Repeatedly toss a coin that comes up heads with probability p. Let A_i be the event that the ith toss is a head. Let $X_i = I[A_i]$. Then

$$\frac{S_n}{n} = \frac{\text{number of heads}}{\text{number of trials}}.$$

Now $\mu = E[I[A_i]] = P(A_i) = p$, so the WLLN states that

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \to 0 \text{ as } n \to \infty,$$

which recovers the intuitive (or frequentist) interpretation of probability.

Strong law of large numbers Why is do we use the word 'weak'? Because there is also a 'strong' form of a law of large numbers, which is

$$P\left(\frac{S_n}{n} \to \mu \text{ as } n \to \infty\right) = 1.$$

This is not the same as the weak form. What does this mean? The idea is that $\omega \in \Omega$ determines

$$\left\{\frac{S_n}{n}, \quad n = 1, 2, \dots\right\}$$

as a sequence of real numbers. Hence it either tends to μ or it does not.

$$P\left(\omega : \frac{S_n(\omega)}{n} \to \mu \text{ as } n \to \infty\right) = 1.$$

We write this as

$$\frac{S_n}{n} \to^{a.s.} \mu,$$

which is read as 'S_n/n *tends almost surely to* μ'.

11.4 Probabilistic proof of Weierstrass approximation theorem

Theorem 11.5 (not examinable)**.** *If f is a continuous real-valued function on the interval $[0, 1]$ and $\epsilon > 0$, then there exists a polynomial function p such that $|p(x) - f(x)| < \epsilon$ for all $x \in [0, 1]$.*

Proof. From Analysis I: A continuous function on $[0, 1]$ is bounded. So assume, WLOG, $|f(x)| \leq 1$. From Analysis II: A continuous function on $[0, 1]$ is uniformly continuous. This means that there exists $\delta_1, \delta_2, \dots$ such that if $x, y \in [0, 1]$ and $|x - y| < \delta_m$ then $|f(x) - f(y)| < 1/m$.

We define the so-called Bernstein polynomials:

$$b_{k,n}(x) = \tbinom{n}{k} x^k (1 - x)^{n-k}, \quad 0 \leq k \leq n.$$

Then take

$$p_n(x) = \textstyle\sum_{k=0}^{n} f(k/n) b_{k,n}(x).$$

Fix an $x \in [0, 1]$ and let X be a binomial random variable with distribution $B(n, x)$. Notice that $p_n(x) = E[f(X/n)]$. Let A be the event $\{|f(X/n) - f(x)| \geq 1/m\}$. Then

$$\begin{aligned}
\Big|p_n(x) - f(x)\Big| &= \Big|E\big[f(X/n) - f(x)\big]\Big| \\
&\leq (1/m)P(A^c) + E\Big[|f(X/n) - f(x)| \;\Big|\; A\Big]P(A) \\
&\leq (1/m) + 2P(A).
\end{aligned}$$

By using Chebyshev's inequality and the fact that $A \subseteq \{|X/n - x| \geq \delta_m\}$,

$$P(A) \leq P(|X/n - x| \geq \delta_m) \leq \frac{x(1-x)}{n\delta_m^2} \leq \frac{1}{4n\delta_m^2}$$

Now choose m and n large enough so that $\frac{1}{m} + \frac{1}{2n\delta_m^2} < \epsilon$ and we have $\left|p_n(x) - f(x)\right| < \epsilon$. We have shown this for all $x \in [0,1]$. □

11.5 Benford's law

A set of numbers satisfies **Benford's law** if the probability that a number begins with the digit k is $\log_{10}\left(\frac{k+1}{k}\right)$.

This is true, for example, of the Fibonacci numbers: $\{F_n\} = \{1, 1, 2, 3, 5, 8, \dots\}$.

Let $A_k(n)$ be the number of the first n Fibonacci numbers that begin with a k. See the table for $n = 10000$. The fit is extremely good.

k	$A_k(10000)$	$\log_{10}\left(\frac{k+1}{k}\right)$
1	3011	0.30103
2	1762	0.17609
3	1250	0.12494
4	968	0.09691
5	792	0.07918
6	668	0.06695
7	580	0.05799
8	513	0.05115
9	456	0.04576

'Explanation'. Let $\alpha = \frac{1}{2}(1 + \sqrt{5})$. If is well-known that when n is large, $F_n \approx \alpha^n$. So F_n and α^n have the same first digit. A number m begins with the digit k if the fractional part of $\log_{10} m$ lies in the interval $[\log_{10} k, \log_{10}(k+1))$. Let $]x[= x - \lfloor x \rfloor$ denote the fractional part of x. A famous theorem of Weyl states the following: If β is irrational, then the sequence of fractional parts $\{]n\beta[\}_{n=1}^{\infty}$ is uniformly distributed. This result is certainly very plausible, but a proper proof is beyond our scope. We apply this with $\beta = \log_{10} \alpha$, noting that the fractional part of $\log_{10} F_n$ is then $]n\beta[$. □

Benford's law also arises when one is concerned with numbers whose measurement scale is arbitrary. For example, whether we are measuring the areas of world lakes in km^2 or miles^2 the distribution of the first digit should surely be the same. The distribution of the first digit of X is determined by the distribution of the fractional part of $\log_{10} X$. Given a constant c, the distribution of the first digit of cX is determined by the distribution of the fractional part of $\log_{10} cX = \log_{10} X + \log_{10} c$. The uniform distribution is the only distribution on $[0,1]$ that does not change when a constant is added to it (mod 1). So if we are to have scale invariance then the fractional part of $\log_{10} X$ must be uniformly distributed, and so must lie in $[0, 0.3010]$ with probability 0.3010.

12 Probability generating functions

Distribution uniquely determined by p.g.f. Abel's lemma. The p.g.f. of a sum of random variables. Tilings. *Dyck words*.

12.1 Probability generating function

Consider a random variable X, taking values $0, 1, 2, \ldots$. Let $p_r = P(X = r)$, $r = 0, 1, 2, \ldots$. The **probability generating function** (p.g.f.) of X, or of the distribution $(p_r,\ r = 0, 1, 2, \ldots)$, is

$$p(z) = E\left[z^X\right] = \sum_{r=0}^{\infty} P(X = r)\, z^r = \sum_{r=0}^{\infty} p_r z^r.$$

Thus $p(z)$ is a polynomial or a power series. As a power series it is convergent for $|z| \le 1$, by comparison with a geometric series, and

$$|p(z)| \le \sum_r p_r |z|^r \le \sum_r p_r = 1.$$

We can write $p_X(z)$ when we wish to give a reminder that this is the p.g.f. of X.

Example 12.1 [*A die*].

$$p_r = \tfrac{1}{6}, \quad r = 1, \ldots, 6,$$
$$p(z) = E\left[z^X\right] = \tfrac{1}{6}\left(z + z^2 + \cdots + z^6\right) = \tfrac{1}{6} z \frac{1 - z^6}{1 - z}.$$

Theorem 12.2. *The distribution of X is uniquely determined by the p.g.f. $p(z)$.*

Proof. We find p_0 from $p_0 = p(0)$. We know that we can differentiate $p(z)$ term by term for $|z| \le 1$. Thus

$$p'(z) = p_1 + 2p_2 z + 3p_3 z^2 + \cdots$$
$$p'(0) = p_1.$$

Repeated differentiation gives

$$\frac{d^i}{dz^i} p(z) = p^{(i)}(z) = \sum_{r=i}^{\infty} \frac{r!}{(r-i)!} p_r z^{r-i}$$

and so $p^{(i)}(0) = i! p_i$. Thus we can recover $p_0, p_1, \ldots$ from $p(z)$. □

Theorem 12.3 (Abel's Lemma)**.**

$$E[X] = \lim_{z \to 1} p'(z).$$

Proof. First prove '$\geq$'. For $0 \leq z \leq 1$, $p'(z)$ is a nondecreasing function of z, and

$$p'(z) = \sum_{r=1}^{\infty} r p_r z^{r-1} \leq \sum_{r=1}^{\infty} r p_r = E[X],$$

so $p'(z)$ is bounded above. Hence $\lim_{z\to 1} p'(z) \leq E[X]$.

Now prove '$\leq$'. Choose $\epsilon \geq 0$. Let N be large enough that $\sum_{r=1}^{N} r p_r \geq E[X] - \epsilon$. Then

$$E[X] - \epsilon \leq \sum_{r=1}^{N} r p_r = \lim_{z\to 1} \sum_{r=1}^{N} r p_r z^{r-1} \leq \lim_{z\to 1} \sum_{r=1}^{\infty} r p_r z^{r-1} = \lim_{z\to 1} p'(z).$$

Since this is true for all $\epsilon \geq 0$, we have $E[X] \leq \lim_{z\to 1} p'(z)$. □

Usually, $p'(z)$ is continuous at $z = 1$, and then $E[X] = p'(1)$.

Similarly we have the following.

Theorem 12.4.

$$E[X(X-1)] = \lim_{z\to 1} p''(z).$$

Proof. Proof is the same as Abel's Lemma but with

$$p''(z) = \sum_{r=2}^{\infty} r(r-1) p z^{r-2}.$$ □

Example 12.5. Suppose X has the Poisson distribution with parameter λ.

$$P(X = r) = \frac{\lambda^r}{r!} e^{-\lambda}, \quad r = 0, 1, \ldots .$$

Then its p.g.f. is

$$E\left[z^X\right] = \sum_{r=0}^{\infty} z^r \frac{\lambda^r}{r!} e^{-\lambda} = e^{-\lambda z} e^{-\lambda} = e^{-\lambda(1-z)}.$$

To calculate the mean and variance of X:

$$p'(z) = \lambda e^{-\lambda(1-z)}, \quad p''(z) = \lambda^2 e^{-\lambda(1-z)}.$$

So

$$E[X] = \lim_{z\to 1} p'(z) = p'(1) = \lambda \quad \text{(since } p'(z) \text{ continuous at } z = 1)$$
$$E[X(X-1)] = p''(1) = \lambda^2$$

$$\begin{aligned} \operatorname{Var} X &= E\left[X^2\right] - E[X]^2 \\ &= E[X(X-1)] + E[X] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned}$$

Theorem 12.6. *Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables with p.g.fs $p_1(z)$, $p_2(z), \ldots,$ $p_n(z)$. Then the p.g.f. of $X_1 + X_2 + \cdots + X_n$ is*

$$p_1(z)p_2(z)\cdots p_n(z).$$

Proof.

$$\begin{aligned} E\left[z^{X_1+X_2+\cdots+X_n}\right] &= E\left[z^{X_1}z^{X_2}\cdots z^{X_n}\right] \\ &= E\left[z^{X_1}\right]E\left[z^{X_2}\right]\cdots E\left[z^{X_n}\right] \\ &= p_1(z)p_2(z)\cdots p_n(z). \end{aligned}$$

□

Example 12.7. Suppose X has a binomial distribution, $B(n, p)$. Then

$$E\left[z^X\right] = \sum_{r=0}^{n} P(X = r)z^r = \sum_{r=0}^{n} \binom{n}{r} p^r(1-p)^{n-r}z^r = (1 - p + pz)^n.$$

This proves that X has the same distribution as $Y_1 + Y_2 + \cdots + Y_n$, where $Y_1, Y_2, \ldots, Y_n$ are i.i.d. Bernoulli random variables, each with

$$P(Y_i = 0) = q = 1 - p, \quad P(Y_i = 1) = p, \quad E\left[z^{Y_i}\right] = (1 - p + pz).$$

Note. Whenever the p.g.f. factorizes it is useful to look to see if the random variable can be written as a sum of other (independent) random variables.

Example 12.8. If X and Y are independently Poisson distributed with parameters λ and μ then:

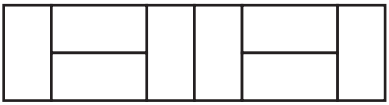
$$E\left[z^{X+Y}\right] = E\left[z^X\right]E\left[z^Y\right] = e^{-\lambda(1-z)}e^{-\mu(1-z)} = e^{-(\lambda+\mu)(1-z)},$$

which is the p.g.f. of a Poisson random variable with parameter $\lambda + \mu$. Since p.g.fs are 1–1 with distributions, $X + Y$ is Poisson distributed with parameter $\lambda + \mu$.

12.2 Combinatorial applications

Generating functions are useful in many other realms.

Tilings. How many ways can we tile a $(2 \times n)$ bathroom with (2×1) tiles?



Say f_n, where

$$f_n = f_{n-1} + f_{n-2} \qquad f_0 = f_1 = 1.$$

Let

$$F(z) = \sum_{n=0}^{\infty} f_n z^n.$$

$$f_n z^n = f_{n-1} z^n + f_{n-2} z^n \implies \textstyle\sum_{n=2}^{\infty} f_n z^n = \sum_{n=2}^{\infty} f_{n-1} z^n + \sum_{n=2}^{\infty} f_{n-2} z^n$$

and so, since $f_0 = f_1 = 1$,

$$F(z) - f_0 - z f_1 = z(F(z) - f_0) + z^2 F(z)$$
$$F(z)(1 - z - z^2) = f_0(1 - z) + z f_1 = 1 - z + z = 1.$$

Thus $F(z) = (1 - z - z^2)^{-1}$. Let

$$\alpha_1 = \tfrac{1}{2}(1 + \sqrt{5}) \qquad \alpha_2 = \tfrac{1}{2}(1 - \sqrt{5}),$$

$$F(z) = \frac{1}{(1 - \alpha_1 z)(1 - \alpha_2 z)} = \frac{1}{\alpha_1 - \alpha_2}\left(\frac{\alpha_1}{(1 - \alpha_1 z)} - \frac{\alpha_2}{(1 - \alpha_2 z)}\right)$$

$$= \frac{1}{\alpha_1 - \alpha_2}\left(\alpha_1 \textstyle\sum_{n=0}^{\infty} \alpha_1^n z^n - \alpha_2 \sum_{n=0}^{\infty} \alpha_2^n z^n\right).$$

The coefficient of z^n, that is f_n, is the **Fibonacci number**

$$f_n = \frac{1}{\alpha_1 - \alpha_2}(\alpha_1^{n+1} - \alpha_2^{n+1}).$$

Dyck words. There are 5 Dyck words of length 6:

$$()()(),\ (())(),\ ()(()),\ ((())),\ (()()).$$

In general, a **Dyck word** of length $2n$ is a balanced string of n '(' and n ')'.

Let C_n be the number of Dyck words of length $2n$. What is this?

In general, $w = (w_1)w_2$, where w, w_1, w_2 are Dyck words.

So $C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}$, taking $C_0 = 1$.

Let $c(x) = \sum_{n=0}^{\infty} C_n x^n$. Then $c(x) = 1 + x c(x)^2$. So

$$c(x) = \frac{1 - \sqrt{1 - 4x}}{2x} = \sum_{n=0}^{\infty} \binom{2n}{n} \frac{x^n}{n+1}.$$

$C_n = \frac{1}{n+1}\binom{2n}{n}$ is the nth **Catalan number**. It is the number of Dyck words of length $2n$, and also has many applications in combinatorial problems. It is the number of paths from $(0, 0)$ to $(2n, 0)$ that are always nonnegative, i.e. such that there are always at least as many ups as downs (heads as tails). We will make use of this result in a later discussion of random matrices in §24.3.

The first Catalan numbers for $n = 0, 1, 2, 3, \ldots$ are $1, 1, 2, 5, 14, 42, 132, 429, \ldots$.

13 Conditional expectation

Conditional distributions. Joint distribution. Conditional expectation and its properties. Marginals. The p.g.f. for the sum of a random number of terms. *Aggregate loss and value at risk*. *Conditional entropy*.

13.1 Conditional distribution and expectation

Let X and Y be random variables (in general, not independent) with **joint distribution**

$$P(X = x, Y = y).$$

Then the distribution of X is

$$P(X = x) = \sum_{y \in \Omega_Y} P(X = x, Y = y).$$

This is called the **marginal distribution** for X.

Assuming $P(Y = y) > 0$, the **conditional distribution** for X given by $Y = y$ is

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

The **conditional expectation** of X given $Y = y$ is,

$$E[X \mid Y = y] = \sum_{x \in \Omega_X} xP(X = x \mid Y = y).$$

We can also think of $E[X \mid Y]$ as the random variable defined by

$$E[X \mid Y](\omega) = E[X \mid Y = Y(\omega)].$$

Thus $E[X \mid Y] : \Omega \to \Omega_X$, (or $: \Omega \to \mathbb{R}$ if X is real-valued).

Example 13.1. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, with $X_i \sim B(1, p)$, and

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then

$$\begin{aligned} P(X_1 = 1 \mid Y = r) &= \frac{P(X_1 = 1, Y = r)}{P(Y = r)} = \frac{P(X_1 = 1, X_2 + \cdots + X_n = r - 1)}{P(Y = r)} \\ &= \frac{P(X_1 = 1)\, P(X_2 + \cdots + X_n = r - 1)}{P(Y = r)} \\ &= \frac{p \cdot \binom{n-1}{r-1} p^{r-1}(1-p)^{n-r}}{\binom{n}{r} p^r (1-p)^{n-r}} = \frac{\binom{n-1}{r-1}}{\binom{n}{r}} = \frac{r}{n}. \end{aligned}$$

So

$$E[X_1 \mid Y = r] = 0 \times P(X_1 = 0 \mid Y = r) + 1 \times P(X_1 = 1 \mid Y = r) = \frac{r}{n}$$

$$E[X_1 \mid Y = Y(\omega)] = \frac{1}{n} Y(\omega)$$

and therefore

$$E[X_1 \mid Y] = \frac{1}{n} Y, \quad \text{which is a random variable, i.e. a function of } Y.$$

13.2 Properties of conditional expectation

Theorem 13.2. *If X and Y are independent then*

$$E[X \mid Y] = E[X].$$

Proof. If X and Y are independent then for any $y \in \Omega_Y$

$$E[X \mid Y = y] = \sum_{x \in \Omega_X} xP(X = x \mid Y = y) = \sum_{x \in \Omega_X} xP(X = x) = E[X]. \qquad \square$$

Theorem 13.3 (**tower property of conditional expectation**)**.** *For any two random variables, X and Y,*

$$E\big[E[X \mid Y]\big] = E[X].$$

Proof.

$$\begin{aligned} E\big[E[X \mid Y]\big] &= \sum_y P(Y = y)\, E[X \mid Y = y] \\ &= \sum_y P(Y = y) \sum_x xP(X = x \mid Y = y) \\ &= \sum_y \sum_x xP(X = x, Y = y) \\ &= E[X]. \end{aligned}$$

$\square$

This is also called the **law of total expectation**. As a special case: if $A_1, \ldots, A_n$ is a partition of the sample space, then $E[X] = \sum_{i:P(A_i)>0} E[X \mid A_i]P(A_i)$.

13.3 Sums with a random number of terms

Example 13.4. Let $X_1, X_2, \ldots$ be i.i.d. with p.g.f. $p(z)$. Let N be a random variable independent of $X_1, X_2, \ldots$ with p.g.f. $h(z)$. We now find the p.g.f. of

$$S_N = X_1 + X_2 + \cdots + X_N.$$

$$
\begin{aligned}
E\left[z^{X_1+\cdots+X_N}\right] &= E\Big[E\left[z^{X_1+\cdots+X_N} \mid N\right]\Big] \\
&= \sum_{n=0}^{\infty} P\left(N=n\right) E\left[z^{X_1+\cdots+X_N} \mid N=n\right] \\
&= \sum_{n=0}^{\infty} P\left(N=n\right) \left(p(z)\right)^n \\
&= h(p(z)).
\end{aligned}
$$

Then for example

$$
E\left[X_1+\cdots+X_N\right] = \frac{d}{dz} h(p(z))\bigg|_{z=1} = h'(1)p'(1) = E\left[N\right] E\left[X_1\right].
$$

Similarly, we can calculate $\frac{d^2}{dz^2} h(p(z))$ and hence $\mathrm{Var}(X_1+\cdots+X_N)$ in terms of $\mathrm{Var}(N)$ and $\mathrm{Var}(X_1)$. This gives

$$
\mathrm{Var}(S_N) = E[N]\,\mathrm{Var}(X_1) + E[X_1]^2\,\mathrm{Var}(N).
$$

13.4 Aggregate loss distribution and VaR

A quantity of interest to an actuary is the **aggregate loss distribution** for a portfolio of insured risks and its **value at risk** (VaR). Suppose that the number of claims that will be made against a portfolio during a year is K, which is Poisson distributed with mean λ, and the severity of loss due to each claim in independent and has p.g.f. of $p(z)$. The aggregate loss is $S_K = X_1 + \cdots + X_K$. The VaR_α at $\alpha = 0.995$ is the value of x such that $P(S_K \geq x) = 0.005$. Aggregate loss of x or more occurs only 1 year in 200.

Now $p_K(z) = \exp(\lambda(z-1))$ and so $E[z^{S_K}] = \exp(\lambda(p(z)-1))$. From this we can recover $P(S_K = x)$, $x = 0, 1, \ldots$, and hence $P(S_K \geq x)$, from which we can calculate the VaR.

In practice it is more convenient to use the numerical method of fast Fourier transform and the **characteristic function**, i.e. the p.g.f. with $z = e^{i\theta}$. Suppose X_i takes values $\{0, 1, \ldots, N-1\}$ with probabilities $p_0, \ldots, p_{N-1}$. Let $\omega = e^{-i(2\pi/N)}$ and

$$
p_k^* = p(\omega^k) = \textstyle\sum_{j=0}^{N-1} p_j e^{-i\frac{2\pi k}{N} j}, \quad k = 0, \ldots, N-1.
$$

For example, suppose X_i is uniform on $\{0, 1, 2, 3\}$. Then we obtain the aggregate distribution of $X_1 + X_2$ (with range $\{0, 1, \ldots, 6\}$) by the *Mathematica* code below.

We start by padding out the distribution of X_1 so that there will be room to accommodate $X_1 + X_2$ taking values up to 6.

In squaring `ps`, we are calculating $((p_0^*)^2, \ldots, (p_6^*)^2)$, i.e. $p(z)^2$ (the p.g.f. of $X_1 + X_2$) for each $z \in \{\omega^0, \omega^1, \omega^2, \ldots, \omega^6\}$, where $\omega = e^{-i(2\pi/7)}$. From these 7 values we can

recover the 7 probabilities: $P(X_1 + X_2 = r)$, $r = 0, \ldots, 6$. For larger problems, the fast Fourier transform method is much quicker than taking powers of the p.g.f. directly. See Appendix B for more details.

```
In[537]:= ps = Fourier[{1 / 4, 1 / 4, 1 / 4, 1 / 4, 0, 0, 0},
            FourierParameters → {1, -1}] // Chop
          InverseFourier[ps^2, FourierParameters → {1, -1}] //
           Rationalize

Out[537]= {1., 0.125 - 0.547661 i, 0.125 + 0.0601968 i, 0.125 - 0.156745 i,
           0.125 + 0.156745 i, 0.125 - 0.0601968 i, 0.125 + 0.547661 i}

Out[538]= {1/16, 1/8, 3/16, 1/4, 3/16, 1/8, 1/16}
```

The VaR measure is widely used in finance, but is controversial and has drawbacks.

13.5 Conditional entropy

Suppose that X and Y are not independent. Intuitively, we think that knowledge of Y will reduce the potential surprise (entropy) inherent in X.

To show this, suppose $P(X = a_i, Y = b_j) = p_{ij}$. Denote the marginals as $P(X = a_i) = \sum_j p_{ij} = \alpha_i$. $P(Y = b_j) = \sum_i p_{ij} = \beta_j$.

The conditional probability of X given $Y = b_j$ is $P(X = a_i \mid Y = b_j) = p_{ij}/\beta_j$. Conditional on knowing $Y = b_j$,

$$H(X \mid Y = b_j) = -\sum_i \frac{p_{ij}}{\beta_j} \log \frac{p_{ij}}{\beta_j}.$$

Now average this over values of Y to get

$$H(X \mid Y) = -\sum_j \beta_j \sum_i \frac{p_{ij}}{\beta_j} \log \frac{p_{ij}}{\beta_j} = -\sum_{i,j} p_{ij} \log \frac{p_{ij}}{\beta_j}.$$

Now we show that, on average, knowledge of Y reduces entropy. This is because

$$\begin{aligned} H(X) - H(X \mid Y) &= -\sum_{i,j} p_{ij} \log \alpha_i + \sum_{ij} p_{ij} \log \frac{p_{ij}}{\beta_j} \\ &= -\sum_{i,j} p_{ij} \log \frac{\alpha_i \beta_j}{p_{ij}} \geq -\log \left(\sum_{i,j} p_{ij} \frac{\alpha_i \beta_j}{p_{ij}} \right) = 0, \end{aligned}$$

where in the final line we use Jensen's inequality.

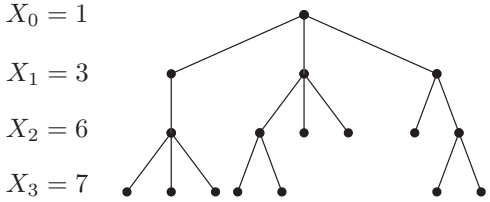
This is only true on average. It is possible that $H(X \mid Y = b_j) > H(X)$ for some j. For example, when playing Cluedo (or conducting a murder investigation), information may be obtained that increases one's uncertainty about who committed the crime.

14 Branching processes

Branching processes. Generating functions. Probability of extinction.

14.1 Branching processes

Branching processes are used to model population growth due to reproduction. Consider a sequence of random variables $X_0, X_1 \ldots$, where X_n is the number of individuals in the n^{th} generation of a population.



Assume the following.

1. $X_0 = 1$.
2. Each individual lives for unit time, and then on death produces k offspring with probability f_k, $\sum_k f_k = 1$.
3. All offspring behave independently.

   $$X_{n+1} = Y_1^n + Y_2^n + \cdots + Y_{X_n}^n,$$

   where Y_i^n are i.i.d. and Y_i^n denotes the number of offspring of the ith member of generation n.

14.2 Generating function of a branching process

Let $F(z)$ be the probability generating function of Y_i^n.

$$F(z) = E\left[z^{Y_i^n}\right] = E\left[z^{X_1}\right] = \sum_{k=0}^{\infty} f_k z^k.$$

Define

$$F_n(z) = E\left[z^{X_n}\right].$$

Then $F_1(z) = F(z)$ the probability generating function of the offspring distribution.

Theorem 14.1.

$$F_{n+1}(z) = F_n(F(z)) = F(F(\cdots(F(z))\ldots)) = F(F_n(z)).$$

Proof.

$$\begin{aligned}
F_{n+1}(z) &= E\left[z^{X_{n+1}}\right] \\
&= E\left[E\left[z^{X_{n+1}} \mid X_n\right]\right] \\
&= \sum_{k=0}^{\infty} P\left(X_n = k\right) E\left[z^{X_{n+1}} \mid X_n = k\right] \\
&= \sum_{k=0}^{\infty} P\left(X_n = k\right) E\left[z^{Y_1^n + Y_2^n + \cdots + Y_k^n}\right] \\
&= \sum_{k=0}^{\infty} P\left(X_n = k\right) E\left[z^{Y_1^n}\right] \cdots E\left[z^{Y_k^n}\right] \\
&= \sum_{k=0}^{\infty} P\left(X_n = k\right) \left(F(z)\right)^k \\
&= F_n(F(z)) \\
&= F(F(\cdots(F(z))\ldots)) \\
&= F(F_n(z)).
\end{aligned}$$

□

Theorem 14.2 (mean and variance of population size)**.** *Let*

$$EX_1 = \mu = \sum_{k=0}^{\infty} k f_k < \infty$$

$$\operatorname{Var}(X_1) = \sigma^2 = \sum_{k=0}^{\infty} (k-\mu)^2 f_k < \infty.$$

Then $E\left[X_n\right] = \mu^n$ *and*

$$\operatorname{Var} X_n = \begin{cases} \dfrac{\sigma^2 \mu^{n-1}(\mu^n - 1)}{\mu - 1}, & \mu \neq 1 \\ n\sigma^2, & \mu = 1. \end{cases} \tag{14.1}$$

Proof. Prove by calculating $F_n'(z)$, $F_n''(z)$. Alternatively

$$\begin{aligned}
E\left[X_n\right] &= E\left[E\left[X_n \mid X_{n-1}\right]\right] \quad \text{(using tower property)} \\
&= E\left[\mu X_{n-1}\right] \\
&= \mu E\left[X_{n-1}\right] \\
&= \mu^n \quad \text{(by induction)}
\end{aligned}$$

$$\begin{aligned}
E\left[\left(X_n - \mu X_{n-1}\right)^2\right] &= E\left[E\left[\left(X_n - \mu X_{n-1}\right)^2 \mid X_{n-1}\right]\right] \\
&= E\left[\operatorname{Var}\left(X_n \mid X_{n-1}\right)\right] \\
&= E\left[\sigma^2 X_{n-1}\right] \\
&= \sigma^2 \mu^{n-1}.
\end{aligned}$$

Thus

$$E\left[X_n^2\right] - 2\mu E\left[X_n X_{n-1}\right] + \mu^2 E\left[X_{n-1}^2\right] = \sigma^2 \mu^{n-1}.$$

Now calculate

$$\begin{aligned} E\left[X_n X_{n-1}\right] &= E\left[E\left[X_n X_{n-1} \mid X_{n-1}\right]\right] \\ &= E\left[X_{n-1} E\left[X_n \mid X_{n-1}\right]\right] \\ &= E\left[X_{n-1} \mu X_{n-1}\right] \\ &= \mu E\left[X_{n-1}^2\right]. \end{aligned}$$

So $E\left[X_n^2\right] = \sigma^2 \mu^{n-1} + \mu^2 E\left[X_{n-1}^2\right]$, and

$$\begin{aligned} \operatorname{Var} X_n &= E\left[X_n^2\right] - E\left[X_n\right]^2 \\ &= \mu^2 E\left[X_{n-1}^2\right] + \sigma^2 \mu^{n-1} - \mu^2 E\left[X_{n-1}\right]^2 \\ &= \mu^2 \operatorname{Var} X_{n-1} + \sigma^2 \mu^{n-1} \\ &= \mu^4 \operatorname{Var} X_{n-2} + \sigma^2(\mu^{n-1} + \mu^n) \\ &= \mu^{2(n-1)} \operatorname{Var} X_1 + \sigma^2(\mu^{n-1} + \mu^n + \cdots + \mu^{2n-3}) \\ &= \sigma^2 \mu^{n-1}(1 + \mu + \cdots + \mu^{n-1}). \end{aligned}$$

□

14.3 Probability of extinction

To deal with extinction we must be careful with limits as $n \to \infty$. Let

$$\begin{aligned} A_n &= [X_n = 0] \quad \text{(event that extinction occurs by generation } n\text{)}, \\ A &= \bigcup_{n=1}^{\infty} A_n \quad \text{(event that extinction ever occurs)}. \end{aligned}$$

We have $A_1 \subset A_2 \subset \cdots$. Let q be the **extinction probability**.

$$q = P\left(\text{extinction ever occurs}\right) = P(A) = \lim_{n\to\infty} P\left(A_n\right) = \lim_{n\to\infty} P\left(X_n = 0\right)$$

Then, since $P(X_n = 0) = F_n(0)$,

$$\begin{aligned} F(q) &= F\left(\lim_{n\to\infty} F_n(0)\right) \\ &= \lim_{n\to\infty} F\left(F_n(0)\right) \qquad \text{(since } F \text{ is continuous, a result from Analysis)} \\ &= \lim_{n\to\infty} F_{n+1}(0), \end{aligned}$$

and thus $F(q) = q$.

Alternatively, using the law of total probability,

$$q = \sum_k P\left(X_1 = k\right) P\left(\text{extinction} \mid X_1 = k\right) = \sum P\left(X_1 = k\right) q^k = F(q).$$

Theorem 14.3. *The probability of extinction, q, is the smallest positive root of the equation $F(q) = q$. Suppose μ is the mean of the offspring distribution. Then*

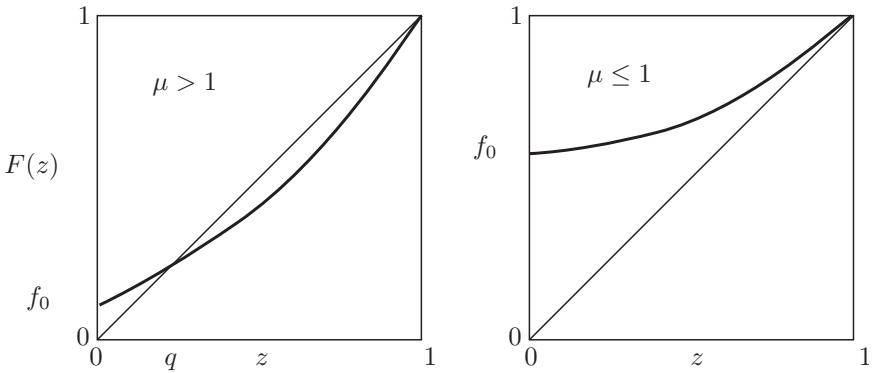
$$\text{If } \mu \leq 1 \text{ then } q = 1, \quad \text{while if } \mu > 1 \text{ then } q < 1.$$

Proof.

$$F(1) = 1, \quad \mu = \sum_{k=0}^{\infty} k f_k = \lim_{z \to 1} F'(z)$$

$$F''(z) = \sum_{k=2}^{\infty} k(k-1) z^{k-2}.$$

Assume $f_0 > 0$, $f_0 + f_1 < 1$. Then $F(0) = f_0 > 0$ and $F'(0) = f_1 < 1$. So we have the following pictures in which $F(z)$ is convex.



Thus if $\mu \leq 1$, there does not exists a $q \in (0, 1)$ with $F(q) = q$.

If $\mu > 1$ then let α be the smallest positive root of $F(z) = z$ then $\alpha \leq 1$. Further,

$$\begin{aligned} & F(0) \leq F(\alpha) = \alpha \quad \text{(since } F \text{ is increasing)} \\ \implies & F(F(0)) \leq \alpha \\ \implies & F_n(0) \leq \alpha, \quad \text{for all } n \geq 1. \end{aligned}$$

So

$$\begin{aligned} & q = \lim_{n \to \infty} F_n(0) \leq \alpha \\ \implies & q = \alpha \quad \text{(since } q \text{ is a root of } F(z) = z). \end{aligned}$$

□

15 Random walk and gambler's ruin

Random walks. Gambler's ruin. Duration of the game. Use of generating functions in random walk.

15.1 Random walks

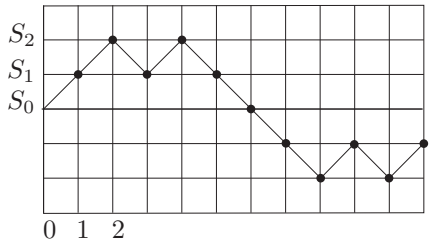
Let $X_1, X_2, \dots$ be i.i.d. r.vs such that

$$X_n = \begin{cases} +1, & \text{with probability } p, \\ -1, & \text{with probability } q = 1 - p. \end{cases}$$

Let

$$S_n = S_0 + X_1 + X_2 + \cdots + X_n$$

where usually $S_0 = 0$. Then $(S_n,\ n = 0, 1, 2, \dots)$ is a 1-dimensional **random walk**.



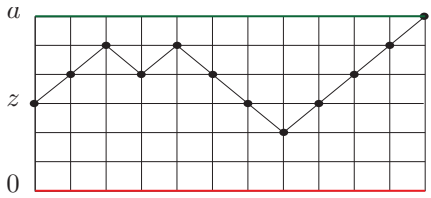
If $p = q = \frac{1}{2}$ then we have a **symmetric random walk**.

15.2 Gambler's ruin

Example 15.1. A gambler starts with an initial fortune of £z, $z < a$ and plays a game in which at successive goes he wins or loses £1 with probabilities p and q, respectively. What is the probability he is bankrupt before reaching a?

This is a random walk starting at z which stops when it hits 0 or a. Let

$$p_z = P(\text{the random walk hits } a \text{ before it hits } 0 \mid \text{start from } z),$$
$$q_z = P(\text{the random walk hits } 0 \text{ before it hits } a \mid \text{start from } z).$$



After the first step the gambler's fortune is either $z+1$ or $z-1$, with probabilities p and q respectively. From the law of total probability

$$p_z = qp_{z-1} + pp_{z+1}, \quad 0 < z < a.$$

Also $p_0 = 0$ and $p_a = 1$. We now solve $pt^2 - t + q = 0$.

$$(pt - q)(t - 1) = 0 \implies t = 1 \text{ or } q/p.$$

The general solution for $p \neq q$ is

$$p_z = A + B\,(q/p)^z$$

and so with the boundary conditions we get

$$p_z = \frac{1 - (q/p)^z}{1 - (q/p)^a}.$$

If $p = q$, the general solution is $A + Bz$ and so

$$p_z = z/a.$$

To calculate q_z, observe that this is the same problem with p, q, z replaced by $q, p, a - z$ respectively. Thus

$$P(\text{hits 0 before } a) = q_z = \begin{cases} \dfrac{(q/p)^z - (q/p)^a}{1 - (q/p)^a}, & \text{if } p \neq q \\ \dfrac{a - z}{a}, & \text{if } p = q. \end{cases}$$

Thus $q_z + p_z = 1$ and so on, as we would expect, the game ends with probability one.

What happens as $a \to \infty$?

$$\begin{aligned} P(\text{path hits 0 ever}) &= P\left(\bigcup_{a=z+1}^{\infty} \{\text{path hits 0 before it hits } a\} \right) \\ &= \lim_{a \to \infty} P\left(\text{path hits 0 before it hits } a\right) \\ &= \lim_{a \to \infty} q_z = \begin{cases} (q/p)^z, & p > q \\ 1, & p \leq q. \end{cases} \end{aligned} \tag{15.1}$$

Let G be the ultimate gain or loss.

$$G = \begin{cases} a - z, & \text{with probability } p_z \\ -z, & \text{with probability } q_z. \end{cases}$$

$$E\left[G\right] = \begin{cases} ap_z - z, & \text{if } p \neq q \\ 0, & \text{if } p = q. \end{cases}$$

Notice that a fair game remains fair: if the coin is fair ($p = q$) then games based on it have expected reward 0.

15.3 Duration of the game

Let D_z be the expected time until the random walk hits 0 or a, starting from z.

D_z is finite, because D_z/a is bounded above by $1/(p^a+q^a)$; this is the mean of geometric random variables (number of windows of size a needed until obtaining a window with all $+1$s or -1s). Consider the first step. By the law of total probability

$$\begin{aligned} D_z = E[\text{duration}] &= E\big[E\left[\text{duration} \mid X_1\right]\big] \\ &= p\,E\left[\text{duration} \mid X_1 = 1\right] + q\,E\left[\text{duration} \mid X_1 = -1\right] \\ &= p(1 + D_{z+1}) + q(1 + D_{z-1}) \\ &= 1 + pD_{z+1} + qD_{z-1}. \end{aligned}$$

This holds for $0 < z < a$ with $D_0 = D_a = 0$.

Let's try for a particular solution $D_z = Cz$.

$$\begin{aligned} Cz &= 1 + pC(z+1) + qC(z-1) \\ \implies C &= \frac{1}{q-p} \quad \text{for } p \neq q. \end{aligned}$$

Consider the homogeneous relation

$$pt^2 - t + q = 0, \quad \text{with roots 1 and } q/p.$$

If $p \neq q$ the general solution is

$$D_z = A + B\,(q/p)^z + \frac{z}{q-p}.$$

Substitute $z = 0$ and $z = a$ to get A and B, hence

$$D_z = \frac{z}{q-p} - \frac{a}{q-p}\frac{1-(q/p)^z}{1-(q/p)^a}, \quad p \neq q.$$

If $p = q$ then a particular solution is $-z^2$. General solution

$$D_z = -z^2 + A + Bz.$$

Substituting the boundary conditions given,

$$D_z = z(a-z), \quad p = q.$$

Example 15.2. Initial capital is z and we wish to reach a before 0.

p	q	z	a	P (ruin)	E [gain]	E [duration]
0.5	0.5	90	100	0.100	0	900.00
0.45	0.55	9	10	0.210	−1.101	11.01
0.45	0.55	90	100	0.866	−76.556	765.56
0.45	0.55	900	1000	≈ 1	−900	9000

15.4 Use of generating functions in random walk

Let's stop the random walk when it hits 0 or a, giving **absorption** at 0 or a. Let

$$U_{z,n} = P\left(\text{r.w. absorbed at 0 at } n \mid \text{starts at } z\right).$$

So

$$\begin{aligned} U_{0,0} &= 1, \\ U_{z,0} &= 0, \quad 0 < z \leq a, \\ U_{0,n} &= U_{a,n} = 0, \quad n > 0. \end{aligned}$$

Consider the generating function

$$U_z(s) = \sum_{n=0}^{\infty} U_{z,n} s^n.$$

Take the recurrence:

$$U_{z,n+1} = pU_{z+1,n} + qU_{z-1,n}, \quad 0 \leq z \leq a,\ n \geq 0,$$

multiply by s^{n+1} and sum over $n = 0, 1, 2 \ldots$, to obtain

$$U_z(s) = psU_{z+1}(s) + qsU_{z-1}(s),$$

where $U_0(s) = 1$ and $U_a(s) = 0$. We look for a solution of the form

$$U_z(s) = \lambda(s)^z,$$

which must satisfy

$$\lambda(s) = ps\lambda(s)^2 + qs.$$

There are two roots:

$$\lambda_1(s),\ \lambda_2(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2ps}.$$

Every solution is of the form

$$U_z(s) = A(s)\lambda_1(s)^z + B(s)\lambda_2(s)^z.$$

Substitute $U_0(s) = 1$ and $U_a(s) = 0$ to find $A(s) + B(s) = 1$ and

$$A(s)\lambda_1(s)^a + B(s)\lambda_2(s)^a = 0.$$

$$U_z(s) = \frac{\lambda_1(s)^a \lambda_2(s)^z - \lambda_1(s)^z \lambda_2(s)^a}{\lambda_1(s)^a - \lambda_2(s)^a}.$$

But $\lambda_1(s)\lambda_2(s) = q/p$ so

$$U_z(s) = (q/p)^z \frac{\lambda_1(s)^{a-z} - \lambda_2(s)^{a-z}}{\lambda_1(s)^a - \lambda_2(s)^a}.$$

Clearly $U_z(1) = q_z$. The same method will find the generating function for absorption probabilities at a, say $V_z(s)$. The generating function for the duration of the game is the sum of these two generating functions. So $D_z = U_z'(1) + V_z'(1)$.

16 Continuous random variables

Continuous random variables. Density function. Distribution function. Uniform distribution. Exponential distribution, and its memoryless property. *Hazard rate*. Relations among probability distributions.

16.1 Continuous random variables

Thus far, we have been considering experiments in which the set of possible outcomes, Ω, is finite or countable. Now we permit a continuum of possible outcomes.

For example, we might spin a pointer and let $\omega \in \Omega$ give the angular position at which it stops, with $\Omega = \{\omega : 0 \leq \omega \leq 2\pi\}$. We wish to define a probability measure P on some subsets of Ω. A sensible choice of P for a subset $[0, \theta]$ is

$$P(\omega \in [0, \theta]) = \frac{\theta}{2\pi}, \quad 0 \leq \theta \leq 2\pi.$$

Definition 16.1. A **continuous random variable** X is a real-valued function $X : \Omega \to \mathbb{R}$ for which

$$P(a \leq X(\omega) \leq b) = \int_a^b f(x)\, dx,$$

where f is a function satisfying

1. $f(x) \geq 0$,
2. $\int_{-\infty}^{+\infty} f(x)\, dx = 1$.

The function f is called the **probability density function** (p.d.f.).

For example, if $X(\omega) = \omega$ is the position at which the pointer stops then X is a continuous random variable with p.d.f.

$$f(x) = \begin{cases} \dfrac{1}{2\pi}, & 0 \leq x \leq 2\pi \\ 0, & \text{otherwise.} \end{cases}$$

Here X is a uniformly distributed random variable; we write $X \sim U[0, 2\pi]$.

Intuition about probability density functions and their uses can be obtained from the approximate relation:

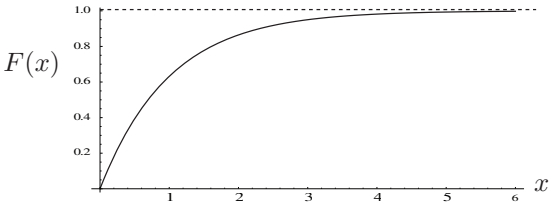
$$P(X \in [x, x + \delta x]) = \int_x^{x+\delta x} f(z)\, dz \approx f(x)\, \delta x.$$

However, remember that $f(x)$ is not a probability. Indeed, $P(X = x) = 0$ for $x \in \mathbb{R}$. So by Axiom III we must conclude $P(X \in A) = 0$ if A is any countable subset of $\mathbb{R}$.

The **cumulative distribution function** (c.d.f.) (or just **distribution function**) of a random variable X (discrete, continuous or otherwise), is defined as

$$F(x) = P\left(X \leq x\right).$$

$F(x)$ is increasing in x and tends to 1.



c.d.f. of $F(x) = 1 - e^{-x}$ for exponential r.v. $\mathscr{E}(1)$

If X is a continuous random variable then

$$F(x) = \int_{-\infty}^{x} f(z)\, dz,$$

and so F is continuous and differentiable.

In fact, the name "continuous random variable" derives from the fact that F is continuous, (though this is actually a shortened form of "absolutely continuous"; the qualifier "absolutely" we leave undefined, but it equivalent to saying that a p.d.f. exists).

We have

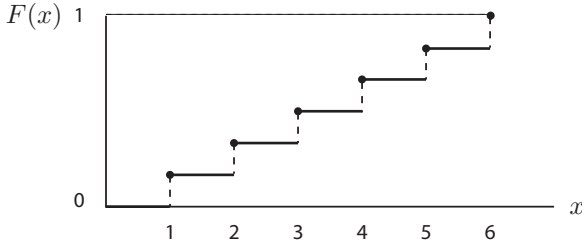
$$F'(x) = f(x)$$

at any point x where the fundamental theorem of calculus applies.

The distribution function is also defined for a discrete random variable,

$$F(x) = \sum_{\omega : X(\omega) \leq x} p_\omega$$

in which case F is a step function.



c.d.f. for X = number shown by rolling a fair die

In either case

$$P\left(a < X \leq b\right) = P\left(X \leq b\right) - P\left(X \leq a\right) = F(b) - F(a).$$

Remark. There exist random variables that are neither discrete or continuous. For example, consider a r.v. with the c.d.f.

$$F(x) = \begin{cases} x, & 0 \le x \le 1/2, \\ 1/2, & 1/2 \le x < 1, \\ 1, & x = 1. \end{cases}$$

The sample space is not countable (so not a discrete r.v.), and there is no p.d.f. (so not a continuous r.v.). There is an **atom** at $x = 1$, as $P(X = 1) = 1/2$. Only discrete r.vs have a p.m.f. and only continuous r.vs have a p.d.f. All random variables have a c.d.f.

16.2 Uniform distribution

The **uniform distribution** on $[a, b]$ has the c.d.f., and corresponding p.d.f.

$$F(x) = \frac{x-a}{b-a}, \qquad f(x) = \frac{1}{b-a}, \quad a \le x \le b.$$

If X has this distribution we write $X \sim U[a, b]$.

16.3 Exponential distribution

The **exponential distribution with parameter** λ has the c.d.f. and p.d.f.

$$F(x) = 1 - e^{-\lambda x}, \qquad f(x) = \lambda e^{-\lambda x}, \quad 0 \le x < \infty.$$

If X has this distribution we write $X \sim \mathscr{E}(\lambda)$. Note that X is nonnegative.

An important fact about the exponential distribution it that it has the **memoryless property**. If $X \sim \mathscr{E}(\lambda)$ then

$$P\left(X \ge x + z \mid X \ge z\right) = \frac{P\left(X \ge x + z\right)}{P\left(X \ge z\right)} = \frac{e^{-\lambda(x+z)}}{e^{-\lambda z}} = e^{-\lambda x} = P\left(X \ge x\right).$$

If X were the life of something, such as '*how long this phone call to my mother will last*', the memoryless property says that after we have been talking for 5 minutes, the distribution of the remaining duration of the call is just the same as it was at the start. This is close to what happens in real life.

If you walk in Cambridge on a busy afternoon the distribution of the time until you next run into a friend is likely to be exponentially distributed. Can you explain why?

The discrete distribution with the memoryless property is the geometric distribution. That is, for positive integers k and h,

$$P\left(X \ge k + h \mid X \ge k\right) = \frac{P\left(X \ge k + h\right)}{P\left(X \ge k\right)} = \frac{q^{k+h}}{q^k} = q^h = P(X \ge h).$$

16.4 Hazard rate

In modelling the lifetimes of components (or humans) a useful notion is the **hazard rate**, defined as

$$h(x) = \frac{f(x)}{1 - F(x)}.$$

Its meaning is explained by

$$P(X \le x + \delta x \mid X > x) = \frac{P(x < X \le x + \delta x)}{P(X > x)} \approx \frac{f(x)\,\delta x}{1 - F(x)} = h(x)\,\delta x.$$

The distribution having constant $h(x) = \lambda$ is the exponential distribution $\mathscr{E}(\lambda)$.

In actuarial science the hazard rate of human life is known as the 'force of mortality'.

16.5 Relationships among probability distributions

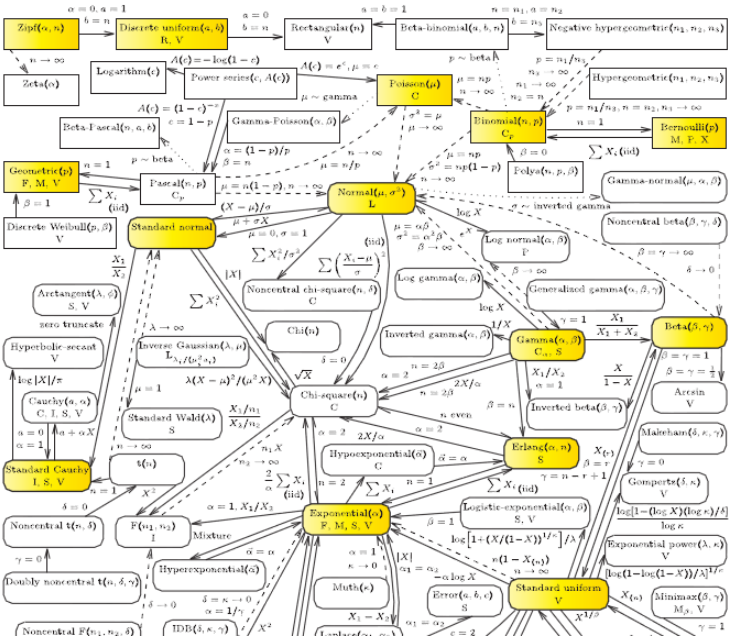
We have met a large number of distributions: Bernoulli, binomial, Poisson, geometric, hypergeometric, Zipf, Benford, uniform, exponential. By the end of the course we will add gamma, Cauchy, beta, normal and multivariate normal.

It is interesting and useful to know how distributions relate to one another. For example, the Poisson is the limit of the binomial as $n \to \infty$ and $np \to \lambda$. And if X and Y are i.i.d. exponential r.vs then $X/(X+Y)$ is uniformly distributed.

At the right is a partial view of Univariate Distribution Relationships, by L. M. Leemis and J. T. McQueston, *The American Statistician*, 2008, 62:45–53. Click above to see the whole picture.

I have highlighted the distributions we meet in this course.

See also the Wikipedia page: Relationships among probability distributions.



17 Functions of a continuous random variable

Distribution of a function of a random variable. Expectation and variance of a continuous random variable. Stochastic ordering. Inspection paradox.

17.1 Distribution of a function of a random variable

Theorem 17.1. *If X is a continuous random variable with p.d.f. $f(x)$ and $h(x)$ is a continuous strictly increasing function with $h^{-1}(x)$ differentiable then $Y = h(X)$ is a continuous random variable with p.d.f.*

$$f_Y(x) = f\left(h^{-1}(x)\right) \frac{d}{dx} h^{-1}(x).$$

Proof. The distribution function of $Y = h(X)$ is

$$P\left(h(X) \leq x\right) = P\left(X \leq h^{-1}(x)\right) = F\left(h^{-1}(x)\right),$$

since h is strictly increasing and F is the distribution function of X. Then

$$\frac{d}{dx} P\left(h(X) \leq x\right)$$

is the p.d.f., which is, as claimed, f_Y. □

Note. It is easier to repeat this proof when you need it than to remember the result.

Example 17.2. Suppose X is uniformly distributed on $[0, 1]$. Consider $Y = -\log X$.

$$P\left(Y \leq y\right) = P\left(-\log X \leq y\right) = P\left(X \geq e^{-y}\right) = \int_{e^{-y}}^{1} 1 \, dx = 1 - e^{-y}.$$

This is $F(y)$ for Y having the exponential distribution with parameter 1.

More generally, we have the following.

Theorem 17.3. *Let $U \sim U[0, 1]$. For any strictly increasing and continuous distribution function F, the random variable X defined by $X = F^{-1}(U)$ has distribution function F.*

Proof.

$$P\left(X \leq x\right) = P\left(F^{-1}(U) \leq x\right) = P\left(U \leq F(x)\right) = F(x).$$ □

Remarks. (i) This is true when F is not strictly increasing, provided we define $F^{-1}(u) = \inf\{x : F(x) \geq u,\ 0 < u < 1\}$.

(ii) This can also be done (but a bit more messily) for discrete random variables:

$$P(X = x_i) = p_i, \quad i = 0, 1, \ldots$$

Let

$$X = x_j \text{ if } \sum_{i=0}^{j-1} p_i \leq U < \sum_{i=0}^{j} p_i, \quad U \sim U[0,1].$$

This is useful when writing a computer simulation in which there are random events.

17.2 Expectation

The **expectation** (or mean) of a continuous random variable X is

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

provided not both of $\int_0^\infty x f(x)\, dx$ and $\int_{-\infty}^0 x f(x)\, dx$ are infinite.

Theorem 17.4. *If X is a continuous random variable then*

$$E[X] = \int_0^\infty P(X \geq x)\ dx - \int_0^\infty P(X \leq -x)\ dx.$$

Proof.

$$\int_0^\infty P(X \geq x)\ dx = \int_0^\infty \left[\int_x^\infty f(y)\, dy\right] dx = \int_0^\infty \int_0^\infty I[y \geq x] f(y)\, dy\, dx$$
$$= \int_0^\infty \int_0^y dx f(y)\, dy = \int_0^\infty y f(y)\, dy.$$

Similarly, $\int_0^\infty P(X \leq -x)\ dx = -\int_{-\infty}^0 y f(y)\, dy$. The result follows. □

In the case $X \geq 0$ this is $EX = \int_0^\infty (1 - F(x))\, dx$.

Example 17.5. Suppose $X \sim \mathscr{E}(\lambda)$. Then $P(X \geq x) = \int_x^\infty \lambda e^{-\lambda t} dt = e^{-\lambda x}$.

Thus $EX = \int_0^\infty e^{-\lambda x}\, dx = 1/\lambda$.

This also holds for discrete random variables and is often **a very useful way to compute expectation**, whether the random variable is discrete or continuous.

If X takes values in the set $\{0, 1, \dots\}$ the theorem states that

$$E[X] = \sum_{n=0}^{\infty} P(X > n) \quad \left(= \sum_{n=1}^{\infty} P(X \geq n) \right).$$

A direct proof of this is as follows:

$$\sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} I[m > n] P(X = m) = \sum_{m=0}^{\infty} \left(\sum_{n=0}^{\infty} I[m > n] \right) P(X = m)$$
$$= \sum_{m=0}^{\infty} mP(X = m) = EX.$$

This result is very useful and well worth remembering!

17.3 Stochastic ordering of random variables

For two random variables X and Y, we say X *is stochastically greater than* Y and write $X \geq_{\text{st}} Y$ if $P(X > t) \geq P(Y > t)$ for all t. Using the result above.

$$X \geq_{\text{st}} Y \implies \sum_{k=0}^{\infty} P(X > k) \geq \sum_{k=0}^{\infty} P(Y > k) \implies EX \geq EY.$$

So stochastic ordering implies expectation ordering.

This is also true for continuous random variables. For example, suppose X and Y are exponential random variables with parameters $1/2$ and 1.

Then $P(X > t) = e^{-t/2} > e^{-t} = P(Y > t)$. So $X \geq_{\text{st}} Y$. Also $EX = 2 > 1 = EY$.

17.4 Variance

The **variance** of a continuous random variable X is defined as for discrete r.vs,

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

The properties of expectation and variance are the same for discrete and continuous random variables; just replace $\sum$ with $\int$ in the proofs.

Example 17.6.

$$\operatorname{Var} X = E\left[X^2\right] - E[X]^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \left(\int_{-\infty}^{\infty} x f(x)\, dx \right)^2.$$

Example 17.7. Suppose $X \sim U[a, b]$. Then

$$EX = \int_a^b x \frac{1}{b-a}\, dx = \tfrac{1}{2}(a+b), \quad \operatorname{Var} X = \int_a^b x^2 \frac{dx}{b-a}\, dx - (EX)^2 = \tfrac{1}{12}(b-a)^2.$$

17.5 Inspection paradox

Suppose that n families have children attending a school. Family i has X_i children at the school, where $X_1, \ldots, X_n$ are i.i.d. r.vs, with $P(X_i = k) = p_k$, $k = 1, \ldots, m$. The average family size is μ. A child is picked at random. What is the probability distribution of the size of the family from which she comes? Let J be the index of the family from which she comes.

$$P(X_j = k \mid J = j) = \frac{P(J = j, X_j = k)}{P(J = j)} = E\left[\frac{p_k \frac{k}{k+\sum_{i\neq j} X_i}}{1/n}\right]$$

which does not depend on j. Thus

$$\frac{P(X_J = k)}{P(X_1 = k)} = E\left[\frac{n}{1 + \sum_{i\neq j} X_i/k}\right] \geq k\frac{n}{k + (n-1)\mu} \quad \text{(by Jensen's inequality).}$$

So

$$\frac{P(X_J = k)}{P(X_1 = k)} \quad \text{is increasing in } k \text{ and greater than 1 for } k > \mu.$$

Using the fact that $\frac{a}{A} \leq \frac{a+b}{A+B}$ when $\frac{a}{A} \leq \frac{b}{B}$ (any $a, b, A, B > 0$)

$$\frac{\sum_{k=1}^m P(X_J = k)}{\sum_{k=1}^m P(X_1 = k)} = \frac{1}{1} \implies \frac{\sum_{k=2}^m P(X_J = k)}{\sum_{k=2}^m P(X_1 = k)} \geq 1$$

$$\vdots$$

$$\implies \frac{\sum_{k=i}^m P(X_J = k)}{\sum_{k=i}^m P(X_1 = k)} \geq 1$$

and hence $P(X_J \geq i) \geq P(X_1 \geq i)$ for all i. So we have proved that X_J is stochastic greater than X_1. From §17.3, this implies that the means are also ordered: $EX_J \geq EX_1 = \mu$. The fact that the family size of the randomly chosen student tends to be greater that that of a normal family is known as the **inspection paradox**.

One also sees this when buses have an average interarrival interval of μ minutes. Unless the interarrival time is exactly μ minutes, the average waiting of a randomly arriving person will be more than $\mu/2$ minutes. This is because she is more likely to arrive in a large inter-bus gap than in a small one. In fact, it can be shown that the average wait will be $(\mu^2 + \sigma^2)/(2\mu)$, where σ^2 is the variance of the interarrival time.

Coda. What do you think of this claim? 'Girls have more brothers than boys do.'

'Proof'. Condition on the family having b boys and g girls. Each girl has b brothers; each boy has $b - 1$ brothers. Now take expected value over all possible b, g. □

18 Jointly distributed random variables

Jointly distributed random variables. Use of pictures when working with uniform distributions. Geometric probability. Bertrand's paradox. Buffon's needle.

18.1 Jointly distributed random variables

For two random variables X and Y the **joint distribution function** is

$$F(x,y) = P\left(X \le x, Y \le y\right), \qquad F : \mathbb{R}^2 \to [0,1].$$

The **marginal distribution** of X is

$$\begin{aligned} F_X(x) &= P\left(X \le x\right) = P\left(X \le x, Y < \infty\right) = F(x,\infty) \\ &= \lim_{y\to\infty} F(x,y). \end{aligned}$$

Similarly, $F_Y(x) = F(\infty, y)$.

We say that $X_1, X_2, \ldots, X_n$ are **jointly distributed continuous random variables**, and have **joint probability density function** f, if for any set $A \subseteq \mathbb{R}^n$

$$P\big((X_1, X_2, \ldots, X_n) \in A\big) = \iint \ldots \int_{(x_1,\ldots,x_n)\in A} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n,$$

and f satisfies the obvious conditions:

$$f(x_1, \ldots, x_n) \ge 0,$$

$$\iint \ldots \int_{\mathbb{R}^n} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1.$$

Example 18.1. The joint p.d.f. when $n = 2$ can be found from the joint distribution.

$$F(x,y) = P\left(X \le x, Y \le y\right) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\, du\, dv$$

and so

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\, \partial y}, \quad \text{provided this is defined at } (x,y).$$

Theorem 18.2. *If X and Y are jointly continuous random variables then they are individually continuous random variables.*

Proof. Since X and Y are jointly continuous random variables

$$P(X \in A) = P(X \in A, Y \in (-\infty, +\infty)) = \int_{x \in A} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy$$
$$= \int_{x \in A} f_X(x)\, dx,$$

where $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$ is the p.d.f. of X. □

18.2 Independence of continuous random variables

The notion of independence is defined in a similar manner as it is defined for discrete random variables. Continuous random variables $X_1, \ldots, X_n$ are independent if

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2) \cdots P(X_n \in A_n)$$

for all $A_i \subseteq \Omega_{X_i}$, $i = 1, \ldots, n$.

[**Note**. Each A_i is assumed measurable, i.e. we can compute $P(X_i \in A_i)$ by using the probability axioms and the fact that for any interval, $P(X_i \in [a,b]) = \int_a^b f(x)\, dx$.]

Let F_{X_i} and f_{X_i} be the c.d.f. and p.d.f. of X_i. Independence is equivalent to the statement that for all $x_1, \ldots, x_n$ the joint distribution function factors into the product of the marginal distribution functions:

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

It is also equivalent to the statement that the joint p.d.f. factors into the product of the marginal densities:

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Theorem 9.2 stated that for independent discrete random variables $E\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n E[X_i]$. By the above, we see that this also holds for independent continuous random variables. Similarly, it is true that $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i)$.

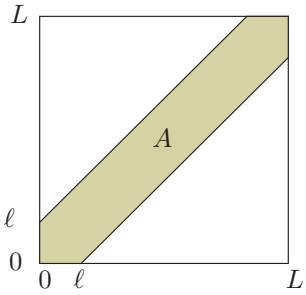
18.3 Geometric probability

The following is an example of what is called **geometric probability**. Outcomes can be visualized and their probabilities found with the aid of a picture.

Example 18.3. Two points X and Y are chosen at random and independently along a line segment of length L. What is the probability that:

$$|X - Y| \leq \ell\,?$$

Suppose that "at random" means uniformly so that $f(x,y) = \dfrac{1}{L^2}, \quad x, y \in [0, L]^2$.



The desired probability is

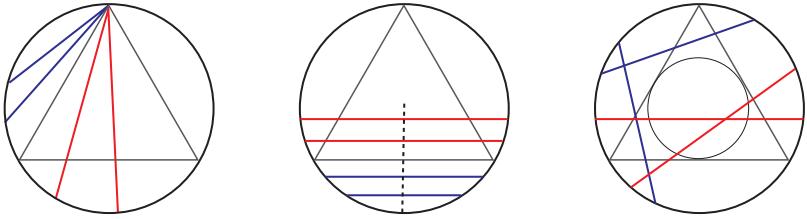
$$= \iint_A f(x,y)\,dx\,dy = \frac{\text{area of A}}{L^2} = \frac{L^2-(L-\ell)^2}{L^2} = \frac{2L\ell-\ell^2}{L^2}.$$

We consider below two further examples of geometric probability. Others are 'A stick in broken at random in two places. What is the probability that the three pieces can form a triangle?'. See also the infamous Planet Zog tripos questions in 2003, 2004.

18.4 Bertrand's paradox

Example 18.4 [*Bertrand's Paradox*]. Posed by Bertrand in 1889: What is the probability that a "random chord" of a circle has length greater than the length of the side of an inscribed equalateral triangle?

The 'paradox' is that there are at least 3 interpretations of what it means for a chord to be chosen 'at random'.



(1) *The endpoints are independently and uniformly distributed over the circumference.* The chord is longer than a side of the triangle if the other chord endpoint lies on the arc between the endpoints of the triangle side opposite the first point. The length of the arc is one third of the circumference of the circle. So answer $= \frac{1}{3}$.

(2) *The chord is perpendicular to a given radius, intersecting at a point chosen uniformly over the radius.* The chord is longer than a side of the triangle if the chosen point is

nearer the center of the circle than the point where the side of the triangle intersects the radius. Since the side of the triangle bisects the radius: answer $= \frac{1}{2}$.

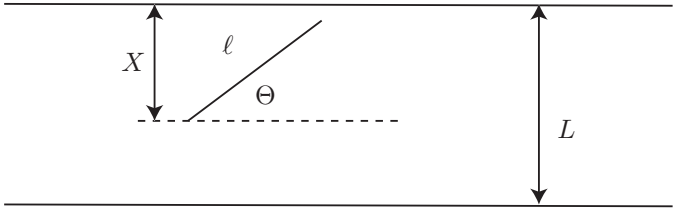
(3) *The midpoint of the chord is chosen uniformly within the circle.* The chord is longer than a side of the inscribed triangle if the chosen point falls within a concentric circle of radius 1/2 the radius of the larger circle. As the smaller circle has 1/4 the area of the large circle: answer $= \frac{1}{4}$.

Suppose we throw long sticks from a large distance onto a circle drawn on the ground. To which of (1)–(3) does this correspond? What is the 'best' answer to Bertrand's question?

18.5 Buffon's needle

Example 18.5. A needle of length ℓ is tossed at random onto a floor marked with parallel lines a distance L apart, where $\ell \leq L$. What is the probability of the event $A =$[the needle intersects one of the parallel lines]?

Answer. Let $\Theta \in [0, \pi]$ be the angle between the needle and the parallel lines and let X be the distance from the bottom of the needle to the line above it.



It is reasonable to suppose that independently

$$X \sim U[0, L], \quad \Theta \sim U[0, \pi).$$

Thus

$$f(x, \theta) = \frac{1}{L}\frac{1}{\pi}, \quad 0 \leq x \leq L \text{ and } 0 \leq \theta \leq \pi.$$

The needle intersects the line if and only if $X \leq \ell \sin \Theta$. So

$$p = P(A) = \int_0^\pi \frac{\ell \sin\theta}{L}\frac{1}{\pi} d\theta = \frac{2\ell}{\pi L}. \qquad \square$$

Suppose we drop a needle n times and it hits the line N times. We might estimate p by $\hat{p} = N/n$, and π by $\hat{\pi} = (2\ell)/(\hat{p}L)$.

In §23.3 we consider how large n must be so that the estimate of π is good, in the sense $P(|\hat{\pi} - \pi| < 0.001) \geq 0.95$.

19 Normal distribution

Normal distribution, Mean, mode and median. Order statistics and their distributions. *Stochastic bin packing*.

19.1 Normal distribution

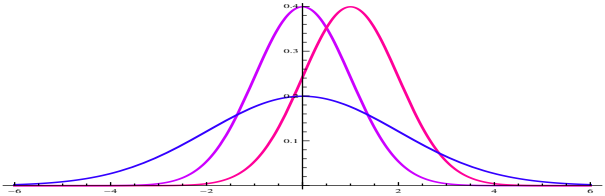
The **normal distribution** (or Gaussian distribution) with parameters μ and σ^2 has p.d.f.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty.$$

To indicate that X has this distribution we write $X \sim N(\mu, \sigma^2)$.

The **standard normal distribution** means $N(0,1)$ and its c.d.f. is usually denoted by $\Phi(x) = \int_{-\infty}^{x} (1/\sqrt{2\pi}) e^{-x^2/2}\, dx$, and $\bar{\Phi}(x) = 1 - \Phi(x)$.

Bell-shaped p.d.fs of $N(0,1)$, $N(1,1)$ and $N(0,2)$:



Example 19.1. We need to check that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

We make the substitution $z = (x-\mu)/\sigma$. So $dx/\sigma = dz$, and then

$$I = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, dz.$$

You are probably familiar with how to calculate this. Look at

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \left[\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx \right] \left[\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\, dy \right] \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy \\ &= \frac{1}{2\pi} \int_{0}^{2\pi} \int_{0}^{\infty} r e^{-\frac{1}{2}r^2}\, dr\, d\theta \\ &= \frac{1}{2\pi} \int_{0}^{2\pi} d\theta = 1. \end{aligned}$$

Therefore $I = 1$.

The expectation is

$$E[X] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$
$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx + \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\mu e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx.$$

The first term is convergent and equals zero by symmetry, so that

$$E[X] = 0 + \mu = \mu.$$

Now let $Z = (X-\mu)/\sigma$. So $\mathrm{Var}(X) = \sigma^2\,\mathrm{Var}(Z)$. Using Theorem 17.1 we see that the density of Z is $(2\pi)^{-1/2}\exp(-z^2/2)$, and so $Z \sim N(0,1)$.

We have $\mathrm{Var}(Z) = E\left[Z^2\right] - E[Z]^2$.

Now $E[Z] = 0$. So using integration by parts to find $E\left[Z^2\right]$,

$$\mathrm{Var}(Z) = \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2}dz = \left[-\tfrac{1}{\sqrt{2\pi}}ze^{-\frac{1}{2}z^2}\right]_{-\infty}^{\infty} + \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}dz = 0 + 1 = 1.$$

Hence $\mathrm{Var}\,X = \sigma^2$.

19.2 Calculations with the normal distribution

Example 19.2 [*The advantage of a small increase in the mean*]. UK adult male heights are normally distributed with mean $70''$ and standard deviation $3''$. In the Netherlands these figures are $71''$ and $3''$. What is $P(Y > X)$, where X and Y are the heights of randomly chosen UK and Netherlands males, respectively?

Answer. Sums of normal r.vs are normally distributed (proved in §22.2). So with $X \sim N(70, 9)$ and $Y \sim N(71, 9)$ we have $Y - X \sim N(1, 18)$.

So $P(Y - X > 0) = \Phi(1/\sqrt{18}) = 0.5931$. This is more than $1/2$ but not hugely so. □

Now suppose that in both countries the Olympic male basketball teams are selected from that portion of men whose height is at least $4''$ above the mean (which corresponds to the 9.1% tallest males of the country). What is the probability that a randomly chosen Netherlands player is taller than a randomly chosen UK player?

Answer. Now we want $P(X < Y \mid X \geq 74, Y \geq 75)$. Let ϕ_X and ϕ_Y be the p.d.fs of $N(70, 9)$ and $N(71, 9)$ respectively. The answer is

$$\frac{\int_{x=74}^{75}\phi_X(x)\,dx\int_{y=75}^{\infty}\phi_Y(y)\,dy + \int_{x=75}^{\infty}\left(\int_{y=x}^{\infty}\phi_Y(y)\,dy\right)\phi_X(x)\,dx}{\int_{x=74}^{\infty}\phi_X(x)\,dx\int_{y=75}^{\infty}\phi_Y(y)\,dy}$$

$$= 0.7558 \quad \text{(computed numerically)}.$$ □

The lesson is that if members of a population A are only slightly better in some activity than members of a population B, then members of A may nonetheless appear much more talented than members of B when one focuses upon the sub-populations of exceptional performers (such as 100m sprinters or Nobel prize winners).

19.3 Mode, median and sample mean

Given a p.d.f. $f(x)$, we say $\hat{x}$ is a **mode** if $f(\hat{x}) \geq f(x)$ for all x, and $\hat{x}$ is a **median** if

$$\int_{-\infty}^{\hat{x}} f(x)\,dx = \int_{\hat{x}}^{\infty} f(x)\,dx = \frac{1}{2}.$$

For a discrete random variable, $\hat{x}$ is a median if

$$P(X \leq \hat{x}) \geq \frac{1}{2} \text{ and } P(X \geq \hat{x}) \geq \frac{1}{2}.$$

For $N(\mu, \sigma^2)$ the mean, mode and median are all equal to μ.

If $X_1, \ldots, X_n$ is a **random sample** from the distribution then the **sample mean** is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

19.4 Distribution of order statistics

Let $Y_1, \ldots, Y_n$ be the values of $X_1, \ldots, X_n$ arranged in increasing order, so $Y_1 \leq \cdots \leq Y_n$. These are called the **order statistics**. Another common notation is $X_{(i)} = Y_i$, so that $X_{(1)} \leq \cdots \leq X_{(n)}$.

The **sample median** is $Y_{\frac{n+1}{2}}$ if n is odd or any value in $\left[Y_{\frac{n}{2}}, Y_{\frac{n}{2}+1}\right]$ if n is even.

The largest is $Y_n = \max\{X_1, \ldots, X_n\}$. If $X_1, \ldots, X_n$ are i.i.d. r.vs with c.d.f. F and p.d.f. f then,

$$P(Y_n \leq y) = P(X_1 \leq y, \ldots, X_n \leq y) = F(y)^n.$$

Thus the p.d.f. of Y_n is

$$g(y) = \frac{d}{dy}F(y)^n = nF(y)^{n-1}f(y).$$

Similarly, the smallest is $Y_1 = \min\{X_1, \ldots, X_n\}$, and

$$P(Y_1 \leq y) = 1 - P(X_1 \geq y, \ldots, X_n \geq y) = 1 - (1 - F(y))^n.$$

Thus the p.d.f. of Y_1 is

$$h(y) = n\left(1 - F(y)\right)^{n-1} f(y).$$

What about the joint density of Y_1, Y_n? The joint c.d.f. is

$$\begin{aligned} G(y_1, y_n) &= P\left(Y_1 \le y_1, Y_n \le y_n\right) \\ &= P\left(Y_n \le y_n\right) - P\left(Y_n \le y_n, Y_1 > y_1\right) \\ &= P\left(Y_n \le y_n\right) - P\left(y_1 < X_1 \le y_n, y_1 < X_2 \le y_n, \ldots, y_1 < X_n \le y_n\right) \\ &= F(y_n)^n - \left(F(y_n) - F(y_1)\right)^n. \end{aligned}$$

Thus the joint p.d.f. of Y_1, Y_n is

$$\begin{aligned} g(y_1, y_n) &= \frac{\partial^2}{\partial y_1 \partial y_n} G(y_1, y_n) \\ &= \begin{cases} n(n-1)\left(F(y_n) - F(y_1)\right)^{n-2} f(y_1) f(y_n), & -\infty < y_1 \le y_n < \infty \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

To see this another way, think of 5 boxes corresponding to intervals $(-\infty, y_1)$, $[y_1, y_1+\delta)$, $[y_1+\delta, y_n)$, $[y_n, y_n+\delta)$, $[y_n+\delta, \infty)$. We find the probability that the counts in the boxes are $0, 1, n-2, 1, 0$ by using the multinomial distribution, and the idea that a point is chosen in box $[y_1, y_1+\delta)$ with probability $f(y_1)\delta$, and so on. See also Example 21.1.

19.5 Stochastic bin packing

The weights of n items are $X_1, \ldots, X_n$, assumed i.i.d. $U[0,1]$. Mary and John each have a bin (or suitcase) which can carry total weight 1. Mary likes to pack in her bin only the heaviest item. John likes to pack the items in order $1, 2, \ldots, n$, packing each item if it can fit in the space remaining. Whose suitcase is more likely to heavier?

Answer. Let Z_M and Z_J be the unused capacity in Mary's and John's bins, respectively.

$$P(Z_M > t) = P(X_1 \le 1-t, \ldots, X_n \le 1-t) = (1-t)^n.$$

Calculation for Z_J is trickier. Let $G_k(x)$ be the probability that $Z_J > t$ given that when John is about to consider the last k items the remaining capacity of his bin is x, where $x > t$. Clearly, $G_0(x) = 1$. We shall prove inductively that $G_k(x) = (1-t)^k$. Assuming this is true at k, and letting $X = X_{n-k+1}$, an inductive step follows from

$$G_{k+1}(x) = P(X > x) G_k(x) + \int_0^{x-t} G_k(x-y)\, dy = (1-t)^{k+1}.$$

Thus, $P(Z_J > t) = G_n(1) = (1-t)^n = P(Z_M > t)$. So, surprisingly, $Z_J =_{\text{st}} Z_M$. □

20 Transformations of random variables

Transformation of random variables. Convolution. Cauchy distribution.

20.1 Transformation of random variables

Suppose $X_1, X_2, \ldots, X_n$ have joint p.d.f. f. Let

$$\begin{aligned} Y_1 &= r_1(X_1, X_2, \ldots, X_n) \\ Y_2 &= r_2(X_1, X_2, \ldots, X_n) \\ &\vdots \\ Y_n &= r_n(X_1, X_2, \ldots, X_n). \end{aligned}$$

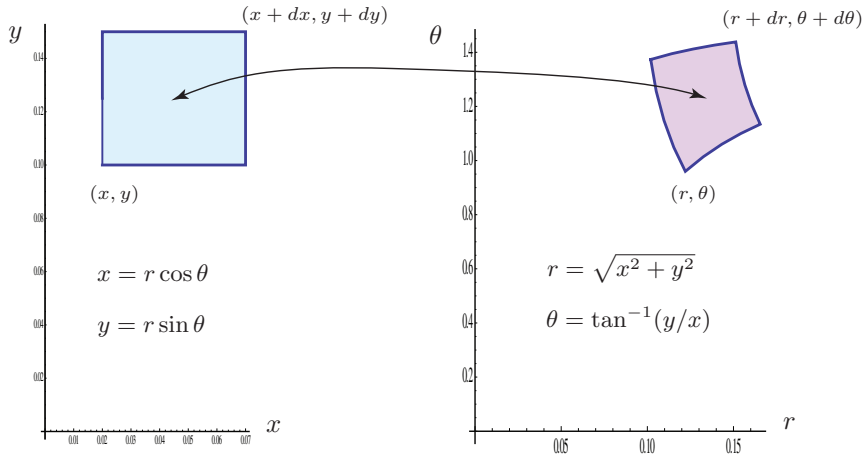
Let $R \subseteq \mathbb{R}^n$ be such that

$$P\left((X_1, X_2, \ldots, X_n) \in R\right) = 1.$$

Let S be the image of R under the above transformation. Suppose the transformation from R to S is 1–1 (bijective). Then there exist inverse functions

$$\begin{aligned} X_1 &= s_1(Y_1, Y_2, \ldots, Y_n) \\ X_2 &= s_2(Y_1, Y_2, \ldots, Y_n) \\ &\vdots \\ X_n &= s_n(Y_1, Y_2, \ldots, Y_n). \end{aligned}$$

The familiar bijection between rectangular and polar coordinates is shown below.



Assume that $\partial s_i/\partial y_j$ exists and is continuous at every point $(y_1, y_2, \ldots, y_n)$ in S. Define the **Jacobian** determinant as

$$J = \frac{\partial(s_1, \ldots, s_n)}{\partial(y_1, \ldots, y_n)} = \det \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}.$$

If $A \subseteq R$ and B is the image of A then

$$\begin{aligned} P\left((X_1, \ldots, X_n) \in A\right) &= \int_A \cdots \int f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \qquad (1) \\ &= \int_B \cdots \int f\left(s_1, \ldots, s_n\right) |J|\ dy_1 \ldots dy_n \\ &= P\left((Y_1, \ldots, Y_n) \in B\right). \qquad (2) \end{aligned}$$

Transformation is 1–1 so (1), (2) are the same. Thus the density for $Y_1, \ldots, Y_n$ is

$$g(y_1, y_2, \ldots, y_n) = \begin{cases} f\left(s_1(y_1, y_2, \ldots, y_n), \ldots, s_n(y_1, y_2, \ldots, y_n)\right) |J|, \\ \qquad \text{if } (y_1, y_2, \ldots, y_n) \in S \\ 0, \quad \text{otherwise.} \end{cases}$$

See also Appendix C.

Example 20.1 [*density of products and quotients*]. Suppose that (X, Y) has density

$$f(x, y) = \begin{cases} 4xy, & \text{for } 0 \le x \le 1,\ 0 < y \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$U = X/Y, \quad V = XY, \quad \text{so } X = \sqrt{UV}. \quad Y = \sqrt{V/U},$$

$$\det \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix} = \det \begin{pmatrix} \frac{1}{2}\sqrt{v/u} & \frac{1}{2}\sqrt{u/v} \\ -\frac{1}{2}\sqrt{v/u^3} & \frac{1}{2}\sqrt{1/(uv)} \end{pmatrix} = \frac{1}{2u}.$$

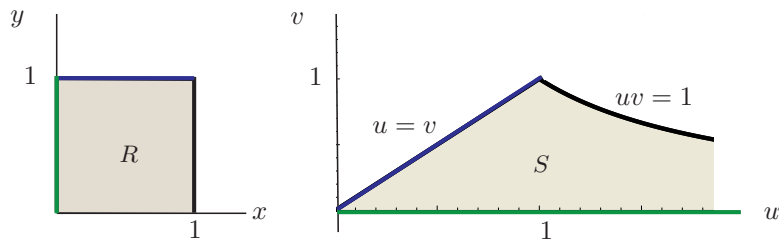
Therefore $|J| = \frac{1}{2u}$. It can sometimes be easier to work the other way and then invert:

$$\det \begin{pmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial v/\partial y \end{pmatrix} = \det \begin{pmatrix} 1/y & -x/y^2 \\ y & x \end{pmatrix} = 2x/y = 2u.$$

Therefore $|J| = \frac{1}{2u}$. So taking S as the regious shown below, we have for $(u, v) \in S$,

$$g(u, v) = \frac{1}{2u}(4xy) = \frac{1}{2u} \times 4\sqrt{uv}\sqrt{\frac{v}{u}} = 2v/u,$$

and $g(u, v) = 0$ otherwise.



Notice that U and V are not independent since

$$g(u,v) = 2(v/u)\, I[(u,v) \in S]$$

is not the product of the two densities.

When the transformations are linear things are simple. Let A be a $n \times n$ invertible matrix. Then

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},$$

$$|J| = \det(A^{-1}) = (\det A)^{-1}.$$

Thus the p.d.f. of $(Y_1, \ldots, Y_n)$ is

$$g(y_1, \ldots, y_n) = \frac{1}{\det A} f(A^{-1}y).$$

20.2 Convolution

Example 20.2. Suppose X_1, X_2 have the p.d.f. $f(x_1, x_2)$ and we wish to calculate the p.d.f. of $X_1 + X_2$.

Let $Y = X_1 + X_2$ and $Z = X_2$. Then $X_1 = Y - Z$ and $X_2 = Z$.

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \text{so } |J| = 1/|\det(A)| = 1.$$

The joint distribution of Y and Z is

$$g(y,z) = f(x_1, x_2) = f(y - z, z).$$

The marginal density of Y is

$$\begin{aligned} g(y) &= \int_{-\infty}^{\infty} f(y - z, z)\, dz, \qquad -\infty < y < \infty, \\ &= \int_{-\infty}^{\infty} f(z, y - z)\, dz \quad \text{(by change of variable)}. \end{aligned}$$

If X_1 and X_2 are independent, with p.d.fs f_1 and f_2 then

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \implies g(y) = \int_{-\infty}^{\infty} f_1(z)f_2(y-z)\,dz.$$

This is called the **convolution** of f_1 and f_2.

20.3 Cauchy distribution

The **Cauchy distribution** has p.d.f.

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

By making the substitution $x = \tan\theta$, we can check that this is a density, since

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)}\,dx = \int_{-\pi/2}^{\pi/2} \frac{1}{\pi}\,d\theta = 1.$$

The Cauchy distribution is an example of a distribution having no mean, since

$$\int_{-\infty}^{\infty} x\frac{dx}{\pi(1+x^2)} = \int_{0}^{\infty} x\frac{dx}{\pi(1+x^2)} + \int_{-\infty}^{0} x\frac{dx}{\pi(1+x^2)} = \infty - \infty.$$

$E[X]$ does not exist because both integrals are infinite.

However, the second moment does exist. It is $E[X^2] = \infty$.

Suppose X and Y are independent and have the Cauchy distribution. To find the distribution of $Z = X + Y$ we use convolution.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\pi^2(1+x^2)(1+(z-x)^2)}\,dx$$

$$= \frac{1/2}{\pi(1+(z/2)^2)} \quad \text{(by using partial fractions)}.$$

We conclude that $\frac{1}{2}Z$ also has the Cauchy distribution.

Inductively, one can show that if $X_1, \ldots, X_n$ are i.i.d. with the Cauchy distribution, then $\frac{1}{n}(X_1 + \cdots + X_n)$ also has the Cauchy distribution. This demonstrates that the central limit theorem does not hold when X_i has no mean.

Facts. (i) If $\Theta \sim U[-\pi/2, \pi/2]$ then $X = \tan\Theta$ has the Cauchy distribution.

(ii) If independently $X, Y \sim N(0,1)$ then X/Y has the Cauchy distribution.

21 Moment generating functions

Transformations that are not 1–1. Minimum of exponentials is exponential. Moment generating functions. Moment generating function of exponential distribution. Sum of i.i.d. exponential random variables and the gamma distribution. *Beta distribution*.

21.1 What happens if the mapping is not 1–1?

What happens if the mapping is not 1–1? Suppose X has p.d.f. f. What is the p.d.f. of $Y = |X|$? Clearly $Y \geq 0$, and so for $0 \leq a < b$,

$$P(|X| \in (a,b)) = \int_a^b (f(x) + f(-x))\,dx \quad \implies f_Y(x) = f(x) + f(-x).$$

Example 21.1. Suppose $X_1, \ldots, X_n$ are i.i.d. r.vs. What is the p.d.f. of the order statistics $Y_1, \ldots, Y_n$?

$$g(y_1, \ldots, y_n) = \begin{cases} n! f(y_1) \cdots f(y_n), & y_1 \leq y_2 \leq \cdots \leq y_n \\ 0, & \text{otherwise.} \end{cases}$$

The factor of $n!$ appears because this is the number of $x_1, \ldots, x_n$ that could give rise to a specific $y_1, \ldots, y_n$.

21.2 Minimum of exponentials is exponential

Example 21.2. Suppose X_1 and X_2 are independent exponentially distributed r.vs, with parameters λ and μ. Let $Y_1 = \min(X_1, X_2)$. Then

$$P(Y_1 \geq t) = P(X_1 \geq t)P(X_2 \geq t) = e^{-\lambda t} e^{-\mu t} = e^{-(\lambda+\mu)t}$$

and so Y is exponentially distributed with parameter $\lambda + \mu$.

Example 21.3. Suppose $X_1, \ldots, X_n$ are i.i.d. r.vs exponentially distributed with parameter λ. Let $Y_1, \ldots, Y_n$ be the order statistics of $X_1, \ldots, X_n$, and

$$\begin{aligned} Z_1 &= Y_1 \\ Z_2 &= Y_2 - Y_1 \\ &\vdots \\ Z_n &= Y_n - Y_{n-1} \end{aligned}$$

To find the distribution of the Z_i we start by writing $Z = AY$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & -1 & 1 \end{pmatrix}.$$

Since $\det(A) = 1$ we have

$$\begin{aligned} h(z_1, \dots, z_n) &= g(y_1, \dots, y_n) \\ &= n! f(y_1) \cdots f(y_n) \\ &= n! \lambda^n e^{-\lambda y_1} \dots e^{-\lambda y_n} \\ &= n! \lambda^n e^{-\lambda(y_1 + \dots + y_n)} \\ &= n! \lambda^n e^{-\lambda(n z_1 + (n-1) z_2 + \dots + z_n)} \\ &= \textstyle\prod_{i=1}^n (\lambda i) e^{-(\lambda i) z_{n+1-i}}. \end{aligned}$$

Thus $h(z_1, \dots, z_n)$ is expressed as the product of n density functions and

$$Z_i \sim \text{exponential}((n + 1 - i)\lambda),$$

with $Z_1, \dots, Z_n$ being independent.

This should also be intuitively obvious. Think of $0, Y_1, Y_2, \dots, Y_n$ as increasing times, separated by $Z_1, Z_2, \dots, Z_n$. Clearly, $Z_1 = Y_1 \sim \mathscr{E}(n\lambda)$ (since Y_1 is the minimum of n i.i.d. exponential r.vs). Then, by the memoryless property of exponential r.vs, things continue after Y_1 in the same way, but with only $n-1$ i.i.d. exponential r.vs remaining.

21.3 Moment generating functions

The **moment generating function** (m.g.f.) of a random variable X is defined by

$$m(\theta) = E\left[e^{\theta X}\right]$$

for those θ such that $m(\theta)$ is finite. It is computed as

$$m(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx$$

where $f(x)$ is the p.d.f. of X.

The m.g.f. is defined for any type of r.v. but is most commonly useful for continuous r.vs, whereas the p.g.f. is most commonly used for discrete r.vs.

We will use the following theorem without proof.

Theorem 21.4. *The moment generating function determines the distribution of X, provided $m(\theta)$ is finite for all θ in some interval containing the origin.*

$E[X^r]$ is called the "r^{th} **moment** of X".

Theorem 21.5. *The r^{th} moment of X is the coefficient of $\theta^r/r!$ in the power series expansion of $m(\theta)$, equivalently, the rth derivative evaluated at $\theta = 0$, i.e. $m^{(r)}(0)$.*

Sketch of proof. $e^{\theta X} = 1 + \theta X + \frac{1}{2!}\theta^2 X^2 + \cdots$. So

$$E\left[e^{\theta X}\right] = 1 + \theta E[X] + \tfrac{1}{2!}\theta^2 E\left[X^2\right] + \cdots . \qquad \square$$

Example 21.6. Let X be exponentially distributed with parameter λ. Its m.g.f. is

$$E\left[e^{\theta X}\right] = \int_0^\infty e^{\theta x}\lambda e^{-\lambda x}dx = \lambda\int_0^\infty e^{-(\lambda-\theta)x}dx = \frac{\lambda}{\lambda-\theta}, \quad \text{for } \theta \le \lambda.$$

The first two moments are

$$E[X] = m'(0) = \left.\frac{\lambda}{(\lambda-\theta)^2}\right|_{\theta=0} = \frac{1}{\lambda}$$

$$E\left[X^2\right] = m''(0) = \left.\frac{2\lambda}{(\lambda-\theta)^3}\right|_{\theta=0} = \frac{2}{\lambda^2}.$$

Thus

$$\operatorname{Var} X = E\left[X^2\right] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Theorem 21.7. *If X and Y are independent random variables with moment generating functions $m_X(\theta)$ and $m_Y(\theta)$ then $X + Y$ has the moment generating function*

$$m_{X+Y}(\theta) = m_X(\theta)\cdot m_Y(\theta).$$

Proof. $E\left[e^{\theta(X+Y)}\right] = E\left[e^{\theta X}e^{\theta Y}\right] = E\left[e^{\theta X}\right]E\left[e^{\theta Y}\right] = m_X(\theta)m_Y(\theta)$. $\square$

21.4 Gamma distribution

Example 21.8 [*gamma distribution*]. Suppose $X_1, \ldots, X_n$ are i.i.d. r.vs each exponentially distributed with parameter λ.

Let $S_n = X_1 + \cdots + X_n$. The m.g.f. of S_n is

$$E\left[e^{\theta(X_1+\cdots+X_n)}\right] = E\left[e^{\theta X_1}\right]\cdots E\left[e^{\theta X_n}\right] = \left(E\left[e^{\theta X_1}\right]\right)^n = \left(\frac{\lambda}{\lambda-\theta}\right)^n.$$

The **gamma distribution**, denoted $\Gamma(n, \lambda)$, with parameters $n \in \mathbb{Z}^+$ and $\lambda > 0$, has density

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}, \quad 0 \le x < \infty.$$

We can check that this is a density, by using integration by parts to show that $\int_0^\infty f(x)\,dx = 1$. Suppose that $Y \sim \Gamma(n, \lambda)$. Its m.g.f. is

$$\begin{aligned} E\left[e^{\theta Y}\right] &= \int_0^\infty e^{\theta x} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}\,dx \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n \int_0^\infty \frac{(\lambda-\theta)^n x^{n-1} e^{-(\lambda-\theta)x}}{(n-1)!}\,dx \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n, \end{aligned}$$

since the final integral evaluates to 1 (the integrand being the p.d.f. of $\Gamma(n, \lambda-\theta)$).

We can conclude that $S_n \sim \Gamma(n, \lambda)$, since the moment generating function characterizes the distribution.

The gamma distribution $\Gamma(\alpha, \lambda)$ is also defined for any $\alpha, \lambda > 0$. The denominator of $(n-1)!$ in the p.d.f. is replaced with the gamma function, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,dx$.

The case in which α is a positive integer is also called the Erlang distribution.

21.5 Beta distribution

Suppose that $X_1, \ldots, X_n$ are i.i.d. $U[0,1]$. Let $Y_1 \le Y_2 \le \cdots \le Y_n$ be the order statistics. The p.d.f. of Y_i is

$$f(y) = \frac{n!}{(i-1)!(n-i)!} y^{i-1}(1-y)^{n-i}, \quad 0 \le y \le 1.$$

Do you see why? Notice that the leading factor is the multinomial coefficient $\binom{n}{i-1,1,n-i}$.

This is the **beta distribution**, denoted $Y_i \sim \text{Beta}(i, n-i+1)$, with mean $i/(n+1)$.

More generally, for $a, b > 0$, $\text{Beta}(a, b)$ has p.d.f.

$$f(x : a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 \le x \le 1.$$

The beta distribution is used by actuaries to model the loss of an insurance risk. There is some further discussion in Appendix D.

22 Multivariate normal distribution

Moment generating function of a normal random variable. Sums and linear transformations of a normal random variable. Bounds on tail probability of a normal distribution. Multivariate and bivariate normal distributions. Multivariate moment generating functions.

22.1 Moment generating function of normal distribution

Example 22.1. The moment generating function of a normally distributed random variable, $X \sim N(\mu, \sigma^2)$, is found as follows.

$$E\left[e^{\theta X}\right] = \int_{-\infty}^{\infty} e^{\theta x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx.$$

Substitute $z = (x-\mu)/\sigma$ to obtain

$$\begin{aligned} E\left[e^{\theta X}\right] &= \int_{-\infty}^{\infty} e^{\theta(\mu+\sigma z)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\ &= e^{\theta\mu+\frac{1}{2}\theta^2\sigma^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\theta\sigma)^2} dz \\ &= e^{\mu\theta+\frac{1}{2}\theta^2\sigma^2}. \end{aligned}$$

The final integral equals 1, as the integrand is the density of $N(\theta\sigma, 1)$.

22.2 Functions of normal random variables

Theorem 22.2. *Suppose X, Y are independent r.vs, with $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. Then*

1. $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$,
2. $aX \sim N(a\mu_1, a^2\sigma_1^2)$.

Proof. 1.

$$\begin{aligned} E\left[e^{\theta(X+Y)}\right] &= E\left[e^{\theta X}\right] E\left[e^{\theta Y}\right] = e^{(\mu_1\theta+\frac{1}{2}\sigma_1^2\theta^2)} e^{(\mu_2\theta+\frac{1}{2}\sigma_2^2\theta^2)} \\ &= \exp\left[(\mu_1+\mu_2)\theta + \tfrac{1}{2}(\sigma_1^2+\sigma_2^2)\theta^2\right] \end{aligned}$$

which is the moment generating function for $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2.

$$E\left[e^{\theta(aX)}\right] = E\left[e^{(\theta a)X}\right] = e^{\mu_1(\theta a)+\frac{1}{2}\sigma_1^2(\theta a)^2}$$
$$= \exp\left[(a\mu_1)\theta + \tfrac{1}{2}a^2\sigma_1^2\theta^2\right],$$

which is the moment generating function of $N(a\mu_1, a^2\sigma_1^2)$. □

22.3 Bounds on tail probability of a normal distribution

The tail probabilities of a normal distribution are often important quantities to evaluate or bound. We can bound the probability in the tail as follows.

Suppose $X \sim N(0,1)$, with density function $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$. Then for $x > 0$,

$$P(X > x) = 1 - \Phi(x) < \int_x^\infty \left(1 + \frac{1}{t^2}\right)\phi(t)\,dt = \frac{1}{x}\phi(x).$$

For example, $1 - \Phi(3) = 0.00135$ and the bound above is 0.00148.

By similar means one can show a lower bound, and hence $\log(1 - \Phi(x)) \sim -\frac{1}{2}x^2$.

22.4 Multivariate normal distribution

Let $X_1, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables with joint density $g(x_1, \ldots, x_n)$.

$$g(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}$$
$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}x^T x}.$$

Here $x^T = (x_1, \ldots, x_n)^T$ is a row vector. Write

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

and consider the vector r.v. $Z = \mu + AX$, where A is an invertible $n \times n$ matrix, so $X = A^{-1}(Z - \mu)$. Here Z and μ are column vectors of n components.

The density of Z is

$$\begin{aligned} f(z_1,\ldots,z_n) &= \frac{1}{(2\pi)^{n/2}}\frac{1}{\det A}e^{-\frac{1}{2}\left(A^{-1}(z-\mu)\right)^T\left(A^{-1}(z-\mu)\right)} \\ &= \frac{1}{(2\pi)^{n/2}\left|\Sigma\right|^{1/2}}e^{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)} \end{aligned}$$

where $\Sigma = AA^T$. This is the **multivariate normal density** (MVN), written as

$$Z \sim MVN(\mu,\Sigma) \quad \text{(or just } N(\mu,\Sigma)\text{)}.$$

The vector μ is EZ. The covariance,

$$\mathrm{Cov}(Z_i,Z_j) = E\left[(Z_i-\mu_i)(Z_j-\mu_j)\right]$$

is the (i,j) entry of

$$\begin{aligned} E\left[(Z-\mu)(Z-\mu)^T\right] &= E\left[(AX)(AX)^T\right] \\ &= AE\left[XX^T\right]A^T \\ &= AIA^T = AA^T = \Sigma \quad \text{(the covariance matrix)}. \end{aligned}$$

If the covariance matrix of the MVN distribution is diagonal, then the components of the random vector Z are independent since

$$f(z_1,\ldots,z_n) = \prod_{i=1}^{n}\frac{1}{(2\pi)^{\frac{1}{2}}\sigma_i}e^{-\frac{1}{2}\left(\frac{z_i-\mu_i}{\sigma_i}\right)^2}, \quad \text{where } \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}.$$

This is a special property of the MVN. We already know that if the joint distribution of r.vs is not MVN then covariances of 0 do not, in general, imply independence.

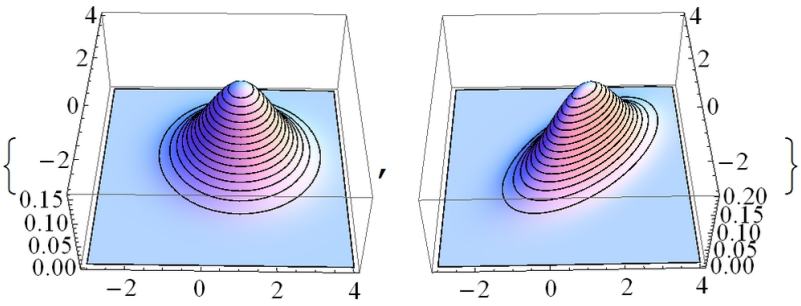
22.5 Bivariate normal

The **bivariate normal random variable** is the multivariate normal with $n = 2$, having covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

$E[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. $\mathrm{Cov}(X_1,X_2) = \sigma_1\sigma_2\rho$.

$$\mathrm{Corr}(X_1,X_2) = \frac{\mathrm{Cov}(X_1,X_2)}{\sigma_1\sigma_2} = \rho, \qquad \Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} \\ -\rho\sigma_1^{-1}\sigma_2^{-1} & \sigma_2^{-2} \end{pmatrix}.$$



Plots of the joint p.d.fs of $MVN\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$, for $\rho = 0$ and $\rho = 0.6$.

The joint distribution is written as

$$f(x_1, x_2) = \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}\sigma_1\sigma_2} \times$$
$$\exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right],$$

where $\sigma_1, \sigma_2 > 0$ and $-1 \leq p \leq 1$.

22.6 Multivariate moment generating function

For random variables $X_1, \ldots, X_n$ and real numbers $\theta_1, \ldots, \theta_n$ we define

$$m(\theta) = m(\theta_1, \ldots, \theta_n) = E\left[e^{(\theta_1 X_1 + \cdots + \theta_n X_n)}\right]$$

to be the **joint moment generating function** of the random variables. It is only defined for those θ for which $m(\theta)$ is finite. Its properties are similar to those of the moment generating function of a single random variable.

The joint moment generating function of the bivariate normal is

$$m(\theta_1, \theta_2) = \exp\Big(\theta_1\mu_1 + \theta_2\mu_2 + \tfrac{1}{2}(\theta_1^2\sigma_1^2 + 2\theta_1\theta_2\rho\sigma_1\sigma_2 + \theta_2^2\sigma_2^2)\Big).$$

23 Central limit theorem

Central limit theorem. Sketch of proof. Normal approximation to the binomial. Opinion polls. Buffon's needle.

23.1 Central limit theorem

Suppose $X_1, \ldots, X_n$ are i.i.d. r.vs, mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$. We know that

$$\mathrm{Var}\,(S_n/\sqrt{n}) = \mathrm{Var}\left(\frac{S_n - n\mu}{\sqrt{n}}\right) = \sigma^2.$$

Theorem 23.1. *Let $X_1, \ldots, X_n$ be i.i.d. r.vs with $E[X_i] = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 < \infty$. Define $S_n = X_1 + \cdots + X_n$. Then for all (a, b) such that $-\infty < a \leq b < \infty$*

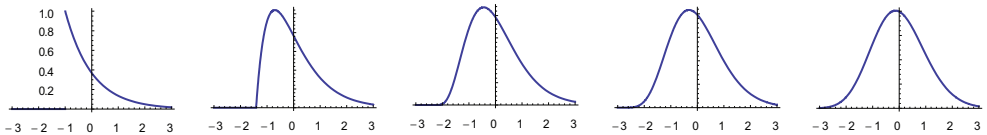
$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$

which is the p.d.f. of a $N[0,1]$ random variable.

We write

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to^{\mathscr{D}} N(0,1).$$

which is read as '*tends in distribution to*'.



Probability density functions of $\frac{S_n - n}{\sqrt{n}}$, $n = 1, 2, 5, 10, 20$,
when $X_1, X_2, \ldots, X_n$ are i.i.d. exponentially distributed with $\mu = \sigma^2 = 1$.

The proof uses the so-called **continuity theorem**, which we quote without proof.

Theorem 23.2 (continuity theorem)**.** *If random variables $X_1, X_2, \ldots$ have moment generating functions $m_i(\theta)$, $i = 1, 2, \ldots$, and $m_i(\theta) \to m(\theta)$ as $i \to \infty$, pointwise for every θ, then X_i tends in distribution to the random variable having m.g.f. $m(\theta)$.*

Sketch proof of Central Limit Theorem. Without loss of generality, take $\mu = 0$ and $\sigma^2 = 1$ (since we can replace X_i by $(X_i - \mu)/\sigma$. The m.g.f. of X_i is

$$\begin{aligned} m_{X_i}(\theta) &= E\left[e^{\theta X_i}\right] \\ &= 1 + \theta E[X_i] + \tfrac{1}{2!}\theta^2 E\left[X_i^2\right] + \tfrac{1}{3!}\theta^3 E\left[X_i^3\right] + \cdots \\ &= 1 + \tfrac{1}{2!}\theta^2 + \tfrac{1}{3!}\theta^3 E\left[X_i^3\right] + \cdots \end{aligned}$$

The m.g.f. of $S_n/\sqrt{n}$ is

$$\begin{aligned} E\left[e^{\theta\frac{S_n}{\sqrt{n}}}\right] &= E\left[e^{\frac{\theta}{\sqrt{n}}(X_1+\cdots+X_n)}\right] \\ &= E\left[e^{\frac{\theta}{\sqrt{n}}X_1}\right]\cdots E\left[e^{\frac{\theta}{\sqrt{n}}X_n}\right] \\ &= E\left[e^{\frac{\theta}{\sqrt{n}}X_1}\right]^n \\ &= \left(m_{X_1}(\theta/\sqrt{n})\right)^n \\ &= \left(1+\tfrac{1}{2}\theta^2\frac{1}{n}+\tfrac{1}{3!}\theta^3 E\left[X^3\right]\frac{1}{n^{3/2}}+\cdots\right)^n \\ &\to e^{\frac{1}{2}\theta^2} \text{ as } n\to\infty \end{aligned}$$

which is the m.g.f. of the $N(0,1)$ random variable. □

Remark. If the m.g.f. is not defined the proof needs the characteristic function:

$$E\left[e^{i\theta X}\right].$$

For example, the Cauchy distribution (defined in §20.3) has no moment generating function, since $E[X^r]$ does not exist for odd valued r. The characteristic function exists and is equal to $e^{-|\theta|}$.

However, the CLT does not hold for Cauchy r.vs since the mean is undefined.

23.2 Normal approximation to the binomial

If $S_n \sim B(n,p)$, so that $X_i = 1$ and 0 with probabilities p and $1-p$, respectively, then

$$\frac{S_n - np}{\sqrt{npq}} \simeq N(0,1).$$

This is called the **normal approximation the binomial** distribution. It applies as $n\to\infty$ with p constant. Earlier we discussed the Poisson approximation to the binomial, which applies when $n\to\infty$ and $np\to\lambda$.

Example 23.3. Two competing airplanes fly a route. Each of n passengers selects one of the 2 planes at random. The number of passengers in plane 1 is

$$S \sim B(n,1/2).$$

Suppose each plane has s seats and let

$$f(s) = P\left(S > s\right) = P\left(\frac{S-\frac{1}{2}n}{\frac{1}{2}\sqrt{n}} > \frac{s-\frac{1}{2}n}{\frac{1}{2}\sqrt{n}}\right).$$

Then

$$\frac{S-np}{\sqrt{npq}} \simeq N(0,1) \implies f(s) \approx 1 - \Phi\left(\frac{2s-n}{\sqrt{n}}\right).$$

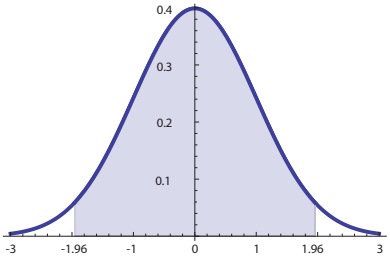
So if $n = 1000$ and $s = 537$ then $(2s-n)/\sqrt{n} = 2.34$. So $\Phi(2.34) \approx 0.99$, and $f(s) \approx 0.01$. Planes need hold 1074 seats, only 74 in excess.

Example 23.4. An unknown fraction of the electorate, p, vote Labour. It is desired to find p within an error not exceeding 0.005. How large should the sample be?

Let the fraction of Labour votes in the sample be $p' = S_n/n$. We can never be certain (without complete enumeration) that $|p'-p| \leq 0.005$. Instead choose n so that the event $|p'-p| \leq 0.005$ has probability ≥ 0.95.

$$P\left(|p'-p| \leq 0.005\right) = P\left(|S_n - np| \leq 0.005n\right) = P\left(\frac{|S_n-np|}{\sqrt{npq}} \leq \frac{0.005\sqrt{n}}{\sqrt{pq}}\right)$$

Choose n such that the probability is ≥ 0.95. It often helps to make a sketch to see what is required:



$$\int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.95.$$

We must choose n so that

$$\frac{0.005\sqrt{n}}{\sqrt{pq}} \geq \Phi^{-1}(0.975) = 1.96.$$

But we don't know p. But $pq \leq \frac{1}{4}$ with the worst case $p = q = \frac{1}{2}$. So we need

$$n \geq \frac{1.96^2}{0.005^2}\frac{1}{4} \simeq 38,416.$$

If we replace 0.005 by 0.03 then $n \simeq 1,068$ will be sufficient. This is typical of the sample size used in commercial and newspaper opinion polls.

Notice that the answer does not depend upon the total population.

23.3 Estimating π with Buffon's needle

Example 23.5. A needle of length ℓ is tossed at random onto a floor marked with parallel lines a distance L apart, where $\ell \leq L$. Recall from Example 18.5 that $p = P(\text{the needle intersects one of the parallel lines}) = 2\ell/(\pi L)$.

Suppose we drop the needle n times. The number of times that it hits a line will be $N \sim B(n,p)$, which is approximately $N(np, np(1-p))$. We estimate p by $\hat{p} = N/n$, which is approximately $N(p, p(1-p)/n)$, and π by

$$\hat{\pi} = \frac{2\ell}{(N/n)L} = \pi \frac{2\ell/(\pi L)}{E(N/n) + (N/n - E(N/n))} = \pi \frac{p}{p + (\hat{p} - p)}$$
$$= \pi \left(1 - \frac{\hat{p} - p}{p} + \cdots\right)$$

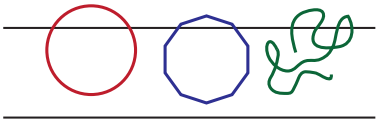
So $\hat{\pi} - \pi$ is approximately distributed as $N(0, \pi^2 p(1-p)/(np^2))$. The variance is minimized by taking p as large as possible, i.e. $\ell = L$. Then

$$\hat{\pi} - \pi \sim N\left(0, \frac{(\pi-2)\pi^2}{2n}\right).$$

$$P(|\hat{\pi} - \pi| < 0.001) \geq 0.95 \iff 0.001\sqrt{\frac{2n}{(\pi-2)\pi^2}} \geq \Phi^{-1}(0.975) = 1.96.$$

This requires $n \geq 2.16 \times 10^7$.

Example 23.6 [*Buffon's noodle*]. Here is another way to show that $p = 2\ell/(\pi L)$.



Notice that a circle of diameter L, no matter where it is placed, intersects the parallel lines exactly twice. Approximate this circle by the boundary of a k-sided regular polygon made up of $k = \pi L/\delta$ rice grains, each of tiny length δ. The probability that a rice grain intersects a line is thus approximately $2/k = 2\delta/(\pi L)$. (For a rigorous proof, one could use two polygons that are inscribed and superscribed to the circle.)

A pin of length ℓ is like ℓ/δ rice grains laid end to end, and so the expected number of times such a pin intersects the lines is $(\ell/\delta) \times 2\delta/(\pi L) = 2\ell/(\pi L)$. At most one rice grain intersects a line, so this must be p, the probability the pin intersects the lines.

This also shows that the expected number of times that a "noodle" of length ℓ crosses the parallel lines is p, irrespective of the shape of the noodle So we might also toss a flexible wet noodle onto the lines, counting the number of crossings N obtained by tossing it n times. Again it is the case that $\hat{\pi} = 2\ell n/(NL) \to \pi$.

24 Continuing studies in probability

Large deviations and Chernoff bound. *Random matrices and Wigner's semicircle law*. I'll perhaps talk about courses in IB. Concluding remarks and wrap up.

24.1 Large deviations

Example 24.1 [*Gambler's success*]. John plays roulette at Las Vegas, betting £1 on red at each turn, which is then doubled, with probability $p < 1/2$, or lost, with probability $q = 1 - p$. The wheel has 38 slots: 18 red, 18 black, 0 and 00; so $p = 18/38$. He tries to increase his wealth by £100.

This is very unlikely, since by (15.1) the probability he should ever be up by 100 is $(p/q)^{100} = (9/10)^{100} = 0.0000265$.

However, suppose this 'large deviation' occurs and after n games he is up by £100. What can we say about n, and about the path followed by John's wealth?

Preliminaries. Let $S_n = X_1 + \cdots + X_n$ be the number of games he wins in the first n, where $X_1, \ldots, X_n$ are i.i.d. $B(1, p)$. His wealth is $W_n = 2S_n - n$. Let $\mu = EX_1 = p$ and $\sigma^2 = \mathrm{Var}(X_i) = pq$. Note that $EW_n = -n/19 < 0$.

He must win $n/2 + 50$ games.

How likely is $P(S_n > na)$ for $a > p$? Using the Chebyshev bound

$$P(S_n > na) = P\left(S_n - n\mu \geq n(a-\mu)\right) \leq \frac{\mathrm{Var}(S_n)}{n^2(a-\mu)^2} = \frac{\sigma^2}{n(a-\mu)^2}.$$

Alternatively, by the Central limit theorem,

$$P(S_n > na)) = P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} > \frac{(a-\mu)\sqrt{n}}{\sigma}\right) \approx 1 - \Phi\left(\frac{(a-\mu)\sqrt{n}}{\sigma}\right).$$

Both show that $P(S_n > na) \to 0$ as $n \to \infty$. □

24.2 Chernoff bound

Another bound is as follows. Let $m(\theta) = Ee^{\theta X_1}$ be the m.g.f. of X_1. Let $\theta > 0$,

$$\begin{aligned} P(S_n > na) &= P\big(e^{\theta S_n} > e^{\theta na}\big) \\ &\leq \frac{E[e^{\theta S_n}]}{e^{\theta na}} \quad \text{(by Markov inequality)} \\ &= \left(\frac{m(\theta)}{e^{\theta a}}\right)^n \\ &= e^{-n[\theta a - \log m(\theta)]}. \end{aligned}$$

Now minimize the right-hand side over θ to get the best bound. This implies

$$P(S_n > na) \le e^{-nI(a)} \quad \text{(the } \textbf{Chernoff bound}\text{)}, \tag{24.1}$$

where $I(a) = \max_{\theta>0}[\theta a - \log m(\theta)]$.

This bound is tight, in the sense that one can also prove that given any $\delta > 0$,

$$P(S_n > na) \ge e^{-n(I(a)+\delta)}, \tag{24.2}$$

for all sufficiently large n. It follows from (24.1)–(24.2) that $\log P(S_n > an) \sim -nI(a)$.

As usual, $\sim$ means that the quotient of the two sides tends to 1 as $n \to \infty$.

This holds for random variables more generally. For example, if $X_i \sim N(0,1)$ then $m(\theta) = e^{\frac{1}{2}\theta^2}$ and $I(a) = \max_\theta \left[\theta a - \frac{1}{2}\theta^2\right] = \frac{1}{2}a^2$. So $\log P(S_n > an) \sim -n\frac{1}{2}a^2$.

For $B(1,p)$ the m.g.f. is $m(\theta) = q + pe^\theta$ and

$$I(a) = \max_\theta[\theta a - m(\theta] = (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p}.$$

The function I is convex in a, with its minimum being $I(p) = 0$.

We can also verify the lower bound (24.2). Let $j_n = \lceil na \rceil$. Then

$$P(S_n > na) = \sum_{i \ge j_n}^{n} \binom{n}{i} p^i (1-p)^{n-i} > \binom{n}{j_n} p^{j_n}(1-p)^{n-j_n}.$$

By applying Stirling's formula on the right hand side we may find:

$$\underline{\lim}_{n\to\infty}(1/n)\log P(S_n > na) \ge -I(a).$$

Hence $\log P(S_n > an) \sim -nI(a)$.

Most likely way to £100. Consider the path on which John's wealth increases to 100. We can argue that this is most likely to look like a straight line. For instance, suppose S_n increases at rate a_1 for n_1 bets, and then rate a_2 for n_2 bets, where $n_1 + n_2 = n$ and $2(n_1a_1 + n_2a_2) - n = 100$. The log-probability of this is about $-n_1I(a_1) - n_2I(a_2)$, which is maximized by $a_1 = a_2$, since I is a convex function.

So the most likely route to 100 is over n bets, with S_n increasing at a constant rate a, and such that $2na - n = 100$. Subject to these constraints $\log P(S_n > an) \approx -nI(a)$ is maximized by $n = 100/(1-2p)$, $a = 1-2p$.

This means it is highly likely that $n \approx 100/(1 - 2\times(18/38)) = 1900$. Interestingly, this is the same as the number of games over which his expected loss would be £100. □

This is an example from the theory of **large deviations**. Informally, we might say that if a rare event does occur then it does so in whatever manner is most likely.

24.3 Random matrices

Random matrices arise in many places: as the adjacency matrix of a random graph, as the sample correlation matrix of a random sample of multivariate data, and in quantum physics, numerical analysis and number theory.

Consider a symmetric $n \times n$ matrix A, constructed by setting diagonal elements 0, and independently choosing each off-diagonal $a_{ij} = a_{ji}$ as 1 or -1 by tossing a fair coin.

A random 10×10 symmetric matrix, having eigenvalues -4.515, -4.264, -2.667, -1.345, -0.7234, 1.169, 2.162, 2.626, 3.279, 4.277.

$$\begin{pmatrix} 0 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & 1 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 0 & -1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 0 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 0 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 & 0 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 & 1 & 0 & 1 & -1 \\ -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 0 & -1 \\ -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 0 \end{pmatrix}$$

Recall that the eigenvalues of a symmetric real matrix are real.

Let Λ be a randomly chosen eigenvalue of a random A. What can we say about Λ?

Since A and $-A$ are equally likely, $E\Lambda^k = 0$ if k is odd.

Consider $k = 4$. Suppose the eigenvalues of A are $\lambda_1, \ldots, \lambda_n$.

$$E[\Lambda^4] = \frac{1}{n} E[\lambda_1^4 + \cdots + \lambda_n^4] = \frac{1}{n} E[\mathrm{Tr}(A^4)].$$

Now

$$E[\mathrm{Tr}(A^4)] = E \sum_{i_1,i_2,i_3,i_4} a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_4} a_{i_4 i_1} = \sum_{i_1,i_2,i_3,i_4} E[a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_4} a_{i_4 i_1}] \qquad (24.3)$$

where the sum is taken over all possible paths of length 4 through a subset of the n indices: $i_1 \to i_2 \to i_3 \to i_4 \to i_1$.

A term in (24.3), $E[a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_4} a_{i_4 i_1}]$, is either 1 or 0. It is 1 iff for each two indices i, j the total number of a_{ij}s and a_{ji}s contained in $\{a_{i_1 i_2}, a_{i_2 i_3}, a_{i_3 i_4}, a_{i_4 i_1}\}$ is even.

Let i, j, k range over triples of distinct indices. $E[a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_4} a_{i_4 i_1}] = 1$ for

- $n(n-1)(n-2)$ terms of the form $E[a_{ij} a_{ji} a_{ik} a_{ki}]$;
- $n(n-1)(n-2)$ terms of the form $E[a_{ij} a_{jk} a_{kj} a_{ji}]$;
- $n(n-1)$ terms of the form $E[a_{ij} a_{ji} a_{ij} a_{ji}]$.

Thus, $E[\Lambda^4 / n^{\frac{4}{2}}] = n^{-\frac{4}{2}-1} E[\mathrm{Tr}(A^4)] = n^{-3}[2n(n-1)(n-2) + n(n-1)] \to 2$ as $n \to \infty$.

The limit 2 is C_2, the number of Dyke words of length 4. These words are ()() and (()), which correspond to the patterns of the first two bullet points above.

This argument easily generalizes to any even k, to show $\lim_{n\to\infty} E[(\Lambda/n^{\frac{1}{2}})^k] \to C_{k/2}$, a Catalan number, and the number of Dyck words of length k (described in §12.2).

This begs the question: what random variable X has sequence of moments

$$\{EX^k\}_{k=1}^{\infty} = \{0, C_1, 0, C_2, 0, C_3, \dots\} = \{0, 1, 0, 2, 0, 5, 0, 14, \dots\}?$$

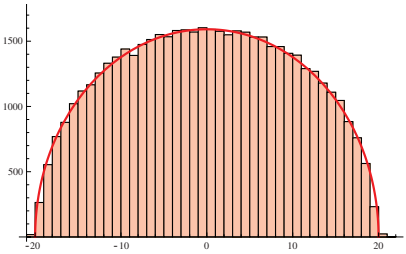
It is easy to check that this is true when X has the p.d.f.

$$f(x) = \tfrac{1}{2\pi}\sqrt{4 - x^2}, \quad -2 \le x \le 2.$$

Here is a histogram of 50,000 eigenvalues obtained by randomly generating 500 random 100×100 matrices. Bin sizes are of width 1. Consistent with the above analysis, $\lambda/\sqrt{100}$ has empirical density closely matching f.

Rescaling appropriately, the red semicircle is

$$g(x) = 50000\tfrac{1}{10} f(\tfrac{x}{10}), \quad -20 \le x \le 20.$$



This result is **Wigner's semicircle law**. Notice that our argument does not really need the assumption that a_{ij} are chosen from the discrete uniform distribution on $\{-1, 1\}$. We need only that $Ea_{ij}^k = 0$ for odd k and $Ea_{ij}^k < \infty$ for even k.

This means that Wigner's theorem is in the same spirit as the Central limit theorem in Lecture 23, which holds for any random variable with finite first two moments. Wigner's theorem dates from 1955, but the finer analysis of the eigenvalues structure of random matrices interests researchers in the present day.

24.4 Concluding remarks

In §24.1–24.3 we have seen some fruits of research in probability in modern times. In doing so we have touched on many topics covered in our course: Markov and Chebyshev inequalities, moment generating function, sums of Bernoulli r.vs, Stirling's formula, normal distribution, gambler's ruin, Dyke words, generating functions, and Central limit theorem.

In Appendix H you can find some notes about further courses in the Mathematical Tripos in which probability features.

I'll give a final overview of the course and wrap up.

A Problem solving strategies

Students sometimes say that the theory in Probability IA is all very well, but that the tripos questions require ingenuity. That is sometimes true, of questions like 2004/II/12: "Planet Zog is a sphere with centre O. A number N of spaceships land at random on its surface. ...". Of course this makes the subject fun.

But to help with this, let's compile a collection of some frequently helpful problem solving strategies.

1. You are asked to find $P(A)$. For example, A might be the event that at least two out of n people share the same birthday. Might it be easier to calculate $P(A^c)$? The simple fact that $P(A) = 1 - P(A^c)$ can sometimes be remarkably useful. For example, in Lecture 1 we calculated the probability that amongst n people no two have the same birthday.
2. You are asked to find $P(A)$. Is there a partition of A into disjoint events, $B_1, B_2, \ldots, B_n$, so that $P(A) = \sum_i P(B_i)$? Probabilities of intersections are usually easier to calculate than probabilities of unions. Fortunately, the inclusion-exclusion formula gives us a way to convert between the two.
3. You are asked to find $P(A)$. Can A be written as the union of some other events? Might the inclusion-exclusion formula be helpful?
4. The expectation of a sum of random variables is the sum of their expectations. The random variables need not be independent. This is particularly useful when applied to indicator random variables. You are asked to find EN, the expected number of times that something occurs. Can you write $N = I_1 + \cdots + I_k$, where this is a sum of indicator variables, each of which is concerned with a distinct way in which N can be incremented? This idea was used Example 8.3 in lectures, where N was the number of couples seated next to one another. It is the way to do the Planet Zog question, mentioned above. You can use it in Examples sheet 2, #11.
5. You are asked to place a bound on some probability. Can you use one of the inequalities that you know? Boole, Bonferroni, Markov, Chebyshev, Jensen, Cauchy-Schwarz, AM-GM).
6. You are asked something about sums of independent random variables. Might a probability generating function help? For example, suppose X has the geometric distribution $P(X = r) = (1/2)^{r+1}$, $r = 0, 1, \ldots$. Is it possible for X to have the same distribution as $Y_1 + Y_2$, where Y_1, Y_2 are some two independent random variables with the same distribution? Hint: what would the p.g.f. of Y_i have to be?
7. This is like 2 above, but with the tower property of conditional expectation. You are asked to find EX. Maybe there is some Y so that $E[X]$ is most easily computed as $E[E[X \mid Y]] = \sum_y E[X \mid Y = y]\, P(Y = y)$.

8. Learn well all the distributions that we cover in the course and understand the relations between them. For example, if $X \sim U[0,1]$ then $-\log X \sim \mathscr{E}(1)$. Learn also all special properties: such the memoryless property of the geometric and exponential distributions. As you approach a question ask yourself, what distribution(s) are involved in this question? Is some special property or relationship useful?

9. You are given the joint density function of continuous random variables $X_1, \ldots, X_n$ and want to prove that they are independent. Try to spot, by inspection, how to factor this as $f(x_1, \ldots, x_n) = f_1(x_i) \cdots f_n(x_n)$, where each f_i is a p.d.f.

10. In questions about transformation of continuous random variables there are a couple things to look out for. Always start with a bijection between $R \subseteq \mathbb{R}^n$ and $S \subseteq \mathbb{R}^n$ and make sure that you specify S correctly. If $Y_1, \ldots, Y_n$ is more variables that really interest you, then, having found the joint density of them all, you can always integrate out the superfluous ones. In computing the Jacobian J, remember that it is sometimes easier to compute $1/J = \partial(y_1, \ldots, y_n)/\partial(x_1, \ldots, x_n)$.

B Fast Fourier transform and p.g.fs

Although not examinable, a study of how the fast Fourier transform (FFT) can be used to sum random variables provides a good exercise with probability generating functions. This method is used in practice in financial mathematics, such as when calculating the aggregate loss distribution of a portfolio of insurance risks.

Suppose we wish to find the distribution of $Y = X_1 + X_2$, where the X_i are i.i.d. and have have p.g.f. $p(z) = p_0 + p_1 z + \cdots + p_{N-1} z^{N-1}$. The p.g.f. of Y is $p_Y(z) = p(z)^2$, which can be found by making $O(N^2)$ multiplications of the form $p_i p_j$. Assuming multiplications take time, we say the time-complexity is $O(N^2)$.

With the Fast Fourier Transform we can reduce the time-complexity to $O(N \log N)$. The steps are as follows.

(a) Compute $p(z)$ at each $z = \omega^0, \omega^1 \ldots, \omega^{2N-1}$. where $\omega = e^{-2\pi i/(2N)}$.
This is the discrete Fourier transform (DFT) of the sequence $(p_0, p_1, \ldots, p_{N-1})$.

(b) Compute $p_Y(z) = p(z)^2$, at each $z = \omega^0, \omega^1 \ldots, \omega^{2N-1}$.

(c) Recover the distribution $(P(Y = y),\ y = 0, \ldots, 2N-2)$.
To do this we use an inverse DFT, for which the calculation is almost the same as doing a DFT, as in step (a).

Step (b) takes $2N$ multiplications. Steps (a) and (c) are computed using the fast Fourier transform in $O\big(2N \log(2N)\big)$ multiplications (of complex numbers).

A feeling for way in which the FFT simplifies the calculation can be obtained by studying below that case $N = 4$, $\omega = e^{-i\pi/2} = -i$. Notice how the 16 multiplications in the left-hand table become many fewer multiplications in the right-hand table. Note also that $(p_0 + p_2, p_0 - p_2)$ is the DFT of (p_0, p_2).

z	$p(z)$
ω^0	$p_0\omega^0 + p_1\omega^0 + p_2\omega^0 + p_3\omega^0$
ω^1	$p_0\omega^0 + p_1\omega^1 + p_2\omega^2 + p_3\omega^3$
ω^2	$p_0\omega^0 + p_1\omega^2 + p_2\omega^4 + p_3\omega^6$
ω^3	$p_0\omega^0 + p_1\omega^3 + p_2\omega^6 + p_3\omega^9$

$=$

$p(z)$		
$(p_0 + p_2)$	$+$	$(p_1 + p_3)$
$(p_0 - p_2)$	$+$	$\omega(p_1 - p_3)$
$(p_0 + p_2)$	$-$	$(p_1 + p_3)$
$(p_0 - p_2)$	$-$	$\omega(p_1 - p_3)$

The key idea is to find DFTs of the sequences $(p_0, p_2, \ldots, N-2)$, and $(p_1, p_3, \ldots, N-1)$ and then combine them to create the DFT of $(p_0, p_1, p_2, p_3, \ldots, N-1)$. Combining takes only $O(N)$ multiplications. Suppose N is a power of 2. We may recursively repeat this trick of division into two half-size problems, until we are making DFTs of sequences of length 2. Because we repeat the trick $\log_2 N$ times, and N multiplications are needed at each stage, the FFT is of time complexity $O(N \log N)$.

C The Jacobian

We quoted without proof in §20.1 the fact that the Jacobian gives the right scaling factor to insert at each point when one wishes to compute the integral of a function over a subset of $\mathbb{R}^n$ after a change of variables. We used this to argue that if $X_1, \dots, X_n$ have joint density f, then $Y_1, \dots, Y_n$ have joint density g, where $g(y_1, \dots, y_n) = f(x_1(y_1, \dots, y_n), \dots, x_n(y_1, \dots, y_n))|J|$, and J (the Jacobian) is the determinant of the matrix whose (i, j)th element is $\partial x_i / \partial y_j$.

This works because: (i) every differentiable map is locally linear, and (ii) under a linear change of coordinates, such as $y = Ax$, a cube in the $x-$coordinate system becomes a parallelepiped in the y-coordinate system. The n-volume of a parallelepiped is the determinant of its edge vectors (i.e. columns of A). For a proof this fact see this nice essay, A short thing about determinants, by Gareth Taylor.

Warning. What follows next is peripheral to Probability IA. It's a digression I set myself to satisfy my curiosity. I know that in IA Vector Calculus you see a proof for change of variables with Jacobian in $\mathbb{R}^2$ and $\mathbb{R}^3$. But so far as I can tell, there is no course in the tripos where this formula is proved for $\mathbb{R}^n$. I have wondered how to make the formula seem more intuitive, and explain the why $|J|$ plays the role it does.

Here now, for the curious, is a motivating argument in $\mathbb{R}^n$. I use some facts from IA Vectors and Matrices.

Let's start with a 1–1 linear map, $y = r(x) = Ax$, where A is $n \times n$ and invertible. The inverse function is $x = s(y) = A^{-1}y$.

Let $Q = A^T A$. This matrix is positive definite and symmetric, and so has positive real eigenvalues.[1] Consider the sphere $S = \{y : y^T y \leq 1\}$. Its preimage is

$$R = \{x : (Ax)^T Ax \leq 1\} = \{x : x^T Q x \leq 1\}.$$

Let $e_1, \dots, e_n$ be orthogonal unit-length eigenvectors of Q, with corresponding eigenvalues $\lambda_1, \dots, \lambda_n$; the eigenvalues are strictly positive. Then

$$R = \left\{x : x = \textstyle\sum_i \alpha_i e_i,\ \alpha \in \mathbb{R}^n,\ \sum_i (\alpha_i \sqrt{\lambda_i})^2 \leq 1\right\},$$

which is an ellipsoid in $\mathbb{R}^n$, whose orthogonal axes are in the directions of the e_is.

We can view this ellipsoid as having been obtained from a unit sphere by rescaling, (i.e. squashing or stretching), by factors $\lambda_1^{-1/2}, \dots, \lambda_n^{-1/2}$ in the directions $e_1, \dots, e_n$, respectively. The volume is altered by the product of these factors. The determinant of a matrix is the product of its eigenvalues, so

$$\frac{\text{vol}(R)}{\text{vol}(S)} = \frac{1}{(\lambda_1 \cdots \lambda_n)^{1/2}} = \frac{1}{\sqrt{\det(A^T A)}} = \frac{1}{|\det(A)|} = |\det(A^{-1})| = |J|,$$

[1] If you don't already know these facts, they are quickly explained. If z is an eigenvector, with eigenvalue λ, and $\bar{z}$ is its complex conjugate, then $Qz = \lambda z$ and $Q\bar{z} = \bar{\lambda}\bar{z}$. So $z^T Q\bar{z} - \bar{z}^T Qz = 0 = (\bar{\lambda} - \lambda)\bar{z}^T z$, hence λ is real. Also, $z \neq 0 \implies Az \neq 0 \implies \lambda z^T z = z^T Qz = (Az)^T (Az) > 0$, so $\lambda > 0$.

where

$$|J| = \frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)}.$$

So $\mathrm{vol}(R) = |J| \, \mathrm{vol}(S)$.

Now reimagine S, not as a unit sphere, but a very tiny sphere centred on $\bar{y}$. The sphere is so small that $f(s(y)) \approx f(s(\bar{y}))$ for $y \in S$. The preimage of S is the tiny ellipsoid R, centred on $\bar{x} = (s_1(\bar{y}), \ldots, s_n(\bar{y}))$. So

$$\begin{aligned} \int_{x \in R} f(x) \, dx_1 \ldots dx_n &\approx f(\bar{x}) \, \mathrm{vol}(R) \\ &= f(s_1(\bar{y}), \ldots, s_n(\bar{y})) |J| \, \mathrm{vol}(S) \qquad \text{(C.1)} \\ &\approx \int_{y \in S} f(s_1(y), \ldots, s_n(y)) |J| \, dy_1 \ldots dy_n. \end{aligned}$$

The above has been argued for the linear map $y = r(x) = Ax$. But locally any differentiable map is nearly linear.

Further details are needed to complete a proper proof. We will need to approximate the integrals over some more general regions R and S within $\mathbb{R}^n$ by sums of integrals over tiny sphere and ellipsoids, making appropriate local linearizations of functions r and f, where $r : R \to S$ and $f : R \to \mathbb{R}^+$. But at least after reaching (C.1) we have an intuitive reason for the formula used in §20.1.

D Beta distribution

The Beta(a, b) distribution has p.d.f.

$$f(x : a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 \le x \le 1,$$

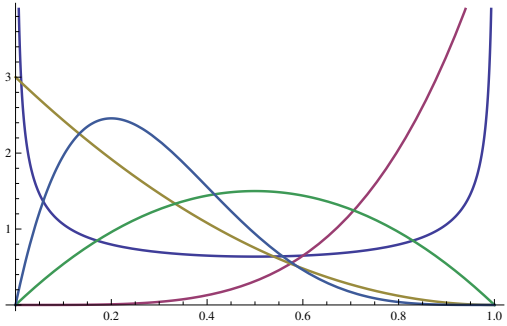
where $a, b > 0$. It is used by actuaries to model the loss of an insurance risk and is a popular choice for the following reasons.

1. The beta distribution has a minimum and a maximum. Reinsurance policies for catastrophic risks are typically written so that if a claim is made then the insurer's exposure will lie between some known minimum and maximum values.

   Suppose a policy insures against a loss event (an earthquake, say) that occurs with a probability of $p = 1 \times 10^{-2}$ per year. If the event occurs the claim will cost the insurer between 100 and 200 million pounds. The annual loss could be modelled by the r.v. $(100 + 2X)Y$, where $Y \sim B(1, p)$ and $X \sim \text{Beta}(a, b)$.

2. The two parameters can be chosen to fit a given mean and variance. The mean is $\frac{a}{a+b}$. The variance is $\frac{ab}{(a+b)^2(a+b+1)}$.

3. The p.d.f. can take many different shapes.



However, the sum of beta distributed r.vs is nothing simple. A portfolio might consist of 10,000 risks, each assumed to be beta distributed. To calculate $P(\sum_{i=1}^{10000} X_i > t)$ one must discretize the distributions and use discrete Fourier transforms, as described in Appendix B.

The moment generating function of the $B(a, b)$ distribution is complicated!

$$m(\theta) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{a+r}{a+b+r} \right) \frac{\theta^k}{k!}.$$

E Kelly criterion

I have not this found a place to include this during a lecture this year, but I leave it in the notes, as some people may enjoy reading it. Maybe I will use this another year.

Consider a bet in which you will double your money or lose it all, with probabilities $p = 0.52$ and $q = 0.48$ respectively. You start with $£X_0$ and wish to place a sequence of bets and maximize the expected growth rate.

If at each bet you wager a fraction f of your capital then $EX_{n+1} = p(1+f)X_n + q(1-f)X_n = X_n + (p-q)fX_n$. This is maximized by $f = 1$, but then there is a large probability that you will go bankrupt. By choosing $f < 1$ you will never go bankrupt.

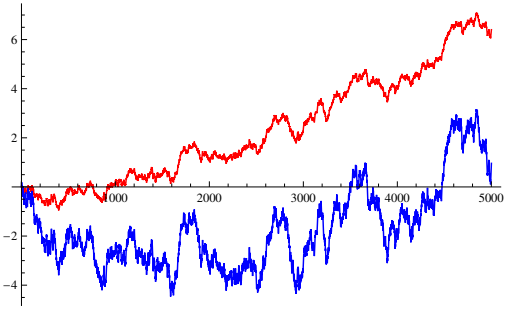
Suppose after n bets you have $£X_n$. You wish to choose f so as to maximize the compound growth rate,

$$\text{CGR} = (X_n/X_0)^{1/n} - 1 = \exp\left(\frac{1}{n}\textstyle\sum_{i=1}^{n} \log(X_i/X_{i-1})\right) - 1.$$

Now $\frac{1}{n}\sum_{i=1}^{n} \log(X_i/X_{i-1}) \to E\log(X_1/X_0)$ as $n \to \infty$ and this should be maximized.

$$E\log(X_1/X_0) = p\log(1+f) + q\log(1-f)$$

which is maximized by $f = p - q$. This is the Kelly criterion.



For $p = 0.52$ a plot of $\log X_n$, $n = 0, \ldots, 5000$,
for Kelly betting $f = 0.04$ (red) and for $f = 0.10$ (blue).

F Ballot theorem

I have not this found a place to include this during a lecture this year, but I leave it in the notes, as some people may enjoy reading it. Maybe I will use this another year.

A famous problem of 19th century probability is Bertrand's ballot problem, posed as the question: "In an election where candidate A receives a votes and candidate B receives b votes with $a > b$, what is the probability that A will be strictly ahead of B throughout the count?"

Equivalently, we are asking for the number of paths of length $a + b$ that start at the origin and end at $T = (a + b, a - b)$. Since every good path must start with an upstep, there are as many good paths as there are paths from $(1, 1)$ to T that never touch the x-axis. The set of paths from $(1, 1)$ to T that do touch the x-axis is in one-to-one correspondence with the set of paths from $(1, -1)$ to T; this is seen by reflecting across the x-axis the initial segment of the path that ends with the step that first touches the x-axis. Subtracting the number of these paths from the number of all paths from $(1, 1)$ to T produces the number of good paths:

$$\binom{a+b-1}{a-1} - \binom{a+b-1}{a} = \frac{a-b}{a+b}\binom{a+b}{a}.$$

So the answer to Bertrand's question is $\frac{a-b}{a+b}$.

Alternatively, we can derive this answer by noticing that the probability that the first step is down, given that the path ends at T is $b/(a + b)$. So the number of paths from $(1, -1)$ to T is $\frac{b}{a+b}\binom{a+b}{a}$. The number that go from $(0, 0)$ to T without returning to the x-axis is therefore $(1 - 2\frac{b}{a+b})\binom{a+b}{a} = \frac{a-b}{a+b}\binom{a+b}{a}$.

G Allais paradox

I have not this found a place to include this during a lecture this year, but I leave it in the notes, as some people may enjoy reading it. This is an interesting paradox in the theory of gambling and utility.

Which of the following would you prefer, Gamble 1A or 1B?

Experiment 1			
Gamble 1A		Gamble 1B	
Winnings	Chance	Winnings	Chance
\$1 million	100%	\$1 million	89%
		Nothing	1%
		\$5 million	10%

Now look at Gambles 2A and 2B. Which do you prefer?

Experiment 2			
Gamble 2A		Gamble 2B	
Winnings	Chance	Winnings	Chance
Nothing	89%	Nothing	90%
\$1 million	11%	\$5 million	10%

When presented with a choice between 1A and 1B, most people choose 1A.

When presented with a choice between 2A and 2B, most people choose 2B.

But this is inconsistent!

Why? Experiment 1 is the same as Experiment 2, but with the added chance of winning \$1 million with probability 0.89 (irrespective of which gamble is taken).

H IB courses in applicable mathematics

Here is some advice about courses in applicable mathematics that can be studied in the second year: Statistics, Markov Chains and Optimization courses.

Statistics. Statistics addresses the question, "What is this data telling me?" How should we design experiments and interpret their results? In the Probability IA course we had the weak law of large numbers. This underlies the frequentist approach of estimating the probability p with which a drawing pin lands "point up" by tossing it many times and looking at the proportion of landings "point up". Bayes Theorem also enters into statistics. To address questions about estimation and hypothesis testing we must model uncertainty and the way data is arises. That gives Probability a central role in Statistics. In the Statistics IB course you will put to good use what you have learned about random variables and distributions this year.

Markov chains. A Markov Chain is a generalization of the idea of a sequence of i.i.d. r.vs., $X_1, X_2, \ldots$. There is a departure from independence because we now allow the distribution of X_{n+1} to depend on the value of X_n. Many things in the world are like this: e.g. tomorrow's weather state follows in some random way from today's weather state. You have already met a Markov chain in the random walks that we have studied in Probability IA. In Markov Chains IB you will learn many more interesting things about random walks and other Markov chains. If I were lecturing the course I would tell you why Polya's theorem about random walk implies that it is possible to play the clarinet in our 3-D world but that this would be impossible in 2-D Flatland.

Optimization. Probability is less important in this course, but it enters when we look at randomizing strategies in two-person games. You probably know the game scissors-stone-paper, for which the optimal strategy is to randomize with probabilities 1/3, 1/3, 1/3. In the Optimization course you will learn how to solve other games. Here is one you will be able to solve (from a recent Ph.D. thesis I read):

I have lost k possessions around my room (keys, wallet, phone, etc). Searching location i costs c_i. "Sod's Law" predicts that I will have lost my objects in whatever way makes the task of finding them most difficult. Assume there are n locations, and cost of searching location i is c_i. I will search until I find all my objects. What is Sod's Law? How do I minimize my expected total search cost? (I will have to randomize.)

Part II. The following Part II courses are ones in which Probability features. If are interested in studying these then do Markov Chains in IB. For Probability and Measure it is also essential to study Analysis II in IB.

Probability and Measure	Stochastic Financial Models
Applied Probability	Coding and Cryptography
Principles of Statistics	Optimization and Control

Index

absorption, 61
aggregate loss distribution, 52
Allais paradox, 106
arcsine law, 5
atom, 64
axioms of probability, 14

ballot theorem, 105
Baye's formula, 23
Bell numbers, 8, 9
Benford's law, 45
Bernoulli distribution, 20
Bernoulli trials, 20
Bertrand's paradox, 72
beta distribution, 85, 103
binomial coefficient, 9
binomial distribution, 20, 31
bivariate normal random variable, 88
Bonferroni's inequalities, 18
Boole's inequality, 15
branching process, 54
    generating function, 54
    probability of extinction, 56
Buffon's needle, 73, 93

Catalan number, 49
Cauchy distribution, 81, 91
Cauchy-Schwarz inequality, 39
Central limit theorem, 90
characteristic function, 52
Chebyshev inequality, 42
Chernoff bound, 94
classical probability, 1
concave function, 38
conditional
    distribution, 50
    expectation, 50
    probability, 22
conditional entropy, 53
continuity theorem, 90
continuous random variable, 62
continuum, 62
convex function, 38
convolution, 52, 80, 81
correlation coefficient, 41
covariance, 40
cumulative distribution function, 63

dependent events, 19
derangement, 17
discrete distribution, 21
discrete Fourier transform, 100
discrete random variable, 27
discrete uniform distribution, 27
disjoint events, 2
distribution, 27
distribution function, 63
Dyck word, 49, 96

Efron's dice, 36
entropy, 41, 53
Erlang distribution, 85
event, 2
expectation, 27, 67
    of a sum, 30, 51, 52
expectation ordering, 36, 68
experimental design, 36
exponential distribution, 64, 82, 84
    memoryless property, 64
    moment generating function of, 84
extinction probability, 56

fast Fourier transform, 53, 100
Fibonacci number, 45, 49
Fundamental rule of counting, 6

gambler's ruin, 23, 58
    duration of the game, 60
gamma distribution, 84, 85
generating functions, 48, 54, 61
geometric distribution, 21, 31
    memoryless property, 64
geometric probability, 71, 73

hazard rate, 65
hypergeometric distribution, 21

Inclusion-exclusion formula, 17, 33
independent
    events, 19
    random variables, 34
independent random variables, 71
indicator function, 32
information entropy, 41, 53
inspection paradox, 69
insurance industry, 21, 103

Jacobian, 79, 101
Jensen's inequality, 38
joint distribution, 50
joint distribution function, 70
joint moment generating function, 89
joint probability density function, 70
jointly distributed continuous random variables, 70

Kelly criterion, 104
Kolmogorov, 14

large deviations, 94, 95
law of total expectation, 51
law of total probability, 23

marginal distribution, 50, 70
Markov inequality, 42
mean, 27
median, 76
memoryless property, 21, 64
mode, 76
moment, 84
moment generating function, 83
    multivariate, 89
multinomial
    coefficient, 10
    distribution, 20
multivariate normal density, 88
mutually exclusive, 2
mutually independent, 20

non-transitive dice, 36
normal approximation the binomial, 91
normal distribution, 74
    bounds on tail probabilities, 87
    moment generating function of, 86

observation, 2
order statistics, 76, 82

partition of the sample space, 23, 51
Poisson distribution, 21, 31
    approximation to the binomial, 21
probabilistic method, 16
probability
    axioms, 14
    density function, 62
    distribution, 14, 20
    generating function, 46
    mass function, 27
    measure, 14
    space, 14

Ramsey number, 16
random matrices, 96
random sample, 76
random variable, 27
    continuous, 62
    discrete, 27
random walk, 58
reflection principle, 5, 105

sample mean, 76
sample median, 76
sample space, 2
Simpson's paradox, 24
standard deviation, 30
standard normal distribution, 74
Stirling's formula, 10
stochastic ordering, 36, 68, 69
strictly convex, 38
strong law of large numbers, 44
subadditive set function, 15
submodular function, 15
sum of random variables, 48, 51, 80
symmetric random walk, 58

tower property of conditional expectation, 51
transformation of random variables, 78, 82, 101

uniform distribution, 62, 64, 68

value at risk, 52
variance, 30, 68
variance of a sum, 35, 40

Weak law of large numbers, 43
Wigner's semicircle law, 97

Zipf's law, 33