Warwick Mathematics Exchange

ST111

Probability A

2022, August 12th

Desync, aka The Big Ree

Contents

Introduction

In *Probability A*, we study the basics of probability, with a main focus on discrete probability distributions. The vast majority of techniques developed in this module are required basic knowledge for further study of probability.

This module is very computational in nature, with many questions simply asking you to calculate the probability of some event happening.

This document is intended to broadly cover all the topics within the Probability A module. Knowledge of statistics past regular A-Levels (from Edexcel FS1 and FS2 o.e.) such as the central limit theorem, Poisson distribution, etc., is not assumed, but prior experience is certainly helpful. Basic familiarity with combinatorics is assumed.

This document is not designed to be a replacement for lecture notes - much of the content is covered in a different order than is taught in the course, so it is not recommended to learn the module from these notes unless you are familiar with most of the content already.

Due to the computational nature of this module, this document mainly consists of a checklist of how to solve different questions, with not too much in the way of theory.

Disclaimer: This document was made by a first year student with a severe dislike of statistics (but this module is mostly just combinatorics anyway). I make *absolutely no guarantee* that this document is complete nor without error. In particular, any content covered exclusively in lectures (if any) will not be recorded here. Additionally, this document was written at the end of the 2022 academic year, so any changes in the course since then may not be accurately reflected.

Notes on formatting

New terminology will be introduced in *italics* when used for the first time. Named theorems will also be introduced in *italics*. Important points will be **bold**. Common mistakes will be underlined. The latter two classifications are under my interpretation. YMMV.

Content not taught in the course will be outlined in the margins like this. Anything outlined like this is not examinable, but has been included as it may be helpful to know alternative methods to solve problems.

History

First Edition: 2022-08-12*
Current Edition: 2022-08-12

Authors

This document was written by R.J. Kit L., a maths student. I am not otherwise affiliated with the university, and cannot help you with related matters.

Please send me a PM on Discord @Desync#6290, a message in the WMX server, or an email to Warwick.Mathematics.Exchange@gmail.com for any corrections. (If this document somehow manages to persist for more than a few years, these contact details might be out of date, depending on the maintainers. Please check the most recently updated version you can find.)

If you found this guide helpful and want to support me, you can buy me a coffee!

(Direct link for if hyperlinks are not supported on your device/reader: ko-fi.com/desync.)

*Storing dates in big-endian format is clearly the superior option, as sorting dates lexicographically will also sort dates chronologically, which is a property that little and middle-endian date formats do not share. See ISO-8601 for more details. This footnote was made by the computer science gang.

1 Combinatorics

Combinatorics is the branch of mathematics that deals with enumeration and discrete structures.

The *multiplication principle* states that if you have n options for one item, and m for a second item, then the total number of ways to choose both items is $n \times m$. In general, if there are k items to be chosen, with n_k options for the kth item, then there are $n_1 \times n_2 \times \cdots \times n_k$ ways to choose the k items. We call these *combinations* of the items. Combinations do not care about the ordering of the items in question.

In contrast, say you are arranging 5 different books on a shelf, each with a different colour, and all snugly fitting together. How many ways are there to arrange the books on the shelf?

We call these different orderings *permutations*, so an equivalent question is to ask how many permutations of 5 objects there are.

For the first position on the shelf, we have 5 options to pick from - any book could go in the first place. Then, for the next space, we have 4 options to pick from, then for the next, 3. We multiply all of these together, giving $5 \times 4 \times 3 \times 2 \times 1 = 120$ ways of arranging the 5 books.

Particularly in combinatorics, multiplying sequences of integers occurs often enough that we have notation for this called the *factorial*: $n! := n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$. This function comes up in a lot of places outside of combinatorics, but here, we restrict the factorial function to take natural inputs only. One thing of note is that $0! = 1$, which corresponds to the idea that if you have zero objects, there is only one permutation - having zero objects.

Another point to remember, is that there is only one *combination* of books here. Regardless of how we arrange them, we will end up with the same 5 books on the shelf - combinations do not care about order.

Now, what if we had 8 different books, but still only 5 spaces on the shelf? Then, we have 8 options for the first position, then 7 for the second,..., down to 4 for the fifth space. We can compactly write this as $\frac{8!}{3!}$.

$$\text{Total permutations} = \frac{(\text{Number of items we can choose from})!}{(\text{Number of items we \underline{can't} choose})!}$$

Or, if we have n items and k spaces, these are called k-permutations of n*, and are denoted with $P(n,k)$ or P^n_k, and,

$$P^n_k = \frac{n!}{(n-k)!}$$

Now, say we have a shelf with 3 spaces, and 7 different books. How many ways to choose 3 books out of 7 do we have?

We know there are $P^7_3 = \frac{7!}{(7-3)!} = 210$ permutations. Each set of 3 objects can be arranged in $3! = 6$ ways, so we've counted each combination exactly 6 times, so the number of combinations is $\frac{7!}{4!3!} = 35$.

$$\text{Total combinations} = \frac{(\text{Number of items we can choose from})!}{(\text{Number of items we \underline{can} choose})! \times (\text{Number of items we \underline{can't} choose})!}$$

We call these k-combinations of n objects, denoted $C(n,k)$, C^n_k, or $\binom{n}{k}$.

$$C^n_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Now, we have 5 books again, with 5 spaces on the shelf, but this time, 3 of the books are indistinguishable. How many permutations do we have now?

*Strictly speaking, a "permutation" should use all the elements of a set, being defined as a bijection from a set to itself. See my guide on MA136, or search "symmetric group" for more.

We still start with the $5! = 120$ permutations from before, but now, every permutation has set of matching versions with the indistinguishable books swapped around. Since those permutations are indistinguishable, we divide them out, giving $\frac{5!}{3!} = 20$ total permutations.

In general, for permutations with repeated elements[*], we start as if the items are all distinguishable, then divide out by the repeated elements swapping with each other:

$$\text{Total permutations with repeated objects} = \frac{\text{(Total number of objects)!}}{\text{(Group 1)!} \times \text{(Group 2)!} \times \text{(Group 3)!} \cdots}$$

Example. How many distinct anagrams of the word "MISSISSIPPI" are there?

The word is 11 letters long, and we have 4 groups of letters, M, I, S and P with multiplicities 1, 4, 4, and 2, respectively, so there are,

$$\frac{11!}{1!4!4!2!} = 32\,560$$

such anagrams.

Of course, when rolling 10 D6 dice, there are 6 possible values for the first die, 6 for the second, 6 for the third, and so on, giving 6^{10} possible permutations, but how many possible *combinations*[†] are there when rolling 10 D6 dice?

We let x_i denote the number of dice in with value i per multisubset. The question is then to find the number of non-negative integer solutions[‡] to $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 10$.

We use a method called the *stars and bars*. A solution to the equation can be represented with x_1 symbols called *stars*, followed by a separator called a *bar*, then x_2 more stars, another bar, and so on. For example, the combination of dice values, 1112234446 (order doesn't matter) would be represented as $\star\star\star|\star\star|\star|\star\star\star||\star$.

In general, any valid solution to $\sum_{i=1}^{n} x_i = k$ is represented by k stars and $n-1$ separating bars. The number of solutions is then the number of permutations of the k stars and $n-1$ bars, which we know from above is,

$$\text{Number of combinations of } n \text{ objects with } k \text{ choices each} = \frac{(k+n-1)!}{k!(n-1)!} = \binom{k+n-1}{k}$$

so there are $\binom{6+10-1}{6} = 5005$ different combinations you could get from rolling 10 D6 dice.

2 Sample Spaces & Probabilities

A *probability space* consists of three elements:

- A *sample space*, Ω, the set of all possible outcomes;
- An *event space*, a family of sets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, with each set representing an *event*;
- A *probability measure*, $\mathbb{P} : \mathcal{F} \to [0,1]$, such that,
  - $\mathbb{P}(\Omega) = 1$;
  - $\mathbb{P}(\emptyset) = 0$;
  - If $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ are countably many disjoint events, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

[*] Since sets cannot contain duplicates of elements, we consider structures called "multisets" instead, and these permutations are technically called *multiset permutations.*

[†] We call these "k-multicombinations" or "k-multisubsets" of n.

[‡] Such equations are called *Diophantine* equations.

and a probability space is *discrete* if Ω is at most countably infinite. An event is *elementary* if it is a set of size 1. $\mathcal{P}(\Omega)$ is also sometimes written as 2^{Ω}.

A set of events are *mutually exclusive* if they are pairwise disjoint - every pair of events is disjoint. Given the definition of a probability measure, this is equivalent to the intersection of any pair of events having probability 0. The empty set is disjoint with every set, including itself.

The complement of an event A, $\Omega \setminus A$, is also written as A' or A^c if the sample space is clear. Note that an event and its complement are mutually exclusive and partition the sample space.

Example. A D6 dice is rolled and a coin is thrown in an experiment. The sample space, Ω, is then $\{1,2,3,4,5,6\} \times \{H,T\}$ where $\times$ is the Cartesian product. $|\Omega| = 12$.

The event space is $\mathcal{P}(\Omega)$, and contains $2^{12} = 4096$ possible events. For example, $\{(1,H),(2,H),(3,H)\} = \{1,2,3\} \times \{H\}$ is the event of the die rolling a number less than or equal to 4 *and* the coin landing on heads.

2.1 Algebra of Sets

> The algebra of sets and boolean/logic statements are isomorphic algebraic structures.
>
> You can transform equations about sets into boolean equations or logic statements by swapping,
>
> $$\begin{array}{ccccc} A \cap B & \Leftrightarrow & a \wedge b & \Leftrightarrow & A \,\text{and}\, B \\ A \cup B & \Leftrightarrow & a \vee b & \Leftrightarrow & A \,\text{or}\, B \\ A^c & \Leftrightarrow & \neg a & \Leftrightarrow & \text{not}\, A \\ \emptyset & \Leftrightarrow & \bot & \Leftrightarrow & 0 \\ U & \Leftrightarrow & \top & \Leftrightarrow & 1 \end{array}$$
>
> which might be helpful for computer scientists or logicians.
>
> For those interested in abstract algebra, these are all complemented distributive lattice structures.

The binary operations of set union, $\cup$, and intersection, $\cap$ are in many ways analogous to the binary operations of addition and multiplication.

- Commutativity;
  - $A \cup B = B \cup A$
  - $A \cap B = B \cap A$
- Associativity;
  - $(A \cup B) \cup C = A \cup (B \cup C)$
  - $(A \cap B) \cap C = A \cap (B \cap C)$
- Distributive property;
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

However, unlike addition and multiplication, union and intersection distribute in both directions.

Two additional properties involve the set containing nothing, the *empty set*, $\emptyset$; and the set containing everything of interest, the *universe set*, U.

- Identity;
  - $A \cup \emptyset = A$

  - $A \cap U = A$
- Complement;
  - $A \cup A^C = U$
  - $A \cap A^C = \emptyset$

so $\emptyset$ and U are the identity elements for union and intersection, respectively. In probability, the universe set is often Ω.

- Idempotency;
  - $A \cup A = A$
  - $A \cap A = A$
- Domination;
  - $A \cup U = U$
  - $A \cap \emptyset = \emptyset$
- Absorption;
  - $A \cup (A \cap B) = A$
  - $A \cap (A \cup B) = A$
- De Morgan's Laws;
  - $(A \cup B)^C = A^C \cap B^C$
  - $(A \cap B)^C = A^C \cup B^C$
- Involution and Complement Laws;
  - $\emptyset^C = U$
  - $U^C = \emptyset$
  - $(A^C)^C = A$

You might notice that all the identities above are given in pairs that can be transformed into each other by interchanging $\cap$ and $\cup$, and $\emptyset$ and U.

These are examples of an extremely powerful property of boolean algebras - the *principle of duality*, which asserts that the the dual of a true statement obtained by interchanging unions/intersections, universes/empty sets and reversing inclusions (for computer scientists, this is the same as reversing a Hasse diagram to get another poset) is also true (note that the involution law is self-dual).

Duality is a concept with uses in a much broader range of applications, particularly in order and category theory.

2.2 Inclusion-Exclusion Principle

Let A and B be sets. Then,

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Proof.

$$\begin{aligned} |A \cup B| &= |A \cup (B \setminus A)| \\ &= |A| + |B \setminus A| \qquad (1) \\ |B| &= |(B \setminus A) \cup (A \cap B)| \end{aligned}$$

$$= |B \setminus A| + |A \cap B| \tag{2}$$

Combining (1) and (2) gives the result. ■

Let A, B and C be sets. Then,

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |C \cap A| + |A \cap B \cap C|$$

In general, to find the cardinality of the union of n sets, we include the cardinality of the sets, exclude the cardinalities of the pairwise intersections, include the cardinalities of the 3-wise intersections, exclude 4-wise, and continue up to n.

Symbolically,

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\emptyset \neq J \subseteq \{1,\cdots,n\}} (-1)^{|J|+1} \left|\bigcap_{j \in J} A_j\right|$$

If all sets are exchanged for events and cardinalities replaced with probability measures, all of the above equations still hold, i.e., $|A \cup B| = |A| + |B| - |A \cap B| \Leftrightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Example. Each square in a 3×3 grid is either shaded or unshaded with equal probability. What is the probability that a 2×2 square is shaded in?

Let a square of dimensions $n \times n$ be denoted S_n.

There are $2^9 = 512$ possible ways to colour the entire S_3 square.

When one S_2 square is shaded in, there are $2^5 = 32$ ways to shade in the remaining 5 squares, and 4 possible positions to place S_2 within S_3, so there are $32 \times 4 = 128$ possible ways to shade in S_3 with 1 S_2 shaded.

If there are 2 S_2 shaded, positioned in opposite corners, there are $2^2 = 4$ ways to shade in the remaining 2 squares, and 2 ways to place the S_2 squares. If the S_2 squares are adjacent, there are 2^3 ways to shade in the remaining 3 squares, and 4 ways to place the S_2 squares. So, for 2 S_2 squares, there are $2^2 \times 2 + 2^3 \times 4 = 40$ ways to shade in the S_3.

3 shaded S_2 squares leaves only 1 square remaining, and there are 4 ways to place 3 S_2 squares, so there are $2^1 \times 4 = 8$ ways to shade 3 S_2 squares.

There is of course only one (1) way to shade 4 S_2 squares at once.

To find the union of these sets, we use the principle of inclusion-exclusion; The set of 2 S_2 squares is contained within certain shading patterns of 1 S_2 square, but we then need to add back in any combinations that would have 3 S_2 squares, then remove the combination which gives 4 S_2 squares again, as it is again contained within 3 S_2 squares. This gives $128 - 40 + 8 - 1 = 95$ ways to have at least 1 S_2 square.

It follows that the probability of having at least 1 S_2 square shaded is $\frac{95}{512}$.

Example. A *derangement* is a permutation with no fixed points: no element appears in its original position.

How many derangements are there of n objects?

For some fixed n and all $1 \leq k \leq n$, let S_k be the set of permutations of n objects. The number of derangements is then the total number of permutations, minus the union of these sets, $n! - |\bigcup_{i=1}^{n} S_i|$

Any intersection of a collection of i of these sets then fixes i objects and contains $(n-i)!$ permutations, and there are $\binom{n}{i}$ such collections, so,

$$\begin{aligned}\left|\bigcup_{i=1}^{n} S_i\right| &= \sum_{\emptyset \neq J \subseteq \{1,\cdots,n\}} (-1)^{|J|+1} \left|\bigcap_{j\in J} S_j\right| \\ &= \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} (n-i)! \\ &= \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{i!(n-i)!} (n-i)! \\ &= n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!}\end{aligned}$$

So the number of derangements is $n! - n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!} = n! \sum_{i=1}^{n} \frac{(-1)^{i}}{i!}$[*].

3 Conditional Probability

Let A be an event, and let B be an event with non-zero probability. The probability of A occurring, given that B has occurred, is written as $\mathbb{P}(A|B)$, and can be calculated with,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

It can be seen as the probability of A occurring within a new sample space over B.

3.1 Independence

Two events, A and B, are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. A finite set of events is *pairwise independent* if every pair of events in the set is independent. A finite set of events is *mutually independent* if every event is independent from every other set and every intersection of every other event.

3.2 Law of Total Probability

Let S be a set. If $\{A_i\}_{i=1}^{\infty}$ are countably many disjoint non-empty sets such that $\bigcap_{i=1}^{\infty} A_i = S$, then $\{A_i\}_{i=1}^{\infty}$ are said to *partition* S.

If A is an event that can be written as a countable partition, $A = \{B_i\}_{i=1}^{\infty}$, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i)$$

or alternatively,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

[*] You'll notice that this is the taylor polynomial for e, so another expression for the number of derangements is $\left[\frac{n!}{e}\right]$, where $[x]$ is the nearest integer to x.

3.3 Bayes' Theorem

For any events, A and B,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

If A and B are independent, this reduces to $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Extended form: Let $\{A_i\}_{i=0}^{n}$ partition the sample space. Then,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=0}^{n}\mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

Proof. By the definition of conditional probabilities,

$$\begin{aligned}
\mathbb{P}(A_i|B) &= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(A_i)} \\
&= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(A_i)} \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \mathbb{P}(B|A_i) \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=0}^{n}\mathbb{P}(B|A_i)\mathbb{P}(A_i)}
\end{aligned}$$

■

Example. Three Prisoners Problem

Three prisoners, A, B, and C are in separate cells, supervised by a warden. Two of them have been sentenced to death and will be executed the following morning, but none of the prisoners know who is to be spared.

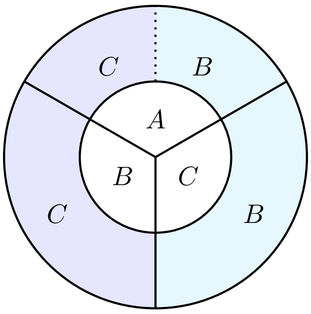
Prisoner A asks the warden what will happen tomorrow. The warden tells A that they won't say anything about A, nor anything about who will live. However, they do say that C is one of the prisoners to be executed.

Prisoner A is pleased as they believe that their own probability of surviving has gone up from 1 in 3, to 1 in 2, as it is now only between A and B who survives.

A secretly tells B the good news, who then reasons that A's chance of surviving is unchanged, while their own chances of survival has gone up to $\frac{2}{3}$.

Which prisoner is correct?

In all cases, the warden will not tell A anything about A's fate. In the case that B is to live, the warden will also not say anything about who will live, so the warden can only say that C will be executed. In the case that C is to live, the warden will not say anything about who will live, so the warden can only say that B will be executed. However, if A is to live, then the warden has a choice. The warden can either say that B or C will be executed.



In the diagram above, the inner ring indicates who lives, with the outer ring indicating who the warden says will be executed. So, the warden says that B will be executed 50% of the time, and that C will be executed the other 50% of the time, corresponding to the right and left side of the circle, respectively. In both cases, A only lives $\frac{1}{3}$ of the time.

We can also write this using Bayes' theorem. Let A, B and C be the events that the corresponding prisoner is not executed, and let X be the event that the warden tells A that C is to be executed.

We see then that the probability that A survives is,

$$\begin{aligned}\mathbb{P}(A|X) &= \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X)} \\ &= \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X|A)\mathbb{P}(A) + \mathbb{P}(X|B)\mathbb{P}(B) + \mathbb{P}(X|C)\mathbb{P}(C)} \\ &= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}} \\ &= \frac{1}{3}\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbb{P}(B|X) &= \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X)} \\ &= \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X|A)\mathbb{P}(A) + \mathbb{P}(X|B)\mathbb{P}(B) + \mathbb{P}(X|C)\mathbb{P}(C)} \\ &= \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}} \\ &= \frac{2}{3}\end{aligned}$$

The denominators are the same in both cases, the difference stemming from the fact that the warden will always state that C is to be executed if B is to live, so $P(X|C) = 1$, but will only do so 50% of the time when A is to live, so $P(X|A) = \frac{1}{2}$.

3.4 Expected Value

The *expected value* of a random variable, X, is the weighted average of all possible values of X.

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p_i$$

where x_i are the possible values of X, and p_i are their corresponding probabilities of occurrence. The expected value is also sometimes denoted μ, particularly when working with normal distributions.

Expectation is linear, so,

$$\mathbb{E}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1} c_i \mathbb{E}(X_i)$$

3.5 Variance

Variance is a measure of dispersion, representing how far a set of numbers is from their mean. Variance is the square of the standard deviation. It is often denoted as $\text{Var}(X)$ or σ^2, and can be calculated from the expected value; $\text{Var}(X) = \mathbb{E}(X^2) - E(X)^2$.

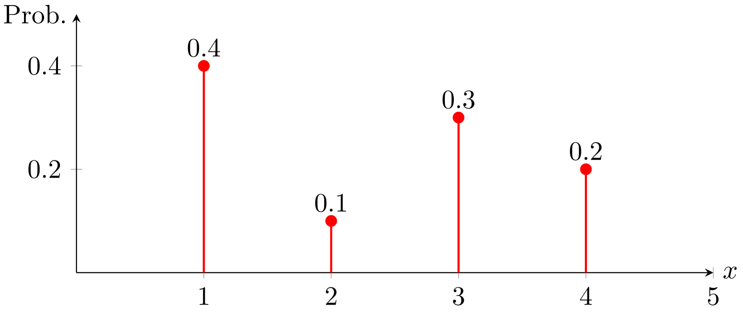
While you will not need to often calculate it by hand, the variance is an important summary statistic, and is frequently used as a parameter in various probability distributions.

4 Probability Distributions

A *random variable* is a quantity whose value depends on the outcome of a random event. Random variables are written in uppercase, with lowercase used to denote specific values the random variables can take.

A *probability mass function* or *discrete density function* is a function that gives the probability that a discrete random variable is equal to some given value. We write $\mathbb{P}(X = x)$ to denote the probability that the random variable X takes the particular value x. Then, the probability mass function, $p_X : \mathbb{R} \to [0,1]$ would be $p_X(x) = \mathbb{P}(X = x)$.

A probability mass function can be drawn on a plot,



Note that all the heights sum to 1, and that the probability mass function is zero at all the real numbers between valid outcomes.

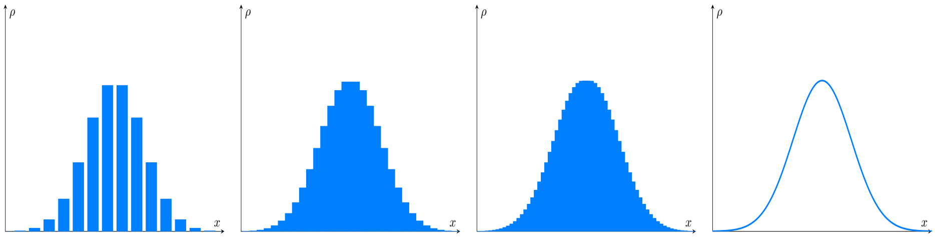
A defining feature of continuous probability distributions is that the probability for a random variable to take any specific value is 0, as there are infinitely many possible values the variable could take.

If every uncountably many particular values in some region all have non-zero probability, then the sum of all those probabilities goes to infinity. If all the probabilities are zero, then the whole sum is again zero, giving no meaningful information about the distribution.

To resolve this problem, we focus not on individual values, but ranges of values that the variable can take instead. For example, for a random variable that takes real values over [0,1], we might divide the interval into 10 parts and ask what the probability of falling into each region is.

When plotting this, rather than using the height of each bar to represent a probability, we use the areas.

As we make the intervals finer and finer, the smaller probability of falling into each interval is captured by the thinner width of each bars, so the height of the bars stay roughly the same as the intervals get smaller. Note that this wouldn't be the case if we used the heights to represent probability - in that case, every bar would shrink, and eventually reach 0 height in the limit. However, using areas, this process approaches a smooth curve.



So although each individual probability goes to zero, the overall shape of the distribution is preserved. With probability being proportional to the area of the bars, the vertical axis needs different units. Calling the width Δx, the height represents some kind of probability per unit in the x direction: $\frac{\text{Prob.}}{\Delta x}$, and we call this a probability density.

The curve that this process approaches is the *probability density function*, and is the continuous analogue to the probability mass function. To get the probability that a random variable lands within an interval, $[a,b]$, we integrate the probability density function between a and b to find the area under the curve. That is to say, if X is a random variable distributed according to the probability density function, f, then $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx$. Note that if $a = b$, then the integral returns 0, and integrating the probability density function over all of space returns 1.

4.1 Finite Discrete Uniform Probability Measures

A finite discrete probability measure is *uniform* if every pair of elementary events are equally likely.

In a discrete uniform probability measure, the probability of an event, A, happening is,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

The plot of the probability mass function of a discrete uniform probability measure is a line of points, all with the same height.

Example. A fair D6 dice is rolled and a fair coin is thrown in an experiment. The event, $A = \{(1,H),(2,H),(3,H)\}$ has probability $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{3}{12} = \frac{1}{4}$.

Example. At a dinner party, 6 guests are seated around a table. Three pairs of hats are randomly distributed to the guests. What is the probability that every guest is sitting next to another guest with the same hat?

If we had a valid arrangement of hats, there are 6 ways to rotate it around the table, and since there are 3 sets of hats, there are 2 possible cycles (i.e., 1-2-3 and 1-3-2 are the two distinct cycles of 3 elements). Each hat can also be swapped around within its own pair, so there are $6 \times 2 \times 2^3$ valid arrangements, and 6! total arrangements, so the probability is $\frac{6\times2\times2^3}{6!} = \frac{2}{15}$.

Most of the questions pertaining to these distributions will effectively reduce down to basic combinatorics and trying to find the size of an event.

You'll notice that we use the cardinality of Ω in the definition of probability for discrete uniform probability measures, and that this doesn't really make sense for infinite discrete sample spaces.

For more general sample spaces, we use the measure of those sets. As will be shown later, the measure of a countable set is zero, so this quotient is still not useful for us.

We could still attempt to define a distribution that assigns the same probability to each elementary event. Let X be a discrete random variable that takes values in a countably infinite set Ω, and suppose such a uniform distribution exists, so there exists some non-negative probability, p such that $\mathbb{P}(X = n) = p$ for all $n \in \Omega$. Since all the n are elementary and Ω is countably infinite, they are disjoint, so we can use the additive property of probability measures.

$$\begin{aligned} 1 &= \mathbb{P}(\Omega) \\ &= \mathbb{P}(X \in \Omega) \\ &= \sum_{n \in \Omega} P(X = n) \\ &= \sum_{n \in \Omega} p \end{aligned}$$

If $p = 0$, then $\sum_{n \in \Omega} p = 0$. If $p > 0$, then $\sum_{n \in \Omega} p = \infty$. In either case, we have a contradiction.

It turns out that there is no way to define a uniform distribution on a countably infinite set. So when someone says "discrete uniform distribution", they mean a finite discrete uniform distribution.

4.2 Continuous Uniform Probability Measures

If Ω is uncountably infinite, then the quotient using cardinalities is not well-defined. We instead use the *measure* of the sets involved. For $\Omega \subseteq \mathbb{R}$, we use the lengths of the sets; for $\Omega \subseteq \mathbb{R}^2$, the areas; and for $\Omega \subseteq \mathbb{R}^3$, the volumes.

Similar to countably infinite sets, not all subsets of $\mathbb{R}^n$ can be assigned a valid probability measure. This generally isn't a problem though, as all the sets we use in this module are measurable. For further reading, search *Vitali sets*, *Hausdorff paradox*, and the *Banach-Tarski theorem.*

Due to additivity, a lot of continuous probability problems reduce down to questions about geometry.

Example. Darts are thrown uniformly at a square with sides 2 units long. A unit circle dartboard is set in the square. What is the probability that any given dart will hit the dartboard?

The area of the dartboard is π, and the area of the square is 4, so the probability that the dart hits is $\frac{\pi}{4}$.

Example. A coin is thrown and lands uniformly on an infinitely large table covered with a regular square grid.

1. If the coin has unit diameter, what is the probability that the the coin does not land on any lines if the squares have side lengths,
   (a) 1?
   (b) 2?
   (c) $n \geq 1$?

Suppose the coin now has radius r, and the squares have side lengths of $L > 2r$.

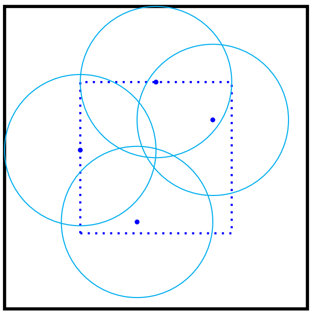
What is the probability that the coin intersects,

1. at most 1 line segment?
2. exactly 2 line segments?
3. exactly 3 line segments?

For all of these questions, we can just consider a single square, or a single intersection point, as the tiling is regular.

If the squares have side lengths 1, and the coin has unit dimeter, then the coin must land on the exact centre of the square, which is a point of zero area, so the probability that the coin does not land any any lines is 0.

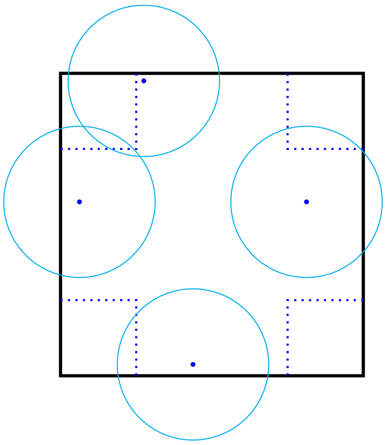
If the squares have side lengths 2, then the centre of the circle get to within radius of the outer square. This traces out another smaller square as our allowable area,



As long as the centre of the circle lands within the smaller square, the circle will not intersect any sides. The side length of the smaller square is the side length of the larger square, minus twice the radius (or minus the diameter) of the circle, so in this case, we have $2-1=1$, so the smaller square has unit length sides, and therefore has unit area. The larger square has area 4, so the probability that the circle does not intersect any lines is $\frac{1}{4}$.

Similarly, if the larger square has side lengths $n \geq 1$, then the smaller square will have side lengths $n-1$, so the probability is $\frac{(n-1)^2}{n^2}$. Or more generally, for a circle of radius r and a square of side length $L > 2r$, $\frac{(L-2r)^2}{L^2}$.

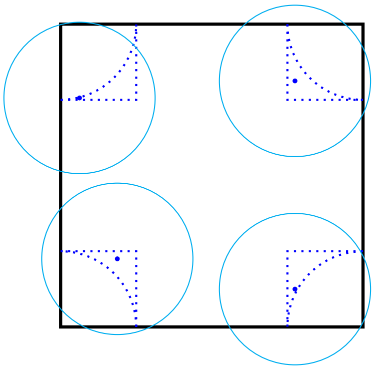
For the probability that the coin intersects at most one line segment, we look at how close the coin can get to the intersection points.



The smaller squares have side lengths equal to the radius of the coin, so the total area is $4r^2$ and the probability of intersecting at most 1 line segment is $\frac{4r^2}{L^2}$.

For intersecting exactly 1 line segment, we just subtract the probability of intersecting zero lines away, so we have $\frac{4r^2-(L-2r)^2}{L^2} = \frac{4r}{L} - 1$.

For intersecting exactly 2 line segments, the centre of the circle has to land in the complement of the region found before, but the circle cannot enclose the corner itself, so the centre cannot be within a radius distance from the corner. The required region is then,



which has area $4r^2 - \pi r^2 = (4-\pi)r^2$, so the probability that the coin intersects exactly two line segments is $\frac{(4-\pi)r^2}{L^2}$.

Given the setup of the grid, it is impossible for the coin to intersect exactly 3 lines, as intersecting more than 2 requires the circle to enclose a corner, which necessarily causes the coin to intersect 4 lines. It follows that the probability of the coin intersecting exactly 3 line segments is 0.

Example. Measure Theory

If a real number in (0,1) is picked uniformly at random, what is the probability that the real number is rational?

The real numbers are uncountably infinite, so this is a continuous probability question, so we need to find the "length" of the rationals over (0,1).

The standard way to do this, is to cover the regions of interest with open intervals, then to add up the lengths of the intervals. One obvious way to do this is to just use (0,1), but we want to do this with the smallest total length possible. Can we do better than a length of 1?

We know that the rationals are a countable set, so there is a bijection between $\mathbb{Q}$ and $\mathbb{N}$. Cantor famously created one such bijection with his zig-zag argument, but we only need the rationals between 0 and 1.

There are many ways to do this, but one organised way is to start with $\frac{1}{2}$, then move onto $\frac{1}{3}$, $\frac{2}{3}$, then $\frac{1}{4}$ and $\frac{3}{4}$, then continuing with the reduced fractions with denominator 5, then 6, and so on. Doing this will list every rational in (0,1) exactly once, creating a bijection between the rationals and the naturals (the indexing set of the sequence).

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \frac{1}{7}, \frac{2}{7}, \ldots$$

Now, we can assign an interval to each rational. Let $\epsilon > 0$, and pick any convergent series, say $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$. Now, $\sum_{i=1}^{\infty} \frac{\epsilon}{2^i} = \epsilon$. Now, if we use the terms of this series as the lengths of intevals covering each rational, we can cover all the rationals in (0,1) using a total length of ϵ, so the length can be arbitrarily small.

We say that the rationals have a *Lebesgue measure* of 0. Doing the same process with the real numbers in (0,1), we find that this interval has a Lebesgue measure of 1. How this works in detail is somewhat involved, requiring more complicated topology techniques, and this question is just meant as a very brief introduction to measure theory, so the proof is omitted.

So, using the measures of the sets instead of cardinalities, we find that the probability that a real number randomly selected from (0,1) is rational is $\frac{0}{1} = 0$. This may be counterintuitive, given that the rationals are dense in the reals, and that it is certainly possible to select a rational from (0,1), but this kind of thing is very common in continuous probability.

We say that an event is said to happen *almost surely* if the set of possible exceptions has measure zero (and *almost never* is defined similarly). Note that this does not preclude the set of exceptions from being non-empty: rational numbers clearly exist between 0 and 1, but this set has measure 0, so we say that a rational is selected almost never, or equivalently, that an irrational is selected almost surely.

4.3 Binomial Distributions

A *Bernoulli trial* is an experiment with exactly two possible outcomes, often labelled "success" and "failure", with the probabilities being the same every time the experiment is conducted.

If we define the random variable X to represent the number of successes in a fixed number of identical Bernoulli trials, then X is distributed *binomially*, and we write $X \sim B(n,p)$, where $n \in \mathbb{N}$ is the number of trials and $p \in [0,1]$ is the probability of success. Then, the following are equivalent notation for the probability mass function for the binomial distribution:

$$p_X(k) = f(k,n,p) = \mathbb{P}(k; n,p) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

A binomial distribution is a valid model for a random variable X if there are two possible outcomes, the number of trials and probability of success is fixed, and the trials are all independent from each other.

The expected value and variance of a random variable distributed binomially with parameters n,p, $X \sim B(n,p)$ is $\mathbb{E}(X) = np$ and $\text{Var}(X) = np(1-p)$ (it is helpful to memorise these values, as they are used a lot, particularly in various approximations to the binomial distribution).

Situations where the binomial distribution could be used:

- Number of tails obtained on a (possibly biased) coin over 10 throws;
- Number of votes obtained by a candidate in a plurality voting election;
- Number of side effects of new medication experienced by 100 patients.

Situations where the binomial distribution is not valid:

- The colour of cards randomly removed from a deck without replacement (not independent - this is the *hypergeometric distribution*)
- The suit of cards randomly removed from a deck with replacement (not binary - this is the *multinomial distribution*)
- Number of times a die is rolled until a 6 is obtained (number of trials is not fixed - this is the *negative binomial* or *geometric distribution*);

4.4 Poisson Distribution

The series definition of the exponential function is,

$$e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Multiplying both sides by e^{-x}, we have,

$$1 = \frac{x^0 e^{-x}}{0!} + \frac{x^1 e^{-x}}{1!} + \frac{x^2 e^{-x}}{2!} + \frac{x^3 e^{-x}}{3!} + \cdots$$

The right hand side sums to 1, so we can use these values as probabilities to define a probability distribution.

Using the probability mass function, $f(k;\lambda) = \mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, this distribution is the *Poisson distribution*, taking a single parameter $\lambda > 0$.

A Poisson distribution is a valid model for a random variable X if events occur independently, singly in space or time, and at a constant average rate such that the mean number of occurrences over an interval is proportional to the length of the interval.

The expected value and variance of a random variable in a poisson distribution with parameter λ, $X \sim \text{Pois}(\lambda)$ is $\mathbb{E}(X) = \lambda = \text{Var}(X)$.

Situations where the Poisson distribution could be used:

- Number of alpha particles emitted by a radioactive source over a given time period;
- Number of patients arriving at an emergency room at a given hour of the day;
- Number of faulty parts manufactured at a factory in a day.

Situations where the Poission distribution is not valid:

- Number of students arriving at a lecture hall (not constant rate, and not independent);
- Number of earthquakes in a country per year (not independent);
- Number of articles published by tenured professors (to be tenured, a professor must have published at least once, so Poisson distribution is not a good fit due to the 0 output.)

4.5 Normal Distribution

The *normal* or *Gaussian distribution* has two parameters: μ, the population mean, and σ^2, the population variance. The distribution is symmetric about the mean, with mean=median=mode.

The probability density function of the normal distribution is,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The expected value and variance of a random variable distributed normally with parameters μ,σ^2, $X \sim N(\mu,\sigma^2)$ is $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

If some data is coded using the formula, $y = \frac{x-a}{b}$, then the mean and standard deviation of the coded data is given given by $\mu_y = \frac{\mu_x - a}{b}$ and $\sigma_y = \frac{\sigma_x}{b}$ (this is true of all random variables, not just normally distributed ones).

The *standard normal distribution* has mean 0 and standard devation 1. If $X \sim N(\mu,\sigma^2)$, then we can *standardise* X with the coding $Z = \frac{X-\mu}{\sigma}$. The resulting z-values are distributed according to the standard normal distribution, $Z \sim N(0,1)$. This works because every normal distribution is a version of the standard normal with the domain stretched by a factor of σ, and then translated by μ.

The probability density function of the standard normal distribution is often denoted $\phi(x)$, and is given by,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Instead of integrating the normal probability density function directly, we often standardise the given data and write the integral in terms of the standard normal density function. If $X \sim N(\mu,\sigma^2)$, then,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx$$

$$= \frac{1}{\sigma} \int_a^b \phi\left(\frac{x-\mu}{\sigma}\right) dx$$

5 Law of Large Numbers

Given a sequence of independent and identically distributed random variables $\{X_i\}_{i=1}^n$ with finite expected value $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n) = \mu < \infty$, define a new random variable $\overline{X}_n = \sum_{i=1}^n \frac{X_i}{n}$. This variable is the *sample mean.*

As expectation is linear, $\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{n\mu}{n} = \mu$, so the sample mean has the same mean as each of the individual variables, as we would expect.

Within statistics, there are various notions of convergence of random variables. These concepts are also called *stochastic convergence* in other areas of maths, and they formalise the idea that a sequence of random events can sometimes settle into some kind of stable behaviour with sufficiently large sample sizes.

We say that a sequence of random variables, $\{X_n\}$, *converges in distribution* or *converges weakly* towards a random variable X, if $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ at which F is continuous, and where F_n and F are the cumulative distribution functions of X_n and X, respectively. This means that we increasingly expect that the next outcome in a sequence of random experiments is modelled better and better by the distribution of X. We use the notation $X_n \xrightarrow{\mathcal{D}} X$ or $X_n \rightsquigarrow X$ to represent this kind of convergence. This kind of convergence is used in the *weak law of large numbers* .

We say that a sequence of random variables, $\{X_n\}$, *converge in probability*, towards a random variable, X, if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$. This means that the probability of an “unusual” outcome becomes smaller and smaller as the sequence progress. We use the notation $X_n \xrightarrow{P} X$ or $\operatorname{plim}(X_n) = X$ to represent this kind of convergence. This kind of convergence is used in the *central limit theorem* .

We say that a sequence of random variables, $\{X_n\}$ converges *almost surely*, *almost everywhere* or *strongly*, towards a random variable X, if $\mathbb{P}(\lim_{n\to\infty} X_n = x) = 1$. This type of convergence is very similar to pointwise convergence from analysis. This form of convergence means that the events for which X_n do not converge to X have probability 0 (the same as randomly selecting a rational from a reals interval; possible, but probability 0 - it has Lebesgue measure 0). We use the notation $X_n \xrightarrow{a.s.} X$ to represent this kind of convergence. This kind of convergence is used in the *strong law of large numbers* .

There is another stronger form of convergence analogous to uniform convergence from analysis called *sure convergence* or, but is rarely used in statistics as the only difference between sure and almost sure convergence in probability is in sets with Lebesgue measure 0.

The forms of convergence above are given in order of strength, with convergence in distribution being the weakest form. There are various other stronger forms of stochastic convergence not covered here.

There is a weak and a strong version of the law of large numbers. Both state that the sample average converges to the expected value;

$$\overline{X}_n \to \mu \text{ as } n \to \infty$$

The difference between the weak and strong versions is in the mode of convergence.

5.1 Weak Law of Large Numbers

The *weak law of large numbers* states that the sample mean converges in probability towards the expected value as the sample size increases;

$$\overline{X}_n \xrightarrow{P} \mu \text{ as } n \to \infty$$

That is, for any given error, $\epsilon > 0$, there exists a sufficiently large sample size that will ensure that the average of the observations, $\overline{X_n}$ will almost always be within ϵ of the expected value, μ, which is the definition of a limit.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \epsilon\right) = 1$$

Equivalently, $\overline{X}_n$ will almost never be further than ϵ of the expected value, μ.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) = 0$$

5.1.1 Bernoulli's Weak Law of Large Numbers

Suppose $X \sim B(n{,}p)$. Then, the expected value is $\mu = np$, so the weak law of large numbers says

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - np\right| > \epsilon\right) = 0$$

However, for binary random variables, such as in the binomial distribution, we can also look at the mean of the proportion of successes, and not just the mean of the number of successes. Doing so, we have,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_n}{n} - p\right| > \epsilon\right) = 0$$

5.2 Strong Law of Large Numbers

The *strong law of large numbers* states that the sample mean converges almost surely to the expected value;

$$\overline{X}_n \overset{a.s.}{\to} \mu \text{ as } n \to \infty$$

That is, $\mathbb{P}(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$.

The weak law simply states that for some large n, $\bar{X}_n$ is likely to be close to μ, but does not preclude the possibility that $|\bar{X}_n - \mu| > \epsilon$ happens infinitely many times (though, likely only at increasingly infrequent intervals for larger and larger n).

The strong law states that this almost surely does not occur (i.e., has Lebesgue measure 1). Note that this does not imply that for any $\epsilon > 0$, there exists N such that $|\bar{X}_n - \mu| < \epsilon$ holds for all $n > N$, since converging almost surely is not uniform convergence.

5.3 Central Limit Theorem

The *classical central limit theorem* states that if $\{X_i\}_{i=1}^n$ is an independent and identically distributed sequence of random samples drawn from a population with mean μ and variance σ^2, then the sample mean $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ converges in distribution to $N\left(\mu, \frac{\sigma^2}{n}\right)$, regardless of the distribution of the population.

6 Approximating the Binomial

With very large number of Bernoulli trials, it quickly becomes intractable to calculate factorials. For large n, we often approximate the binomial distribution with other, computationally easier distributions.

6.1 Poisson Limit Theorem

Let p_n be a sequence of real numbers in [0,1] such that the sequence np_n converges to some limit $\lambda < \infty$. Then,

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

That is to say, if $X \sim B(n{,}p)$, n is large, and p is small, then X is approximately $\sim \text{Pois}(np)$ (recall that np is the expected value of X).

6.2 De Moivre–Laplace Theorem

The *De Moivre–Laplace theorem* is a special case of the central limit theorem. If $X \sim B(n,p)$, then, as $n \to \infty$, X converges in distribution to $N(\mu,\sigma^2)$, where μ is the expected value of X, which is np, and σ^2 is the variance of X, which is $np(1-p)$.

In other words, if $X \sim B(n,p)$, then for large n, X is approximately $\sim N\left(np, \sqrt{np(1-p)}^2\right)$.

Because the normal distribution is continuous, while the binomial is discrete, you need to apply a *continuity correction* when calculating probabilities. If $X \sim N(n,p)$ and $Y \sim N\left(np, \sqrt{np(1-p)}^2\right)$, then,

- $\mathbb{P}(X = a) \approx \mathbb{P}(a - 0.5 < Y < a + 0.5)$;
- $\mathbb{P}(X > a) \approx \mathbb{P}(Y > a + 0.5)$;
- $\mathbb{P}(X \geq a) \approx \mathbb{P}(Y > a - 0.5)$;
- $\mathbb{P}(X < a) \approx \mathbb{P}(Y > a - 0.5)$;
- $\mathbb{P}(X \leq a) \approx \mathbb{P}(Y < a + 0.5)$;

7 Closing Remarks & Condensed Summary

Apart from some new terminology, this module is almost entirely A-Level content, other otherwise relatively elementary combinatorics.

A lot of extra information is included in this document, but learning it does make a lot of the syllabus easier to remember - in particular, expected values, data coding and the central limit theorem.

$$\text{Total permutations} = \frac{\text{(Number of items we can choose from)!}}{\text{(Number of items we }\underline{\text{can't}}\text{ choose)!}}$$

If we have n items and k spaces, these are called k-permutations of and are denoted with $P(n,k)$ or P^n_k, and,

$$P^n_k = \frac{n!}{(n-k)!}$$

$$\text{Total combinations} = \frac{\text{(Number of items we can choose from)!}}{\text{(Number of items we }\underline{\text{can}}\text{ choose)!} \times \text{(Number of items we }\underline{\text{can't}}\text{ choose)!}}$$

We call these k-combinations of n objects, denoted $C(n,k)$, C^n_k, or $\binom{n}{k}$.

$$C^n_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\text{Total permutations with repeated objects} = \frac{\text{(Total number of objects)!}}{\text{(Group 1)!} \times \text{(Group 2)!} \times \text{(Group 3)!} \cdots}$$

$$\text{Number of combinations of } n \text{ objects with } k \text{ choices each} = \frac{(k+n-1)!}{k!(n-1)!} = \binom{k+n-1}{k}$$

A *probability space* consists of three elements:

- A *sample space*, Ω, the set of all possible outcomes;
- An *event space*, a family of sets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, with each set representing an *event*;
- A *probability measure*, $\mathbb{P} : \mathcal{F} \to [0,1]$, such that,
  - $\mathbb{P}(\Omega) = 1$;
  - $\mathbb{P}(\emptyset) = 0$;
  - If $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ are countably many disjoint events, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

and a probability space is *discrete* if Ω is at most countably infinite. An event is *elementary* if it is a set of size 1. $\mathcal{P}(\Omega)$ is also sometimes written as 2^{Ω}.

A set of events are *mutually exclusive* if they are pairwise disjoint - every pair of events is disjoint. Given the definition of a probability measure, this is equivalent to the intersection of any pair of events having probability 0. The empty set is disjoint with every set, including itself.

The complement of an event A, $\Omega \setminus A$, is also written as A' or A^c if the sample space is clear. Note that an event and its complement are mutually exclusive and partition the sample space.

The algebra of sets and boolean/logic statements are isomorphic algebraic structures. You can transform

> equations about sets into boolean equations or logic statements by swapping,
>
> $$\begin{array}{ccccc} A \cap B & \Leftrightarrow & a \wedge b & \Leftrightarrow & A \,\text{and}\, B \\ A \cup B & \Leftrightarrow & a \vee b & \Leftrightarrow & A \,\text{or}\, B \\ A^c & \Leftrightarrow & \neg a & \Leftrightarrow & \text{not}\, A \\ \emptyset & \Leftrightarrow & \bot & \Leftrightarrow & 0 \\ U & \Leftrightarrow & \top & \Leftrightarrow & 1 \end{array}$$
>
> which might be helpful for computer scientists or logicians..

Inclusion-Exclusion Principle: Let A and B be sets. Then,

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Proof.

$$\begin{aligned} |A \cup B| &= |A \cup (B \setminus A)| \\ &= |A| + |B \setminus A| \qquad (1) \\ |B| &= |(B \setminus A) \cup (A \cap B)| \\ &= |B \setminus A| + |A \cap B| \qquad (2) \end{aligned}$$

Combining (1) and (2) gives the result. ■

Let A, B and C be sets. Then,

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |C \cap A| + |A \cap B \cap C|$$

In general, to find the cardinality of the union of n sets, we include the cardinality of the sets, exclude the cardinalities of the pairwise intersections, include the cardinalities of the 3-wise intersections, exclude 4-wise, and continue up to n.

Symbolically,

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\emptyset \neq S \subseteq \{1,\cdots,n\}} (-1)^{|J|+1} \left|\bigcap_{s \in S} A_s\right|$$

If all sets are exchanged for events and cardinalities replaced with probability measures, all of the above equations still hold, i.e., $|A \cup B| = |A| + |B| - |A \cap B| \Leftrightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

A *derangement* is a permutation with no fixed points: no element appears in its original position. There are $n! \sum_{i=1}^{n} \frac{(-1)^i}{i!}$ derangements of n objects.

Proof. For some fixed n and all $1 \leq k \leq n$, let S_k be the set of permutations of n objects. The number of derangements is then the total number of permutations, minus the union of these sets, $n! - |\bigcup_{i=1}^{n} S_i|$

Any intersection of a collection of i of these sets then fixes i objects and contains $(n-i)!$ permutations, and there are $\binom{n}{i}$ such collections, so,

$$\begin{aligned} \left|\bigcup_{i=1}^{n} S_i\right| &= \sum_{\emptyset \neq J \subseteq \{1,\cdots,n\}} (-1)^{|J|+1} \left|\bigcap_{j \in J} S_j\right| \\ &= \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} (n-i)! \\ &= \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{i!(n-i)!} (n-i)! \end{aligned}$$

$$= n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!}$$

So the number of derangements is $n! - n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!} = n! \sum_{i=1}^{n} \frac{(-1)^{i}}{i!}$. ■

Conditional Probability: Let A be an event, and let B be an event with non-zero probability. The probability of A occurring, given that B has occurred, is written as $\mathbb{P}(A|B)$, and can be calculated with,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

It can be seen as the probability of A occurring within a new sample space over B.

Two events, A and B, are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. A finite set of events is *pairwise independent* if every pair of events in the set is independent. A finite set of events is *mutually independent* if every event is independent from every other set and every intersection of every other event.

Let S be a set. If $\{A_i\}_{i=1}^{\infty}$ are countably many disjoint non-empty sets such that $\bigcap_{i=1}^{\infty} A_i = S$, then $\{A_i\}_{i=1}^{\infty}$ are said to *partition* S.

Law of Total Probability: If A is an event that can be written as a countable partition, $A = \{B_i\}_{i=1}^{\infty}$, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i)$$

or alternatively,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Bayes' Theorem: For any events, A and B,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

If A and B are independent, this reduces to $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Extended form: Let $\{A_i\}_{i=0}^{n}$ partition the sample space. Then,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=0}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

Proof. By the definition of conditional probabilities,

$$\begin{aligned}
\mathbb{P}(A_i|B) &= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(A_i)} \\
&= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(A_i)} \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \mathbb{P}(B|A_i) \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}
\end{aligned}$$

$$= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=0}^{n}\mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

■

The *expected value* of a random variable, X, is the weighted average of all possible values of X.

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p_i$$

where x_i are the possible values of X, and p_i are their corresponding probabilities of occurrence. The expected value is also sometimes denoted μ, particularly when working with normal distributions.

Expectation is linear, so,

$$\mathbb{E}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1} c_i \mathbb{E}(X_i)$$

Variance is a measure of dispersion, representing how far a set of numbers is from their mean. Variance is the square of the standard deviation. It is often denoted as $\text{Var}(X)$ or σ^2, and can be calculated from the expected value; $\text{Var}(X) = \mathbb{E}(X^2) - E(X)^2$.

While you will not need to often calculate it by hand, the variance is an important summary statistic, and is frequently used as a parameter in various probability distributions.

A *random variable* is a quantity whose value depends on the outcome of a random event. Random variables are written in uppercase, with lowercase used to denote specific values the random variables can take.

A *probability mass function* or *discrete density function* is a function that gives the probability that a discrete random variable is equal to some given value. We write $\mathbb{P}(X = x)$ to denote the probability that the random variable X takes the particular value x. Then, the probability mass function, $p_X : \mathbb{R} \to [0,1]$ would be $p_X(x) = \mathbb{P}(X = x)$.

A *probability density function* is the continuous analogue to the probability mass function. To get the probability that a random variable lands within an interval, $[a,b]$, we integrate the probability density function between a and b to find the area under the curve. That is to say, if X is a random variable distributed according to the probability density function, f, then $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx$. Note that if $a = b$, then the integral returns 0, and integrating the probability density function over all of space returns 1.

A finite discrete probability measure is *uniform* if every pair of elementary events are equally likely.

In a discrete uniform probability measure, the probability of an event, A, happening is,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

The plot of the probability mass function of a discrete uniform probability measure is a line of points, all with the same height.

If Ω is uncountably infinite, then the quotient using cardinalities is not well-defined. We instead use the *measure* of the sets involved. For $\Omega \subseteq \mathbb{R}$, we use the lengths of the sets; for $\Omega \subseteq \mathbb{R}^2$, the areas; and for $\Omega \subseteq \mathbb{R}^3$, the volumes.

A *Bernoulli trial* is an experiment with exactly two possible outcomes, often labelled “success” and “failure”, with the probabilities being the same every time the experiment is conducted.

If we define the random variable X to represent the number of successes in a fixed number of identical Bernoulli trials, then X is distributed *binomially*, and we write $X \sim B(n,p)$, where $n \in \mathbb{N}$ is the number of trials and $p \in [0,1]$ is the probability of success. Then, the following are equivalent notation for the probability mass function for the binomial distribution:

$$p_X(k) = f(k,n,p) = \mathbb{P}(k; n,p) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

A binomial distribution is a valid model for a random variable X if there are two possible outcomes, the number of trials and probability of success is fixed, and the trials are all independent from each other.

The expected value and variance of a random variable distributed binomially with parameters n,p, $X \sim B(n,p)$ is $\mathbb{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$ (it is helpful to memorise these values, as they are used a lot, particularly in various approximations to the binomial distribution).

Using the probability mass function, $f(k; \lambda) = \mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, this distribution is the *Poisson distribution*, taking a single parameter $\lambda > 0$.

A Poisson distribution is a valid model for a random variable X if events occur independently, singly in space or time, and at a constant average rate such that the mean number of occurrences over an interval is proportional to the length of the interval.

The expected value and variance of a random variable in a poisson distribution with parameter λ, $X \sim \mathrm{Pois}(\lambda)$ is $\mathbb{E}(X) = \lambda = \mathrm{Var}(X)$.

The *normal* or *Gaussian distribution* has two parameters: μ, the population mean, and σ^2, the population variance. The distribution is symmetric about the mean, with mean=median=mode.

The probability density function of the normal distribution is,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The expected value and variance of a random variable distributed normally with parameters μ,σ^2, $X \sim N(\mu,\sigma^2)$ is $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

If some data is coded using the formula, $y = \frac{x-a}{b}$, then the mean and standard deviation of the coded data is given given by $\mu_y = \frac{\mu_x - a}{b}$ and $\sigma_y = \frac{\sigma_x}{b}$ (this is true of all random variables, not just normally distributed ones).

The *standard normal distribution* has mean 0 and standard devation 1. If $X \sim N(\mu,\sigma^2)$, then we can *standardise* X with the coding $Z = \frac{X-\mu}{\sigma}$. The resulting z-values are distributed according to the standard normal distribution, $Z \sim N(0,1)$. This works because every normal distribution is a version of the standard normal with the domain stretched by a factor of σ, and then translated by μ.

The probability density function of the standard normal distribution is often denoted $\phi(x)$, and is given by,

$$\phi(x = \frac{1}{\sqrt{2\pi}} e^{-} \frac{x^2}{2}$$

Instead of integrating the normal probability density function, we often standardise the given data and write the integral in terms of the standard normal density function. If $X \sim N(\mu,\sigma^2)$, then,

$$\begin{aligned}\mathbb{P}(a \le X \le b) &= \int_a^b f(x)\,dx \\ &= \frac{1}{\sigma}\int_a^b \phi\left(\frac{x-\mu}{\sigma}\right)\,dx\end{aligned}$$

iven a sequence of independent and identically distributed random variables $\{X_i\}_{i=1}^n$ with finite expected value $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n) = \mu < \infty$, define a new random variable $\overline{X}_n = \sum_{i=1}^n \frac{X_i}{n}$. This variable is the *sample mean.*

As expectation is linear, $\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{n\mu}{n} = \mu$, so the sample mean has the same mean as each of the individual variables, as we would expect.

The *weak law of large numbers* states that the sample mean converges in probability towards the expected value as the sample size increases;

$$\overline{X}_n \xrightarrow{P} \mu \text{ as } n \to \infty$$

That is, for any given error, $\epsilon > 0$, there exists a sufficiently large sample size that will ensure that the average of the observations, $\overline{X_n}$ will almost always be within ϵ of the expected value, μ, which is the definition of a limit.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \epsilon\right) = 1$$

Equivalently, $\overline{X}_n$ will almost never be further than ϵ of the expected value, μ.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) = 0$$

Bernoulli's Weak Law of Large Numbers: Suppose $X \sim B(n{,}p)$. Then, the expected value is $\mu = np$, so the weak law of large numbers says

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - np\right| > \epsilon\right) = 0$$

However, for binary random variables, such as in the binomial distribution, we can also look at the mean of the proportion of successes, and not just the mean of the number of successes. Doing so, we have,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_n}{n} - p\right| > \epsilon\right) = 0$$

The *strong law of large numbers* states that the sample mean converges almost surely to the expected value;

$$\overline{X}_n \overset{a.s.}{\to} \mu \text{ as } n \to \infty$$

That is, $\mathbb{P}(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$.

The weak law simply states that for some large n, $\bar{X}_n$ is likely to be close to μ, but does not preclude the possibility that $|\bar{X}_n - \mu| > \epsilon$ happens infinitely many times (though, likely only at increasingly infrequent intervals for larger and larger n).

The strong law states that this almost surely does not occur (i.e., has Lebesgue measure 1). Note that this does not imply that for any $\epsilon > 0$, there exists N such that $|\bar{X}_n - \mu| < \epsilon$ holds for all $n > N$, since converging almost surely is not uniform convergence.

The *classical central limit theorem* states that if $\{X_i\}_{i=1}^n$ is an independent and identically distributed sequence of random samples drawn from a population with mean μ and variance σ^2, then the sample mean $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ converges in distribution to $N\left(\mu, \frac{\sigma^2}{n}\right)$, regardless of the distribution of the population.

With very large number of Bernoulli trials, it quickly becomes intractable to calculate factorials. For large n, we often approximate the binomial distribution with other, computationally easier distributions.

Poisson Limit Theorem: Let p_n be a sequence of real numbers in [0,1] such that the sequence np_n converges to some limit $\lambda < \infty$. Then,

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

That is to say, if $X \sim B(n,p)$, n is large, and p is small, then X is approximately $\sim$ Pois(np) (recall that np is the expected value of X).

The *De Moivre–Laplace theorem* is a special case of the central limit theorem. If $X \sim B(n,p)$, then, as $n \to \infty$, X converges in distribution to $N(\mu,\sigma^2)$, where μ is the expected value of X, which is np, and σ^2 is the variance of X, which is $np(1-p)$.

In other words, if $X \sim B(n,p)$, then for large n, X is approximately $\sim N\left(np, \sqrt{np(1-p)}^2\right)$.

Because the normal distribution is continuous, while the binomial is discrete, you need to apply a *continuity correction* when calculating probabilities. If $X \sim N(n,p)$ and $Y \sim N\left(np, \sqrt{np(1-p)}^2\right)$, then,

- $\mathbb{P}(X = a) \approxeq P(a - 0.5 < Y < a + 0.5)$;
- $\mathbb{P}(X > a) \approxeq P(Y > a + 0.5)$;
- $\mathbb{P}(X \geq a) \approxeq P(Y > a - 0.5)$;
- $\mathbb{P}(X < a) \approxeq P(Y > a - 0.5)$;
- $\mathbb{P}(X \leq a) \approxeq P(Y < a + 0.5)$;