Curtin College

DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

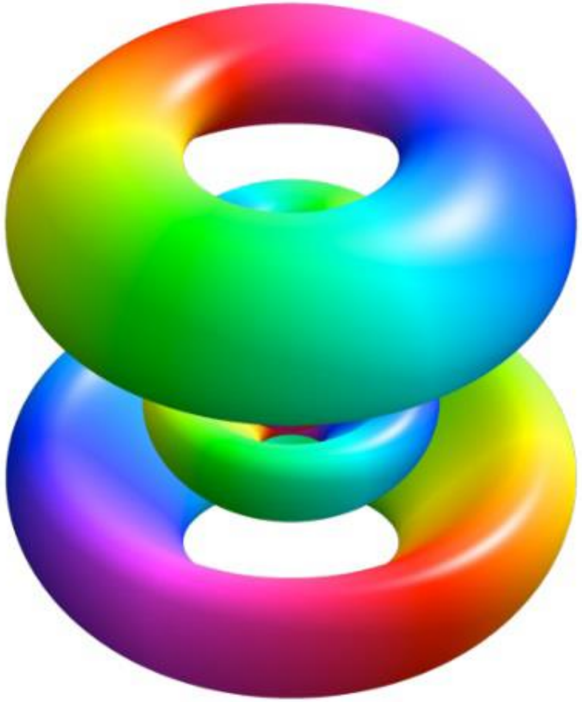
Table of Contents

Why Probability?

- Probability is the mathematics of uncertainty and chance.
- Probability concepts are common in everyday life:
    - probability of rain tomorrow
    - chance that your team will win next weekend
- likelihood that an arrested person is guilty
- What are your intuitive notions of probability?

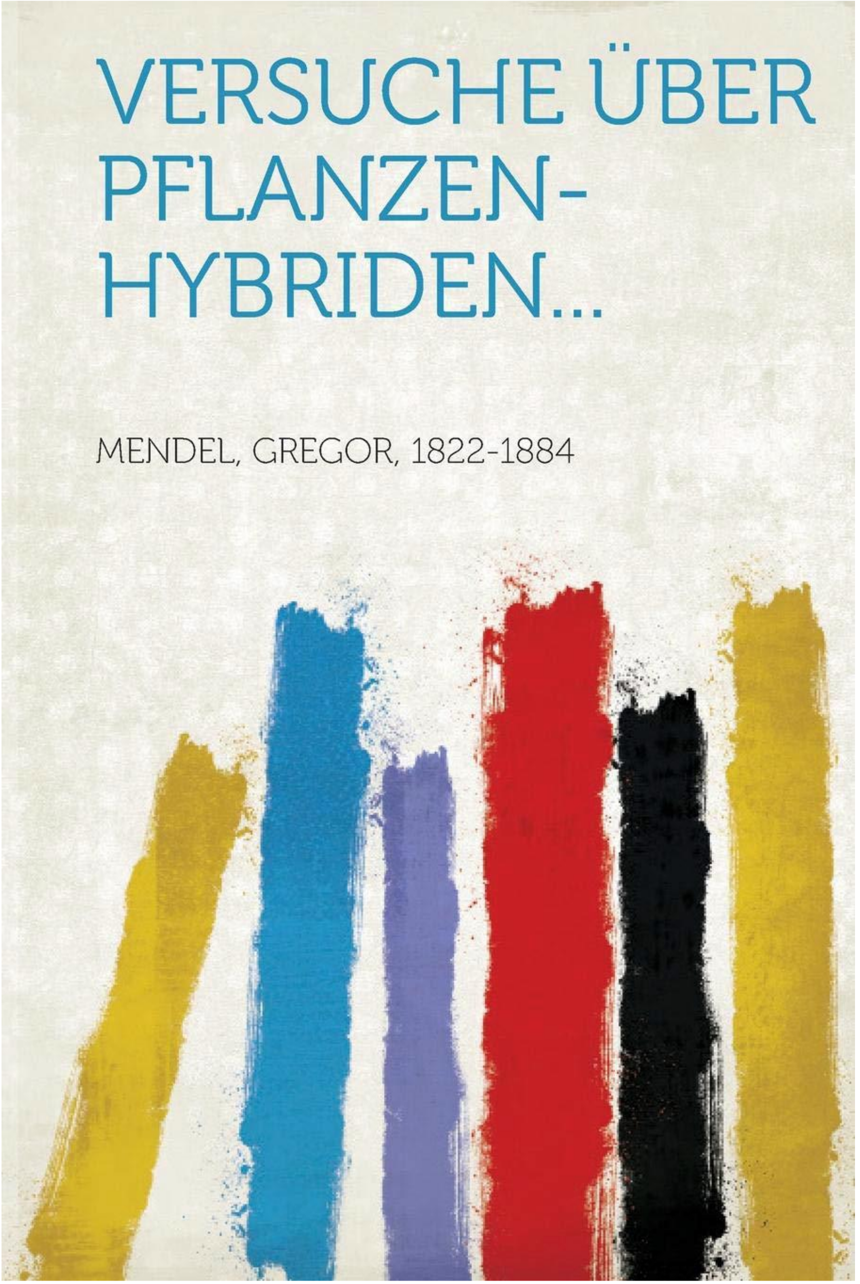
Probability and Quantum Mechanics…

- Electron orbits are representations of the space in which an electron is likely to be found
- The result of wave functions, where the amplitude of the wave function is interpreted as the probability of finding an electron at any point in space



Probability in Genetics

- Fundamental laws of genetics are stated as probability laws, i.e., what is the likelihood of a particular trait – eye colour, for example – being passed on from one generation to the next?



Probability in Weather and Climate

- Forecasts are computed and expressed in terms of probability.
- What is the probability of rain tomorrow?
- Will extreme rainfall events become more likely in the future, and how extreme will they be?

Historical origins

- A gambler’s dispute in 1654 led to the creation of a mathematical theory of probability by two French mathematicians, Blaise Pascal and Pierre de Fermat.
- Antoine Gombaud, (AKA Chevalier de M´er´e) a French noble man and a writer with an interest in gaming and gambling questions, called Pascal’s attention to an apparent misconception concerning a popular dice game.

A Game of Dice

- The game consisted of throwing a pair of dice 24 times. The problem was to decide whether or not to bet even money on the occurrence of at least one double six during the 24 throws. An established gambling rule suggested that betting even money on a double six in 24 throws would be profitable. But de M´er´e did his own calculations and cast doubt on the rule.
- This problem led to an exchange of letters between Pascal and Fermat in which the fundamental principles of probability were formulated for the first time.
- Questions:
  - What is the probability of getting at least one six in 4 throws of a die?
  - What is the probability of getting at least one double-six in 24 throws of a pair of dice?

What is Probability Used for?

- Epidemiology: The spread of virus among humans
- Agriculture: Crop trials, plant genetics
- Meteorology: Weather prediction
- Sports strategies: Probabilities are used for planning and gaining competitive advantage.
- Insurance options: Risk assessment
- Games and recreational activities: What are the chances of getting a full house in a poker hand?
- Economics and finance: Performance of stock market
- And many more. . .

Probability and Statistics

- As the systematic study of randomness and uncertainty, probability provides:
  - methods for quantifying the chance of different outcomes.
  - models for describing and quantifying randomness.
- Inferential statistics investigates the world by examining limited data. Since the data are limited, statistical results are unavoidably uncertain. Probability lets us quantify and control this unavoidable uncertainty.
- Inferential statistics commonly asks two basic questions:
  - given the data, what is our best estimate of a population parameter (e.g., mean, proportion, variance)?
  - do the data display a chance result, or is the world different to what we thought?
- Probability theory provides a principled framework for answering such questions.

What is Statistics Used for?

- Medical research: Drug trials, public health, genetics
- Agriculture: Crop trials, plant genetics
- Forensic science: DNA evidence
- Astronomy: Galaxy distribution, exoplanets
- Geology
- Internet security
- Environmental science: Ecology, Climate
- Many articles in scientific journals use statistics to analyze data and draw conclusions.
- Climate is a statistical description of weather conditions and their
- variations, including both averages and extremes – Australian Academy of Science The Science of Climate Change, August 2010

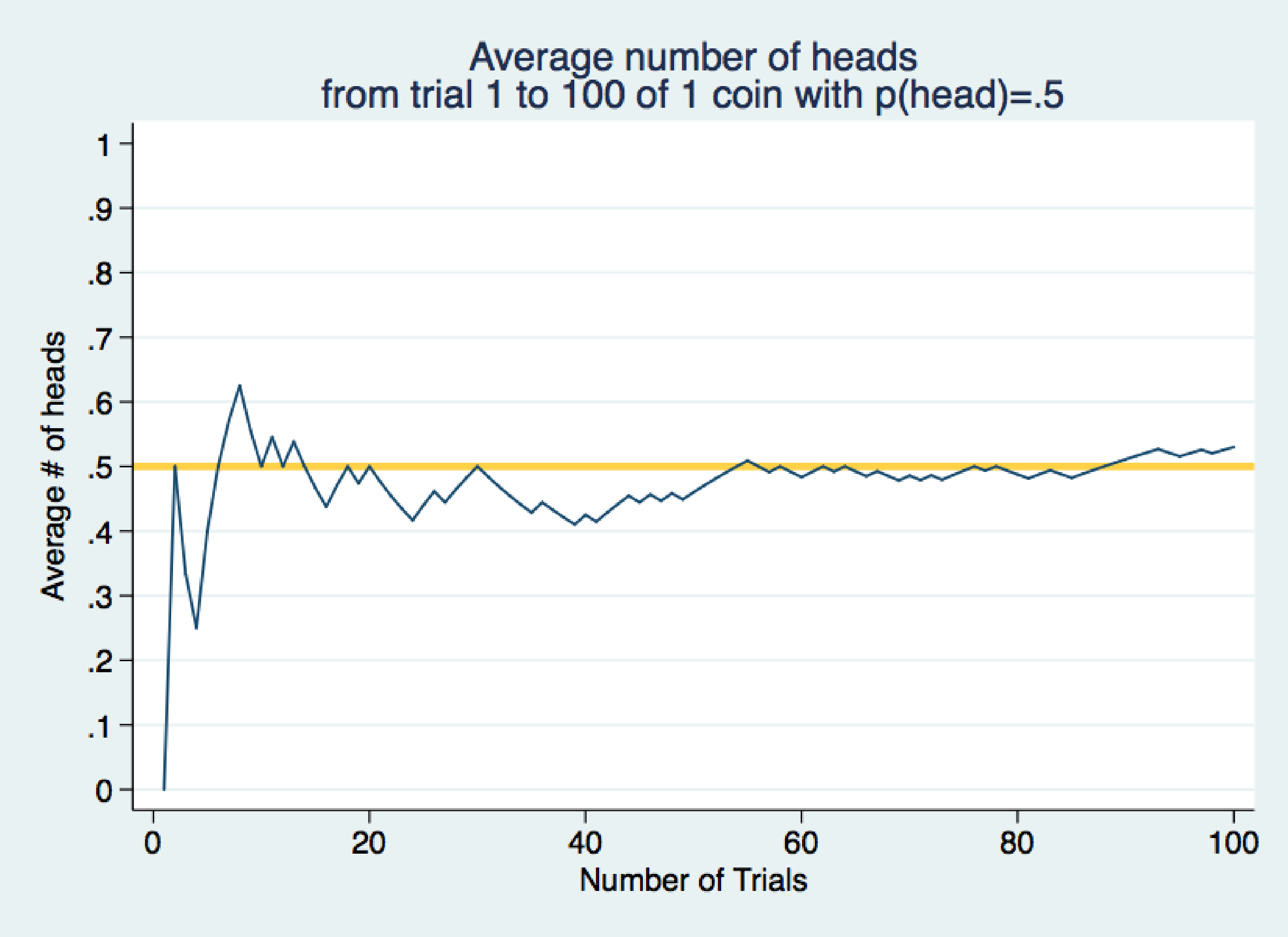
Unit Objectives

- Introduce the theory of probability and random variables
- Provide you with basic tools for analysing data
- Introduce you to statistics as a way of thinking
- Gain understanding of the role of statistical inference in drawing conclusions about the world
- By the end of the unit, you should
  - Have a good grounding that will allow you to take more advanced units;
  - Have a firm grasp of the fundamental principles of probability;
  - Be on your way to becoming a statistically literate citizen who can cast a critical eye over claims made in the media and elsewhere.

Basic Ideas of Probability

- Chance behaviour is unpredictable in the short run, but has a regular and predictable pattern in the long run.
- We call a phenomenon random if individual outcomes are uncertain.
- We can apply probability if, despite the randomness, there is a regular distribution of outcomes in a large number of repetitions.
- The probability of any outcome of a chance process is the proportion of times the outcome would occur in a very long series of repetitions.
- Mathematical probability is an idealization based on imagining what would happen in an indefinitely long series of trials.

Law of Large Numbers in a Coin Toss Experiment



Probability Models

- Descriptions of chance behaviour contain two parts: a list of possible outcomes and a probability for each outcome.
- The sample space S of a random phenomenon is the set of all possible outcomes.
- An event is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.
- A probability model is a mathematical description of a random phenomenon consisting of two parts: a sample space S and a way of assigning probabilities to events.

Probability Rules

- Any probability is a number between 0 and 1.
- All possible outcomes together must have probability 1.
- If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.
- The probability that an event does not occur is 1 minus the
- probability that the event does occur.

Probability Rules

- Mathematically,
- *Rule 1.* The probability $P(A)$ of any event A satisfies $0 \leq P(A) \leq 1$.
- *Rule 2.* If *S* is the sample space in a probability model, then $P(S) = 1$.
- *Rule 3.* If *A* and *B* are disjoint, $P(A \cup B) = P(A) + P(B)$. This is
- the addition rule for disjoint events.
- *Rule 4.* For any event *A*, $P(A\ does\ not\ occur) = 1 - P(A)$.

Sample Spaces, Outcomes and Events

Example 1

- Toss a coin once: $S = \{H, T\} = \{Head, Tail\}$
- The number of customers arriving at a bank on a randomly selected day: $S = \{0, 1, 2, \ldots\}$
- Measure someone's systolic blood pressure: $S = \{x : x \geq 0\} = [0, \infty)$

Remark 1

- In the first example, the sample space is finite.
- In the second example, the sample space is countably infinite.
- In the third example, the sample space is uncountably infinite.
- A set that has either a finite or a countably infinite number of elements will be
- called countable.
- In the first two examples the sample space is said to be discrete (countable), while
- in the last example S is said to be continuous (uncountably infinite, uncountable).

Union, Intersection and Complement

- An event is just a set of outcomes, so that relationships and results from
- elementary set theory can be used to study events.

Definition 2

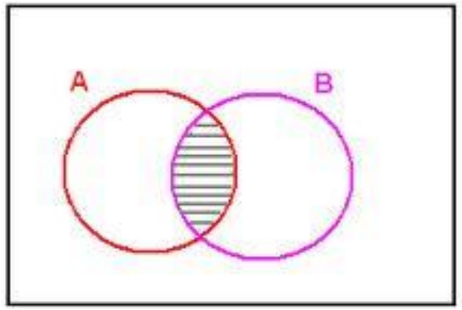
- The union of two events *A* and *B*, denoted by $A \cup B$ and read *"A or B"*, is the event consisting of all outcomes that are either in *A* or in *B* or in both events.

- The intersection of two events *A* and *B*, denoted by $A \cap B$ and read *"A and B"*, is the event consisting of all outcomes that are in both *A* and *B*.

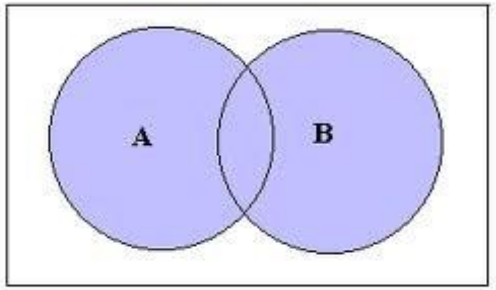
- The complement of an event *A*, denoted by A', A^c or $\overline{A}$, is the set of all outcomes in *S* that are not contained in *A*.

Union, Intersection and Complement

Intersection: $A \cap B$ is the set of points that are both in *A* and in *B*.

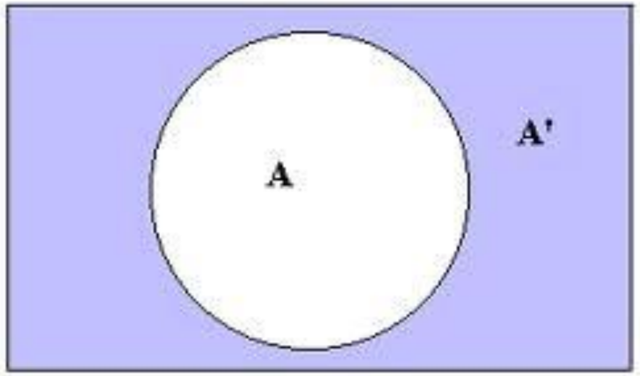


Union: $A \cup B$ is the set of points that are in *A* or *B* or both.

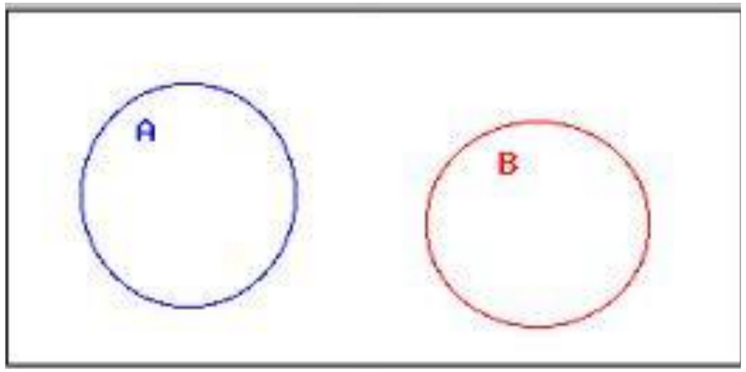


Complementary and Disjoint Events

- The complement of an event A, denoted by A', A^c or $\overline{A}$, is the set of all outcomes in S that are not contained in A.



Definition 3:
When A and B have no outcomes in common, they are said to be mutually exclusive or disjoint events. ($A \cap B = \emptyset$)



Exercises
1. What is the complement of A^c ?
2. What is the intersection of A and A^c ?

De Morgan's Laws

Example 2

- Two coins are tossed.
- Let A denote the event that the first toss yields an H.

(a) What is the sample space?

$$S = \{HH, HT, TH, TT\}$$

(b) What is the complement of A?

$$A = \{HH, HT\}, \qquad A^c = \{TH, TT\}$$

(c) Write down an event that is disjoint with A.

A^c is disjoint with A. $\{TH\}$ and $\{TT\}$ are also disjoint with A.

De Morgan's laws:

For two events $A, B \subseteq S$:

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

Example 3

A coin is flipped 10 times. Writing Heads as H and Tails as T, a possible outcome is HHHTHHTTHT, and the sample space is the set of all possible strings of length 10 of Hs and Ts. We can encode H as 1 and T as 0, so that an outcome is a sequence $(s_1, \ldots s_{10})$ with $s_j \in \{0, 1\}$, and the sample space is the set of all such sequences.

Now let's look at some events:

- Let A_1 be the event that the first flip is Heads. As a set,
  - $A_1 = \{(1, s_2, \ldots s_{10}): s_j \in \{0, 1\}, \text{for } 2 \leq j \leq 10\}$. This is a subset of the sample space, so it is indeed an event;
  - Saying that A_1 occurs is the same as saying that the first flip is Heads. Similarly, let A_j be the event that the jth flip is Heads for j = 2, 3, . . . , 10.
- Let B be the event that at least one flip was Heads.
  - As a set $B = \bigcup_{j=1}^{10} A_j$

De Morgan's Laws

- Let C be the event that all flips were Heads.
  - As a set $C = \bigcap_{j=1}^{10} A_j$
- Let D be the event that there were at least two consecutive Heads.

As a set, $\mathrm{D} = \bigcup_{j=1}^{9} (A_j \cap A_{j+1})$

Events can be described using words or in set notation.

Events and occurrences

sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occured	$s_{actual} \in A$
something must happen	$s_{actual} \in S$

New events from old events

A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$

Relationships between events

A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \emptyset$
$A_1, \ldots, A_n$ are a partition of S	$A_1 \cup \cdots \cup A_n = S$, where
$A_i \cap A_j = \emptyset$ for $i \neq j$	

The Birthday Problem

There are k people in a room. Assume:

- there are 365 days in a year (we ignore leap years)
- all days are equally likely to be birthdays
- people's birthdays are independent
- there are no twins in the room

What is the probability that two or more people in the group have the same birthday?

Focus on the complement of the desired event, i.e., no shared birthdays.

- P(no shared birthdays for 2 people) = $\frac{364}{365}$
- P(no shared birthdays for 3 people) = $\frac{364}{365} \times \frac{363}{365}$
- P(no shared birthdays for 4 people) = $\frac{364}{365} \times \frac{363}{365} \times \frac{362}{365}$
- P(no shared birthdays for n people) = $\frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365-n+1}{365}$
- Then P(at least one shared birthday) = 1 − P(no shared birthdays)

Naıve Definition of Probability

The naıve definition applies when every outcome in S is equally likely to occur.

Definition 4 (Naıve definition of probability)
Let A be an event for an experiment with a finite sample space S. The naıve probability of A is

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

Remark 3
P(A) can be defined in this way only if S is a finite set. Experiments in which each outcome is equally likely to occur are also called Laplace experiments.

Naıve Definition of Probability

There are several important types of problems where the naıve definition is applicable:

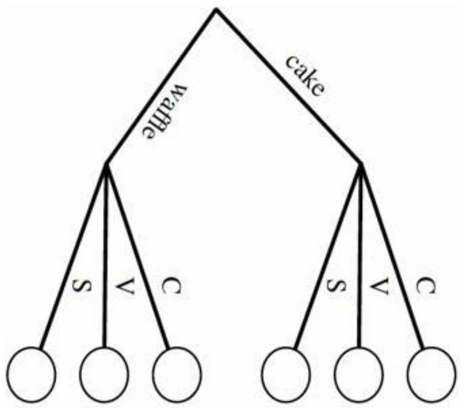
- when there is symmetry in the problem that makes outcomes equally likely. . .
- when the outcomes are equally likely by design. For example, consider a survey of n randomly-selected people from a large population. . .
- we might make predictions based on the naıve definition, compare the predictions to observed data, and so assess whether it is reasonable toassume equally likely outcomes.

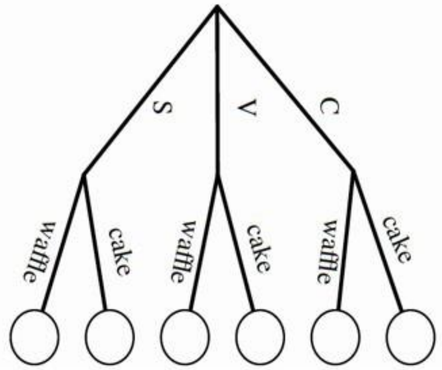
Counting Methods

- Calculating the naıve probability of event A involves counting the number of outcomes in A and in the sample space S.
- Enumerating outcomes is OK when the list is small, but in many real cases such lists are prohibitively long. General counting methods allow us to calculate numbers of outcomes without enumeration.
- Often, we can count the number of outcomes by multiplication.
  - Consider a compound experiment consisting of two sub-experiments such that the first has n_1 outcomes and for each possible outcome for the first, the second has n_2 outcomes. Then the compound experiment has $n_1 \times n_2$ possible outcomes.
  - This rule is known as the fundamental principle of counting.

An Example of the Multiplication Rule

- At your local cinema, you can buy ice cream with a choice of Bacio, Hazelnut, and Tiramisu flavours, and with wafer or waffle cones. By the multiplication rule, there are 2 × 3 = 6 choices, and you can visualize the decision process with a tree diagram.
- It doesn’t matter whether you choose the cone first or the flavour first.
- As long as there is the same number of choices of the second element for each first element, the product rule is valid even when the set of possible second elements depends on the first, i.e. the option of flavours are different for different types of cones.





General Multiplication Rule

- If a compound experiment consists of k sub-experiments such that the ith sub-experiment has n_i outcomes and for each combination of outcomes for the others, then the compound experiment has $k \prod_{i=1}^{k} n_i = n_1 \times n_2 \times \cdots \times n_k$ possible outcomes.
- This can also be stated another way: Suppose that a set consists of ordered collections of k-elements (k-tuples) such that there are n_1 possible choices for the first element, for each choice of the first element, there are n_2 possible choices of the second element, etc., and for each possible choice of the first k − 1 elements, there are n_k choices of the kth element. Then there are $n_1 \times n_2 \times \cdots \times n_k$ possible k-tuples.

Counting examples

If the set S has n distinct elements, how many possible subsets of S exist?

- For each element of S, we can choose whether or not to include it a subset. This gives us 2 possibilities for each of the n elements. So from the product rule, it follows that there are 2n subsets.
- If we sample with replacement from S until we have k objects, n^k such samples are possible.

Using the 26 letters in the English alphabet, a 3-letter 'word' is generated at random. What is the probability that the word contains neither 'a' nor 'e'?

- Here $|S| = 26^3$. If A is the set of words that have no 'a' or 'e', then $|A| = 24^3$. Hence by the naıve definition of probability,

$$P(A) = \frac{24^3}{26^3} = \left(\frac{12}{13}\right)^3 \approx 0.7865$$

Examples

- If we sample without replacement k objects from a set of n objects, there are $n \times (n-1) \times \cdots (n-k+1)$ possible outcomes if $k \leq n$ (and zero possibilities if $k > n$).
- In the 'words' example, what if we now specify that the words cannot contain the same letter more than once?
- Now $|S| = 26 \times 25 \times 24$ and $|A| = 24 \times 23 \times 22$. Thus

$$P(A) = \frac{26 \times 25 \times 24}{24 \times 23 \times 22}$$

- This leads us to the definition of permutations. Any ordered sequence of k objects taken from a set of n distinct objects is called a permutation of size k of the objects.
- We first introduce the factorial notation: n!, pronounced as 'n factorial', is defined as $n! = n \times (n-1) \times \cdots 2 \times 1$.

Permutations

The number of permutations of k objects selected from n is denoted by P_k^n (or $P_{n,k}$) and is given by

$$P_k^n = n \times (n-1) \times \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

- Ten markers are available for grading a test. The test has four questions, and the professor wishes to select a different marker to grade each question with only one marker per question. In how many ways can markers be chosen to grade the test?
- Here n = 10 and k = 4, so the number of different ways markers can be assigned is

$$P_4^{10} = \frac{10!}{6!} = 5040$$

Permutations

- A deck of 52 cards is shuffled and 5 cards are dealt face up in a row. Find the probability that four aces present in adjacent position. The total number of arrangements is

$$P_5^{52} = 311875200$$

- The number of arrangements where the four aces are present and are in adjacent position is

$$48 \times 4! \times 2! = 2304.$$

- Thus, the required probability is

$$\frac{2304}{311875200} = 0.000007388.$$

- How many permutations are there of the word STATISTICS?
- First, we think of the repeated letters as distinct to get 10!.

Then we divide it by the factors by which the repetitions occur to get the number of permutations to be

$$\frac{10!}{3!3!2!} = 50400$$

Combinations

- In permutations, order matters. In the 'words' example, 'pqr' is different from 'rpq'.
- In many situations, the order in which objects occur does not matter, in which case 'pqr' and 'rpq' would be considered the same.
- An unordered subset of k objects selected from n distinct objects is called a combination of size k.
- The number of combinations of size k that can be formed from n distinct objects denoted by

$$\binom{n}{k} = C_k^n = C_{n,k}$$

- For given n and k, the number of combinations is smaller than the number of permutations (unless k = 1, in which case they are equal), because when order is disregarded, some of the permutations correspond to the same combination.

Combinations

In the set $\{A, B, C, D, E\}$, the number of permutations of size 3 is

$$\frac{5!}{(5-3)!} = 60$$

- Consider a single subset $\{A, B, C\}$. There are $3! = 6$ permutations that corresponding to a single combination.
- Similarly, for any other combination of size 3 there are 6 permutations, each obtained by ordering the three objects.
- Thus, we're overcounting by a factor of 6, and hence the number of combinations is

$$\frac{5!}{(5-3)!} \times \frac{1}{3!}$$

- That is:

$$\{A, B, C\}\ \{A, B, D\}\ \{A, B, E\}\ \{A, C, D\}\ \{A, C, E\}$$
$$\{A, D, E\}\ \{B, C, D\}\ \{B, C, E\}\ \{B, D, E\}\ \{C, D, E\}$$

Combinations

More generally, when there are n distinct objects, the number of combinations of size k is given by

$$\binom{n}{k} = \frac{P_k^n}{k!} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{1 \times 2 \times 3 \times \ldots k}$$

- $\binom{n}{0} = \binom{n}{n} = 1$ because there is only one way to choose a set of all n elements or of no elements.
- $\binom{n}{1} = n$ since there are n subsets of size 1.
- $\binom{n}{k} = \binom{n}{n-k}$. For instance, $\binom{20}{18} = \binom{20}{2} = 190.$
- How many committees of two actuarial science students and one data science student be formed from four actuarial science students and three data science students?

Answer:

$$\binom{4}{2} \times \binom{3}{1} = 6 \times 3 = 18$$

Combinations

- In how many different ways can 6 tosses of a coin yield 2 heads and 4 tails?

Answer:

$$\binom{6}{2} = 15$$

- Eight politicians meet at a fund-raising dinner. How many greetings can be exchanged if each politician shakes hands with every other politician exactly once?

Answer:

A handshake corresponds to an unordered sample of size 2 chosen from the 8 politicians. So the answer is

$$\binom{8}{2} = 28$$

- If there are 20 points on a paper such that no three are collinear,
  - how many line segments can one create with two of these points as endpoints?
  - how many triangles can one form with three of these points as vertices?

Answer:

$$\binom{20}{2} = 190 \quad and \binom{20}{3} = 1140$$

Respectively.