50008

Probability and Statistics
Imperial College London

Contents

Chapter 1

Elementary Probability Theory

Probability theory is a mathematical formalism to describe and quantify uncertainty.

Uses of probability include examples such as:

- Finding distribution of runtimes & memory usage for software.
- Response times for database queries.
- Failure rate of components in a datacenter.

1.1 Sample Spaces and Events

Sample Space **Definition 1.1.1**

The set of all possible outcomes of a random experiment. The set is usually denoted with set notation, and can be finite, countably or uncountably infinite.

For example:

Experiment	Sample Space
Coin Toss	$S = \{Heads, Tails\}$
6-Sided Dice Roll	$S = \{1, 2, 3, 4, 5, 6\}$
2 Coin Tosses	$S = \{(H, H), (H, T), (T, H), (T, T)\}$
Choice of Odd number	$S = \{x \in \mathbb{N} \mid \exists y \in \mathbb{N}.[2y + 1 = x]\}$

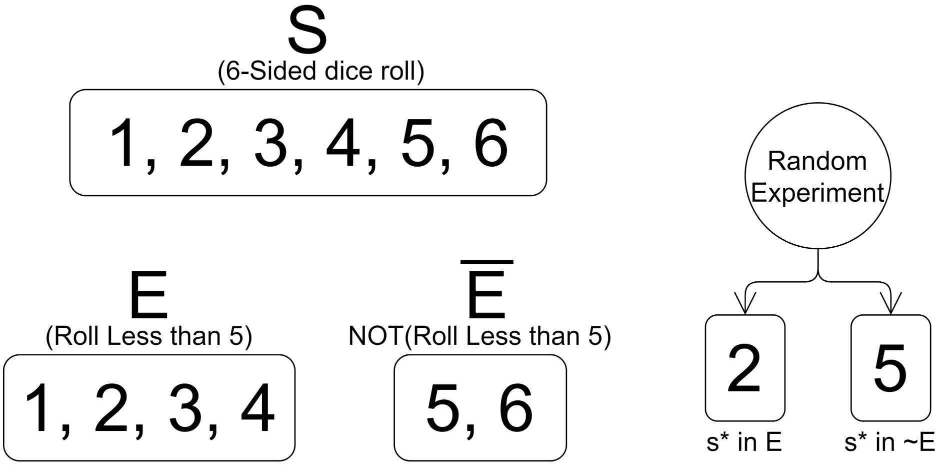
Event **Definition 1.1.2**

Any subset of the sample space $E \subseteq S$ (a set of possible outcomes).

- **null event**($\emptyset$) Empty event, can be used for impossible events.
- **universal event** (S) Event contains entire sample space and is therefore certain.
- **elementary events** Singleton subsets of the sample space (contain one element).

For example:

Event	Set of Event	Sample Space
6-Sided Dice Rolls 1	$E = \{1\}$	$S = \{1, 2, 3, 4, 5, 6\}$
6-Sided Dice Rolls Even	$E = \{2, 4, 6\}$	$S = \{1, 2, 3, 4, 5, 6\}$
6-Sided Dice Rolls 7	$E = \emptyset$	$S = \{1, 2, 3, 4, 5, 6\}$
2 Coin toss get 2 Tails	$E = \{(T, T)\}$	$S = \{(H, H), (H, T), (T, H), (T, T)\}$
Random Natural Number is 4	$E = \{4\}$	$S = \mathbb{N}$



- If we perform a random experiment with outcome $S* \in S$. If $s* \in E$, then event E has occurred.
- If E has not occurred ($s* \notin E$) then $s* \in \overline{E}$.
- The set $\{s*\}$ is an elementary event.
- Null event $\emptyset$ never occurs, the universal event S always occurs.

1.1.1 Set Operations on Events

- **Union / Or**
  $$\bigcup_i E_i = \{s \in S | \exists i.[s \in E_i]\}$$
  Occurs if at least one of the events E_i has occurred (has union of event sets).

  If 4 is rolled on a 6-sided dice, then union of (is 3) and (is 4) occurred.
- **Intersection / And**
  $$\bigcap_i E_i = \{s \in S | \forall i.[s \in E_i]\}$$
  Occurs if all the events E_i occur.

  If 4 is rolled on a 6-sided dice, the intersection of (is even) and (is 4) occurred.
- **Mutual Exlusion**
  $$E_1 \cap E_2 = \emptyset$$
  If sets are disjoint, then they are mutually exclusive (cannot occur simultaneously).

  For a 6-sided dice the events (is 4) and (is 6) are mutually exclusive.

1.1.2 Probability

When determining the probability of every subset $E \subseteq S$ occurring:

- S **is Finite** Can easily assign probabilites.
- S **is countable** Can assign probabilites.
- S **is uncountably infinite**
  Can initially assign some collection of subsets probabilities, but it then becomes impossible to define probabilities on reamining subsets.

  Cannot make probabilities sum to 1 with reasonably axioms.

For this reason when defining a probability function on sample space S, we must define the collection of subsets we will measure.

The subsets are referred to as $\mathcal{F}$ and must be:

1. nonempty ($S \in \mathcal{F}$)
2. closed under complements $E \in \mathcal{F} \Rightarrow \overline{E} \in \mathcal{F}$
3. closed under countable union $E_1, E_2, \dots \in \mathcal{F} \Rightarrow \bigcup_i E_i \in \mathcal{F}$

A collection of sets is known as σ-algebra.

Probability Measure **Definition 1.1.3**

A function $P : \mathcal{F} \to [0, 1]$ on the pair $(S, \mathcal{F})$ such that:

Axiom 1. $\forall E \in \mathcal{F}.[0 \leq P(E) \leq 1]$
Axiom 2. $P(S) = 1$
Axiom 3. Countably additive, for **disjoint** sets $E_1, E_2, \dots \in \mathcal{F}$: $P(\bigcup_i E_i) = \sum_i P(E_i)$

$P(E)$ provides the probability (between 0 and 1 inclusive) that a given event occurs.

From the axioms satisfied by a *probability measure* we can derive that:

1. $P(\overline{E}) = 1 - P(E)$
2. $P(\emptyset) = 0$
3. For any events E_1 and E_2: $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

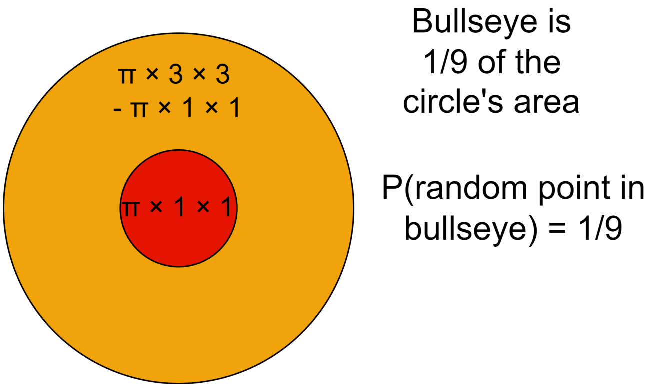
1.2 Interpretations of Probability

1.2.1 Classical Interpretation

Given S is finite and the *elementary events* are equally likely:

$$P(E) = \frac{|E|}{|S|}$$

We can also extend this *uniform probability distribution* to infinite spaces by considering measures such as area, mass or volume.



1.2.2 Frequentist Interpretation

Through repeated observations of identical random experiments in which E can occur, the proportion of experiments where E occurs tends towards the probability of E.

At an infinite number experiments, the proportion of occurrences of E is equal to $P(E)$.

Central Limit Theorem ***Extra Fun!* 1.2.1**

This can also be considered in terms of *central limit theorem*, where the greater the sample size taken from some distribution (with defined mean μ), the closer the mean of the sample to the distribution's mean. (more readings results in less variance in the sample means as they converge on the distribution's mean)

1.2.3 Subjective Interpretation

Probability is the degree of belief held by an individual.

For example if gambling: Option 1: E occurs win £1, $\overline{E}$ occurs win £0
Option 2: Regardless of outcome get £$P(E)$.

Either outcome, the gambler receives £$P(E)$. The value of $P(E)$ is the value for which the individual is indifferent about the choice between option 1 or 2. It is the *individuals probability* of event E occurring.

1.3 Joint Events and Conditional Probability

We commonly need to consider *Join Events* (where two events occur at the same time).

Independent Events **Definition 1.3.1**

Two events are independent if the occurence of one does not affect the other. Given E_1 and E_2 are independent:

$$E_1 \textit{ and } E_2 \textit{ independent} \Leftrightarrow P(E_1 \textit{ occurrs and } E_2 \textit{ occurs}) = P(E_1) \times P(E_2)$$

More generally, the set of events $\{E_1, E_2, \dots\}$ are independent if for any finite subset $\{E_{i_1}, E_{i_2}, \dots, E_{i_n}\}$:

$$p(\bigcap_{j=1}^{n} E_{i_j}) = \prod_{j=1}^{n} P(E_{i_j})$$

If E_1 and E_2 are independent, then so are $\overline{E_1}$ and E_2.

For example with a coin toss, subsequent coin tosses do not effect the next coin toss's probability of heads.

We can show that if E_1 and E_2 are independent, so are $\overline{E_1}$ and E_2:

(1) $F = (E_1 \cap E_2) \cup (\overline{E_1} \cap E_2)$ By set operations
(2) $P(E_2) = P(E_1 \cap E_2) + p(\overline{E_1} \cap E_2)$ As **1** was a disjoint union, Axiom 3
(3) $P(\overline{E_1} \cap E_2) = P(E_2) - P(E_1 \cap E_2)$
(4) $P(\overline{E_1} \cap E_2) = P(E_2) - P(E_1) \times P(E_2)$
(5) $P(\overline{E_1} \cap E_2) = P(E_2) \times (1 - P(E_1)$
(6) $P(\overline{E_1} \cap E_2) = P(E_2) \times P(\overline{E_1})$ By $P(\overline{E}) = 1 - P(E)$

We can show that $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$:

(1) $E_1 \cup E_2 = E_1 \cup (E_2 \cap \overline{E_1})$ From set theory
(2) $P(E_1 \cup E_2) = P(E_1 \cup (E_2 \cap \overline{E_1}))$ By Axiom 3
(3) $P(E_1 \cup E_2) = P(E_1) + P(E_2 \cap \overline{E_1})$
(4) $P(E_2 \cap \overline{E_1}) = P(E_2) - P(E_1 \cap E_2)$ By **3** of the previous proof and as E_1 and E_2 are independent

Dice for Money **Example Question 1.3.1**

We can construct a *Probability Table*:

		Dice						Totals
		1	2	3	4	5	6	
Coin	H	$1/12$	$1/12$	$1/12$	$1/12$	$1/12$	$1/12$	$1/2$
	T	$1/12$	$1/12$	$1/12$	$1/12$	$1/12$	$1/12$	$1/2$
Totals		$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	

We can determine the probability of any event by summing the probabilities of elementary events represented by cells in the table.

$P(H)$ is called a *marginal probability*, as it the probability of one event occurring irrespective of the other (the dice in this case).

$P((H, 3))$ is called a *joint probability* as it involves both events (dice roll and the coin toss).

Roll of the Die **Example Question 1.3.2**

A crooked die (called a top) has the same faces on either side.

We flip the coin, then if it is heads we use the normal die, else we use the top.

		Dice						Totals
		1	2	3	4	5	6	
Coin	H	1/12	1/12	1/12	1/12	1/12	1/12	1/2
	T	1/6	0	1/6	0	1/6	0	1/2
Totals		1/4	1/12	1/4	1/12	1/4	1/12	

We can now see that $P(\{(H,3)\}) \neq P(\{H\}) \times P(\{3\})$ and hence they are dependent, as the dice roll depends on the coin toss.

1.4 Conditional Probability

For two events E and F in *sample space* S, where $P(F) \neq 0$:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Probability of E given F is the probability of both occurring over the probability of F.

Independence ***Extra Fun!* 1.4.1**

If E and F are independent:

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E) \times P(F)}{P(F)} = P(E)$$

Conditional Independence **Definition 1.4.1**

$P(\bullet|F)$ defines a probability measure obeying the axioms of probability on set F (When have just reduced S to F).

Three events E_1, E_2, F are conditionally independent if and only if:

$$P(E_1 \cap E_2|F) = P(E_1|F) \times P(E_2|F)$$

W **Example Question 1.4.1**

hat is the probability the dice rolls a 3 given the dice rolls an odd number?

$$P(\{3\}|\{1,3,5\}) = \frac{P(\{3\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{P(\{3\})}{P(\{1,3,5\})} = \frac{1/6}{1/2} = \frac{1}{3}$$

Go big or go home! **Example Question 1.4.2**

Throw a die from each hand. What is the probability the die thrown from the left is larger than the die thrown from the right.

The sample space is:

$$S = \left\{ \begin{matrix} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{matrix} \right\}$$

We want the event such that the left value of the pair is larger.

For value 1 there are 0 possible, for 2 there is 1 and so on.

$$(1:0),(2:1),(3:2),(4:3),(5:4),(6:5)$$

Hence there are $0+1+2+3+4+5=15$ possible pairs with the left larger than the right.

$$P(E)=\frac{15}{36}=\frac{5}{12}$$

However if we know the left or right die, we can determine a new probability. For example if we know the left die is 4 then we know there are 6 pairs with the left as 4, and 3 of those pairs have a smaller right.

$$P(E|4)=\frac{3}{6}=\frac{1}{2}$$

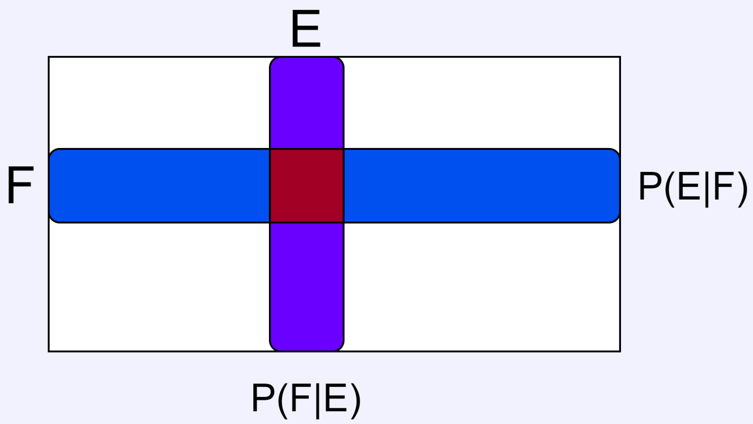
Bayes Theorem **Definition 1.4.2**

For two events E and F we have:

$$P(E\cap F)=P(F)\times P(E|F)=P(F)\times\frac{P(E\cap F)}{P(F)}=P(E)\times P(F|E)=P(E)\times\frac{P(E\cap F)}{P(E)}$$

Hence we can deduce:

$$P(E|F)=\frac{P(E)\times P(F|E)}{P(F)}$$

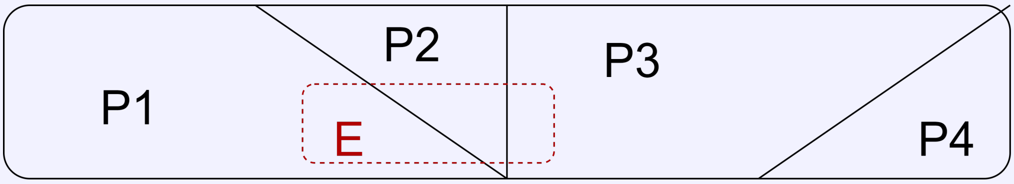


Partition Rule **Definition 1.4.3**

Given a set of events $\{F_1, F_2, \dots\}$ which forms a partition of S (disjoint sets that contain all of F).

For any event $E \subseteq S$:

$$P(E) = \sum_i P(E|F_i) \times P(F_i)$$



Proof:

(1) $E = E \cap S = E \cap \bigcup_i F_i = \bigcup_i (E \cap F_i)$ By set theory and disjointness of partitions.
(2) $P(E) = P(\bigcup_i (E \cap F_i))$
(3) $P(E) = \sum_i P(E \cap F_i)$ By axiom 3 and disjointness of partitions.
(4) $P(E) = \sum_i P(E|F_i) \times P(F_i)$

Law of Total Probability **Definition 1.4.4**

Given some event E and events $\{F_1, F_2, \dots\}$:

$$P(E) = \sum_i P(E \cap F_i)$$

For example the 6-Sided dice, $E = H$ and $F = [\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}]$, the marginal probability is the same as the sum of all cells in row H.

Using complement as a partition we can deduce that:

$$P(E) = P(E \cap F) + P(E \cap \overline{F})$$

$$P(E) = P(E|F) \times P(F) + P(E|\overline{F}) \times P(\overline{F})$$

1.4.1 Terminology Recap

- **Conditional Probabilities** Of the form $P(E|F)$.
- **Joint Probabilities** Of the form $P(E \cap F)$.
- **Marginal Probabilities** Of the form $P(E)$.

Chapter 2

Random Variables

Probability Space **Definition 2.0.1**

$$(S, \mathcal{F}, P)$$

Models a random experiment where probability measure $P(E)$ is defined on subsets $E \subseteq S$ belonging to sigma algebra $\mathcal{F}$.

Within a sample space we can study quantities that are a function of randomly occurring events (e.g temperature, exchange rates, gambling scores).

Random Variable **Definition 2.0.2**

A *random variable* is a mapping from the sample space to the real numbers, for example *random variable* X:

$$X : S \to \mathbb{R}$$

Each element in the sample space $s \in S$ is assigned to a numerical value by $X(s)$.

When referring to the value of a random variable we use its name, e.g X in $P(5 < X \leq 30)$

- **Simple** Finite set of possible outcomes. (e.g dice faces)
- **Discrete** Countable outcomes/support/range. (e.g distance (m))
- **Continuous** Can be a continuous range (e.g temp)

Single Fair Dice Roll **Example Question 2.0.1**

$S = \{1, 2, 3, 4, 5, 6\}$, for any $s \in S . P(\{s\}) = \frac{1}{6}$.
We can define random variable X such that:

$$X(1) = 1, X(2) = 2, X(3) = 3, X(4) = 4, X(5) = 5, X(6) = 6$$

Then we can use X:

$$P_X(1 < X \leq 5) = P(\{2, 3, 4, 5\}) = {}^2\!/_3$$

$$P_X(X \in \{2, 3\}) = P(\{2, 3\}) = {}^1\!/_3$$

We can also define random variable Y such that:

$$Y(\epsilon) = \begin{cases} 0 & \epsilon \text{ is odd} \\ 1 & \epsilon \text{ is even} \end{cases}$$

And hence:

$$P_Y(Y = 0) = P(\{1, 3, 5\}) = {}^1\!/_2$$

2.1 Induced Probability

The probability measure P defined on a sample space S induces a probability distribution on the random variable in $\mathbb{R}$ (distribution of its outcomes).

$$S_X = \{s \in S | X(s) \leq x\}$$

Such that:

$$P_X(X \geq x) = P(S_X)$$

Note that unless there is ambiguity, $P_X(\dots)$ will often be written as $P(\dots)$.

Heads and Tails **Example Question 2.1.1**

We define random variable $X : \{H, T\} \to \mathbb{R}$ over the *continuum* $\mathbb{R}$ such that:

$$X(T) = 0 \text{ and } X(H) = 1$$

$$S_X = \begin{cases} \emptyset & \text{if } x < 0 \\ \{T\} & \text{if } 0 \leq x < 1 \\ \{H, T\} & \text{if } x \geq 1 \end{cases}$$

X represents the number of heads flipped.

$$P_X(X \leq x) = P(S_X) = \begin{cases} P(\emptyset) = 0 & \text{if } x < 0 \\ P(\{T\}) = {}^1\!/_2 & \text{if } 0 \leq x < 1 \\ P(\{H, T\}) = 1 & \text{if } x \geq 1 \end{cases}$$

Now we can use X to compactly show probabilities.

$$P_X(X = 1) = {}^1\!/_2$$

Multiple Coin Flips **Example Question 2.1.2**

$$S = \{TTT, TTH, THT, HTT, THH, HHT, HTH, HHH\}$$

We can define X (number of heads):

$$X(s) = \begin{cases} 0 & s = TTT \\ 1 & s \in \{TTH, THT, HTT\} \\ 2 & s \in \{THH, HHT, HTH\} \\ 3 & s = HHH \end{cases}$$

Hence given 3 coin tosses:

$P_X(X > 1)$	More than one head
$P_X(X < 3)$	Not all heads
$P_X(X \leq 1)$	At least one head

Support/Range **Definition 2.1.1**

The set of all possible values of a random variable X:

$$\mathbb{X} \equiv supp(X) \equiv X(S) = \{x \in \mathbb{R} | \exists s \in S . X(s) = x\}$$

As S contains all possible experiment outcomes, $supp(X)$ contains all possible values/outcomes for the random variables X.

$$P_X(X \leq x) \text{ is defined for all } x \in supp(X)$$

2.2 Cumulative Distributions

Cumulative Distribution Function (F_X) **Definition 2.2.1**

The cumulative distribution function (cfd) of a random variable X is the probability where X takes some value less than or equal to some x:

$$F_X : \mathbb{R} \to [0,1] \text{ such that } F_X(x) = P_x(X \leq x)$$

To be a valid cfd, 3 criteria must be met:

1. **Probability between 0 and 1** $\forall x \in \mathbb{R}. 0 \leq F_X(x) \leq 1$
2. **Monotonicity** $\forall x_1, x_2 \in \mathbb{R} x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$
3. **Infinite Bounds** $F_X(-\infty) = 0, F_X(\infty) = 1$

For any random variable a *cfd* is right-continuous (a result of monotonicity).

$$x_1 > x_2 > x_3 ... > x \Rightarrow F_X(x_1) >= F_X(x_2) >= ... >= F_X(x)$$

We can determine the probability over finite intervals using the cumulative distribution:

$$\text{for } (a,b] \subseteq \mathbb{R} \; P_X(a < X \leq b) = F_X(b) - F_X(a)$$

Distributions

Probability Mass Function (p_X) **Definition 2.2.2**

Also called *probability function* gives the probability that a discrete random variable is exactly equal to a value.

The sample space S is mapped onto elements in the *support* of X (one-to-one).

We can then partition the sample space into a countable, disjoint collection od event subsets:

$$s \in E_i \Leftrightarrow X(s) = x_i, i = 1, 2 \dots$$

A probability mass function is valid if and only if:

1. **No negative probabilities** $\forall x \in supp(X).\ p_X(x) \geq 0$
2. **Probabilities sum to 1** $\sum_{x \in supp(x)} p_X(x) = 1$

2.3 Discrete Random Variable

For a *discrete random variable* we define the probability mass function as:

$$p_X(x_i) = P(X = x_i) = P(E_i) \text{ where } x_i \in supp(X) \text{ and } x_i \text{ is the outcome of event } E_i$$

We can also define using *cfds*:

$$F_X(x_i) = \sum_{j=1}^{i} p_X(x_j) \Leftrightarrow p_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \;\; \text{where } i = 2, 3 \dots$$

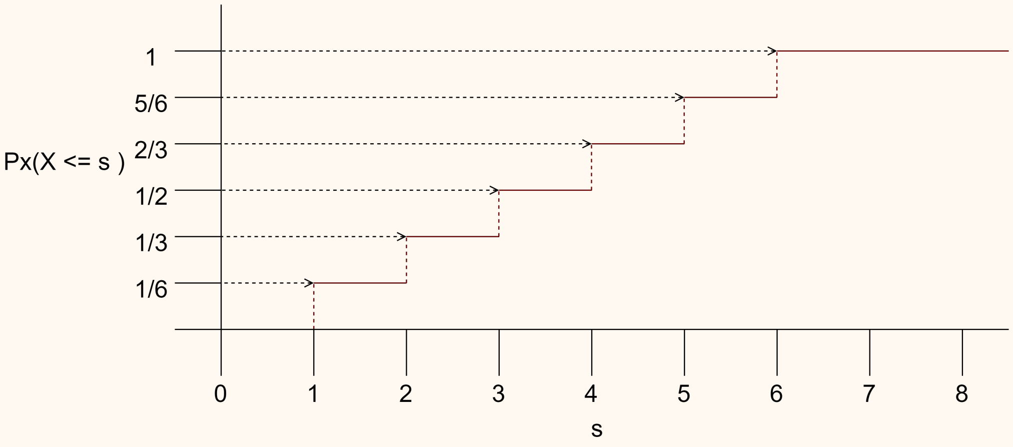
Or more simply:

$$p_X(x_i) = P_X(X = x_i) = P(X \leq x_i) - P(X \leq x_{i-1}) = F_X(x_i) - F_X(x_{i-1})$$

When graphed, F_X is a monotonically increasing, stepped function with jumps at points in $S(X)$.

Six Sided Dice **Example Question 2.3.1**

Here we have X representing the value of the dice roll. We can plot the cumulative distribution (showing probability a dice roll is less than or equal to a given value).



Discrete CFDs have several properties:

- **Limiting Cases**
$$\lim_{x\to-\infty} F_X(x) = 0 \quad \lim_{x\to\infty} F_X(x) = 1$$
At ∞ the whole set of outcomes is covered, probabilities sum to 1. At $-\infty$ none are covered.
- **Continuous from the right**
$$\text{For } x \in \mathbb{R} \lim_{h\to 0^+} F_X(x+h) = F_X(x)$$
Moving from the right to the left the probability will reduce and tend towards the value.
- **Non-Decreasing**
$$a < b \Rightarrow F_X(a) \leq F_X(b)$$
As it is cumulative, the value can only grow larger moving right.
- **Can cover a range**
$$\text{For } a < b.\ P(a < X \leq b) = F_X(b) - F_X(a)$$

Poisson Distribution — Definition 2.3.1

A discrete probability distribution expressing the probability of a given number of events occuring in a fixed time interval, given a constant mean.

$$Pois(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{where } k \text{ is the number of occurrences}$$

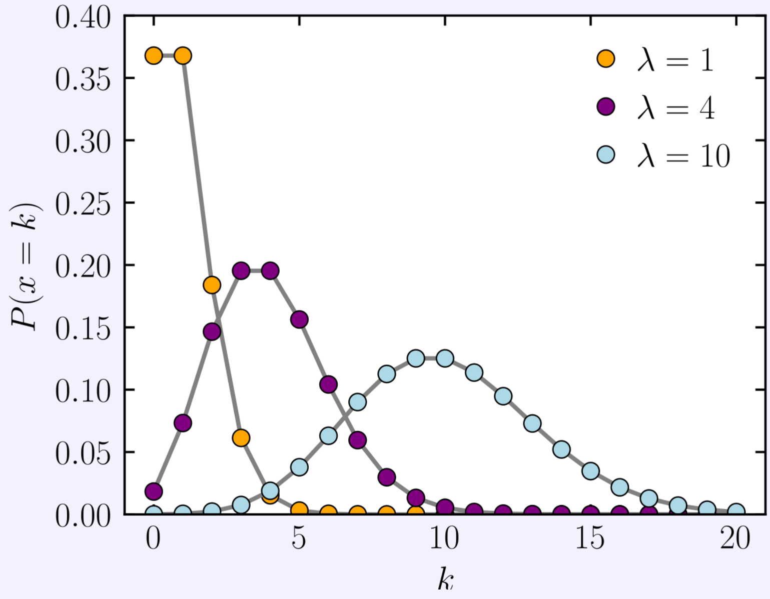
e.g What is the probability exactly 7 people buy pizzas at a stall in one hour, given on average is 4 people per hour?

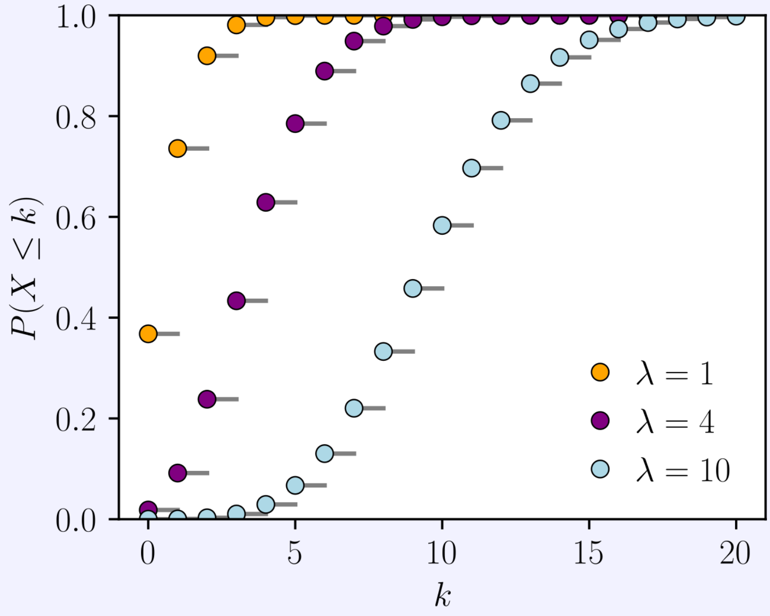
$$X \approx Poisson(4)$$

For a poisson distribution the mean (expected) and variance are equal.

$$E(X) = Var(X)$$

$$P(X = 7) = \frac{4^7 e^{-4}}{7!}$$





2.4 Link with Statistics

We can consider a set of data as realisations of a random variable defined on some underlying population of the data.

- Frequency histogram is an empirical estimate for the *pmf*.
- Cumulative histogram is an empirical estimate of the *cdf*.

2.5 Expectation

Expected Value **Definition 2.5.1**

The expectation of a *discrete random variable* X is:

$$E_X(X) = \sum_x xp(x)$$

Also referred to as μ_X it is the mean value of the distribution.

$$E(g(X)) = \sum_x g(x)p_X(x)$$

$$E(a \times X + b) = a \times E(X) + b$$

$$E(a \times g(X) + b \times f(X)) = a \times E(g(X)) + b \times E(f(X))$$

Given another distribution Y:

$$E(X + Y) = E(X) + E(Y)$$

Dice Rolls **Example Question 2.5.1**

Given random variable X representing the value of a dice roll:

$$X(n) = n \text{ where } 1 \geq n \geq 6$$

$$P(X = x) = \begin{cases} {}^1\!/\!{}_6 & 1 \geq n \geq 6 \\ 0 & otherwise \end{cases}$$

We can get the expected as:

$$E(X) = {}^1\!/\!{}_6 \times 1 + {}^1\!/\!{}_6 \times 2 + {}^1\!/\!{}_6 \times 3 + {}^1\!/\!{}_6 \times 4 + {}^1\!/\!{}_6 \times 5 + {}^1\!/\!{}_6 \times 6 = {}^{21}\!/\!{}_6 = 3.5$$

We can base scoring on the dice roll:

$$score(x) = 4 \times x + 2$$

Hence we can calculate that the expected score is $E(score(X)) = 4 \times 3.5 + 2 = 16$.

Dice and Coins **Example Question 2.5.2**

Given random variable D of a fair dice, and fair coin C:

$$P(D = x) = \begin{cases} {}^1\!/\!{}_6 & 1 \geq n \geq 6 \\ 0 & otherwise \end{cases} \text{ and } P(C = x) = \begin{cases} {}^1\!/\!{}_2 & x \in \{H, T\} \\ 0 & otherwise \end{cases}$$

Given $score = dice\ roll + 1$ if coin flip is heads what is the expected score?

$$E(D) = 3.5 \quad E(C) = 0.5 \quad E(score) = 3.5 + 2 * 0.5 = 4.5$$

2.6 Variance

Moment **Definition 2.6.1**

A function which measures the shape of a function's graph.

The n^{th} moment of a random variable is the expected value of its n^{th} power:

$$n^{th} \text{ moment of } X = \mu_X(n) = E(X^n) = \sum_x x^n p(x)$$

- **First Moment** The expected value.
- **Central Moment** The variance ($E[(X - E(X))^2]$)
- **Standardized Moment** The skew ($\frac{E(X-E(X))^3}{sd(X)^3}$)

Variance **Definition 2.6.2**

The expectation of the deviation from the expected/mean value squared.

$$Var(X) = Var_X(X) = \sigma_X^2 = E[(X - E(X))^2] = E(X^2) - (E(X))^2$$

Note that:

$$Var(a \times X + b) = a^2 Var(X)$$

Standard Deviation **Definition 2.6.3**

The square root of the variance.

$$\sigma_X = sd_X(X) = \sqrt{Var_X(X)}$$

Dice Roll **Example Question 2.6.1**

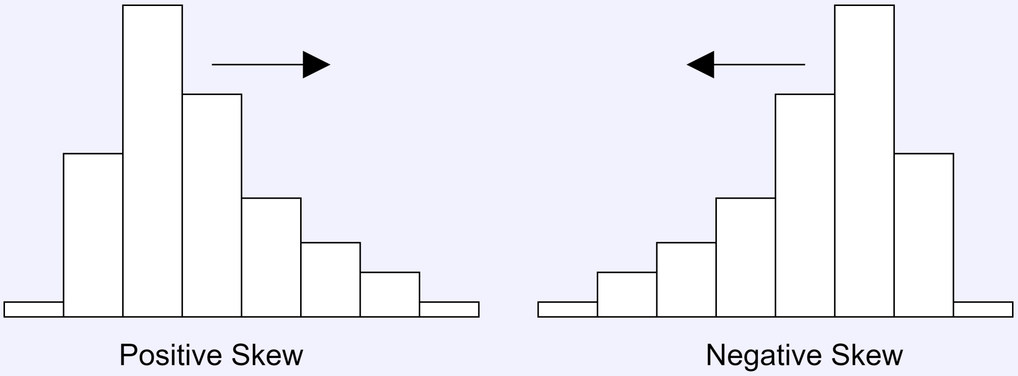
For a random variable representing a dice X:

$$Var(X) = E(X^2) - (E(X^2)) = \sum_x x^2 p(x) - (\sum_x xp(x))^2 = {}^{91}/_{6} - {}^{49}/_{4} = {}^{35}/_{12}$$

Skewness **Definition 2.6.4**

A measure of asymmetry (the standardized moment):

$$\gamma_1 = \frac{E(X-E(X))^3}{sd(X)^3} = \frac{E(X-\mu)}{\sigma^3} \text{ where } \mu = E(X), \sigma = Sd(X)$$



2.7 Sum of Random Variables

Given random variables $X_1, X_2, \dots, X_n$ (not necessarily independent, and potentially from different distributions), the sum is:

$$\text{The sum } S_n = \sum_{i=1}^{n} X_i \quad \text{and the average is } \frac{S_n}{n}$$

(The sum of the outcomes from all random variables)

The expected/mean value of S_n (expected value of the sum of all the random variables) is:

$$E(S_n) = \sum_{i=1}^{n} E(X_i) \quad \text{and} \quad E(\frac{S_n}{n}) = \frac{\sum_{i=1}^{n} E(X_i)}{n}$$

- **All independent**

$$Var(S_n) = \sum_{i=1}^{n} Var(X_i) \quad \text{and} \quad Var(\frac{S_n}{n}) = \frac{\sum_{i=1}^{n} Var(X_i)}{n^2}$$

- **All independent and Identically Distributed**
  Given that for all i, $E(X_i) = \mu_X$ and $Var(X_i) = \sigma_X^2$:

$$E(\frac{S_n}{n}) = \mu_X \quad \text{and} \quad Var(\frac{S_n}{n}) = \frac{\sigma_X^2}{n}$$

Important Discrete Random Variables

Bernouli Distribution **Definition 2.7.1**

For an experiment with only two outcomes, encoded as 1 and 0.

For$X \sim Bernoulli(p)$ where $x \in S(X) = \{0, 1\}$ and $0 \leq p \leq 1$:

PMF	*Expected*	*Variance*
$p_X(x) = p^x(1-p)^{1-x}$	$\mu = E(X) = p$	$\sigma^2 = Var(X) = p(1-p)$

Binomial Distribution **Definition 2.7.2**

Given n trials with two options, binomial models the number of outcomes. (e.g 3 coin tosses, number of ways to get 2 heads out of total outcomes).

For $X \sim Bionomial(n, p)$ where X takes values $0, 1, 2, \dots, n$ and $0 \leq p \leq 1$:

PMF	*Expected*	*Variance*	*Skewness*
$p_X(x) = \binom{n}{x} p^x(1-p)^{n-x}$	$\mu = E(X) = np$	$\sigma^2 = Var(X) = np(1-p)$	$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}$

Note that choice is: $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$

Poisson Distribution **Definition 2.7.3**

Given a constant mean number of events per fixed itme interval, provides probabilities of different numbers of events occuring. (e.g sell on average 6 cookies an hour, what is the probability 10 cookies are sold in a given hour).

For $X \sim Poisson(\lambda)$ where λ is the mean number of events and $\lambda > 0$:

PMF	*Expected*	*Variance*	*Skewness*
$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$	$\mu = E(X) = \lambda$	$\sigma^2 = Var(X) = \lambda$	$\gamma_1 = \frac{1}{\sqrt{\lambda}}$

Note that for poisson the skew is always positive (but decreases as λ increases), and $E(X) \equiv Var(X)$.

Geometric Distribution **Definition 2.7.4**

A potentially infinite number of trials to get an outcome (e.g attempts required to shoot a target, given probability of hit).

We can consider it infinite Bernoulli trials $X_1, X_2, \dots,$ where $X = \{i | X_i = 1\}$ (X is number of attempts to get outcome 1).

For $X \sim Geometric(p)$ where X takes all values in $\mathbb{Z}^+ = \{1, 2, \dots\}$ and $0 \leq p \leq 1$:

PMF	*Expected*	*Variance*	*Skewness*
$p_X(x) = p(1-p)^{x-1}$	$\mu = E(X) = \frac{1}{p}$	$\sigma^2 = Var(X) = \frac{1-p}{p^2}$	$\gamma_1 = \frac{2-p}{\sqrt{1-p}}$

Alternatively we can consider the number of trials *before* getting an outcome:
If $X \sim Geometric(P)$ consider $Y = X - 1$ where Y takes values $\mathbb{N} = \{0, 1, 2, \dots\}$:

PMF	*Expected*	*Variance*	*Skewness*
$p_Y(x) = p(1-p)^y$	$\mu = E(Y) = \frac{1-p}{p}$	Unchanged	Unchanged

Discrete Uniform Distribution **Definition 2.7.5**

Where a discrete number of outcomes are equally likely (e.g fair dice, colour wheel).

For $X \sim U(\{1, 2, \dots, n\})$:

PMF	*Expected*	*Variance*	*Skewness*
$p_X(x) = \frac{1}{n}$	$\mu = E(X) = \frac{n+1}{2}$	$\sigma^2 = Var(X) = \frac{n^2-1}{12}$	$\gamma_1 = 0$

2.8 Poisson Limit Theorem

We can use the *Binomial Distribution* to approximate the *Poisson Distribution*:

$$Poisson(\lambda) \approx Binomial(n, p) \text{ when } \lambda = np \text{ and } n \text{ is very large, } p \text{ is very small}$$

This is as for a *Poisson distribution* mean and variance are equal and for binomial, mean is np and variance $np(1-p)$ so as p gets smaller (and n larger) $np \approx np(1-p)$.

Chapter 3

Continuous Random Variables

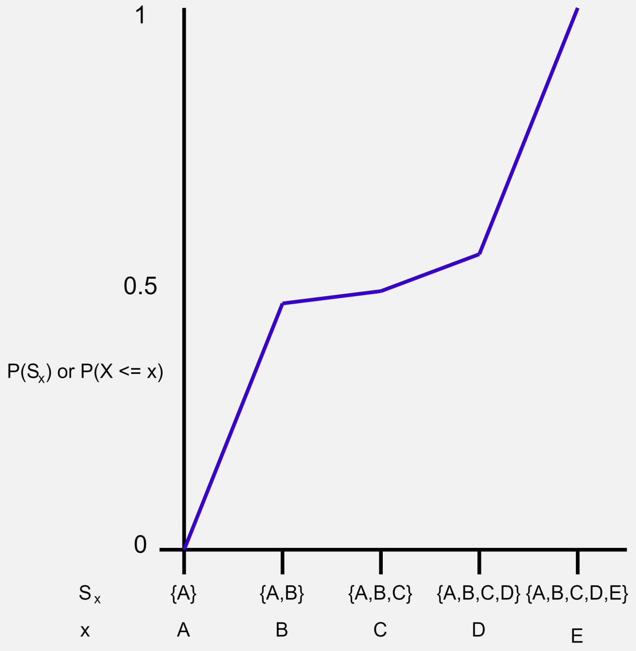
For continuous random variables we want to track quantities in $\mathbb{R}$ (e.g temperature, volume, other probabilities).

Induced Probability Terms ***Extra Fun!*** **3.0.1**

$$S_x = \{s \in S | X(s) \leq x\}$$

$$P_X(X \leq x) = P(S_x) = F_X(x)$$

S_x is the elements of the sample space up to and including x. Hence the probability of getting S_x is the cumulative probability.



Probability Density Function — Definition 3.0.1

For a random variable $X : S \to \mathbb{R}$ the induced probability is defined as:

$$P_X((-\infty, x]) = P(S_X) = F_X(x)$$

A variable X is *absolutely continuous* if $\exists f_X : \mathbb{R} \to \mathbb{R}$ such that:

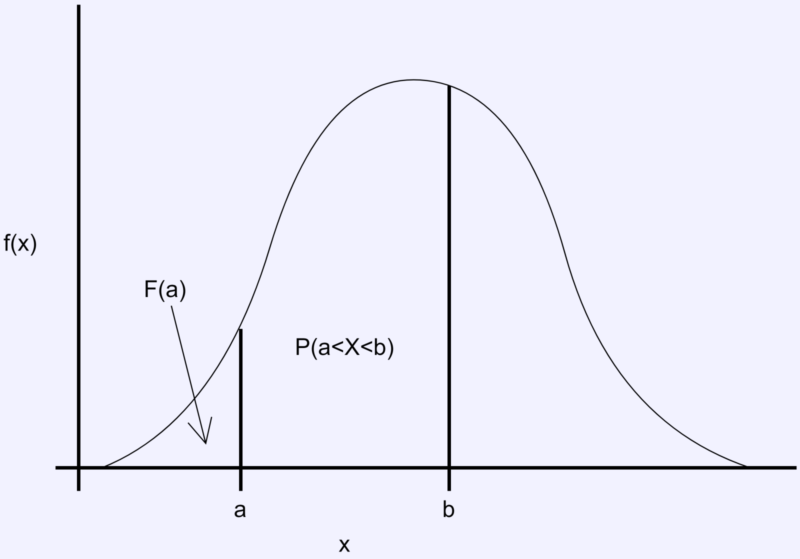
$$F_X(x) = \int_{u=-\infty}^{x} f_X(u)du$$

$$f_x(x) = F'(x) = \frac{d}{dx}F_X(x)$$

Where f_X is the *probability density function* (*pdf*).

To find probability that $X \in (a, b]$:

$$P_X(a < X \leq b) = P_X(X \leq b) - P_X(X \leq a) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$$



- We can use $<$ and $\leq$ interchangeably as $P(X = x) = 0 \Leftrightarrow P(X \leq x) \equiv P(X < x)$.
- Probability of any event is zero: $P_X(X = y) = 0$, any elementary event $\{x\}$ where $x \in \mathbb{R}$ has zero probability.
- However the sum of a range of events probabilities is not zero.
- Hence the range of a continuous random variable is uncountable (i.e as $\mathbb{R}$ is also).

$$\forall x \in \mathbb{R}. f_X(x) >= 0 \;\text{ and }\; \int_{-\infty}^{\infty} f_X(x)d_x = 1$$

Defining a continuous random variable — Example Question 3.0.1

Given some continuous random variable x with a probability density function given as:

$$f(x) = \begin{cases} cx^2 & 0 < x < 3 \\ 0 & otherwise \end{cases}$$

For some unknown constant c

To find the value of c we use the requirement that the cumulative distribution must sum to 1:

$$\int_0^3 cx^2 = 1 \rightsquigarrow [\frac{cx^3}{3}]_0^3 = 1 \rightsquigarrow (9c) - 0 = 1 \rightsquigarrow c = {}^1\!/\!_9$$

Hence:

$$f(x) = \begin{cases} \frac{x^2}{9} & 0 < x < 3 \\ 0 & otherwise \end{cases}$$

Hence we can specify the cumulative probability distribution as:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x^3}{27} & 0 < x < 3 \\ 1 & x \geq 0 \end{cases}$$

We can then calculate probabilities using the cumulative distribution:

$$P(1 < X < 2) = F(2) - F(1) = \frac{2^3}{27} - \frac{1^3}{27} = \frac{7}{27} \approx 0.259$$

3.1 Mean, Variance and Quantiles

Expected (Continuous) **Definition 3.1.1**

The *mean* or *expected* of a continuous random variable X:

$$\mu_X = E_X(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

For a function of interest that is applied to the random variable $g : \mathbb{R} \to \mathbb{R}$:

$$E_X(g(X)) \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- $E(aX + b) = aE(X) + b$
- $E(g(X) + h(X)) = E(g(X)) + E(h(X))$

Variance (Continuous) **Definition 3.1.2**

The variance of a continuous random variable X:

$$\sigma_X^2 = Var_X(X) = E((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

We can show this as:

$$\begin{aligned} Var_X(X) &= \int_{-\infty}^{\infty} x^2 f_X(x) ds - \mu_X^2 \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

For a linear transformation:

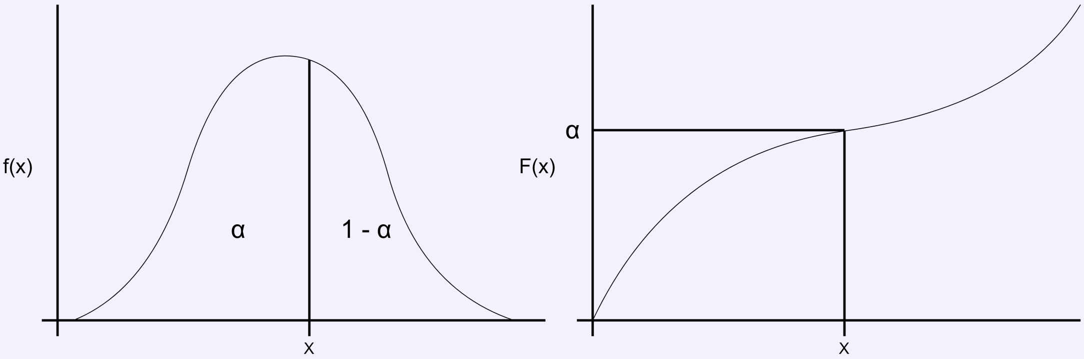
$$Var(aX + b) = a^2 Var(X)$$

Quartiles — Definition 3.1.3

The lower, upper quartiles and median are points

For a continuous random variable X, we define the α-*Quantile* $Q_X(\alpha)$ where $0 \leq \alpha \leq 1$ as the lowest X such that:

$$P(X \leq Q_X(\alpha)) = \alpha \text{ or in other words } Q_X(\alpha) = F_X^{-1}(\alpha)$$



Using Q_X we can define some standard quantiles:

- **Quartiles** Lower Quartile ($\alpha = 1/4$), Median ($\alpha = 1/2$) and Upper Quartile ($\alpha = 3/4$)
- **Percentiles** The nth percentile: $\alpha = \dfrac{n}{100}$

Basic continuous random variable — Example Question 3.1.1

Given continuous random variable X:

$$f(x) = \begin{cases} \frac{x^2}{9} & 0 < x < 3 \\ 0 & otherwise \end{cases}$$

We can calculate the expected:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{-\infty}^{0} xf(x)dx + \int_{0}^{3} xf(x)dx + \int_{3}^{\infty} xf(x)dx \\ &= \int_{-\infty}^{0} x \times 0dx + \int_{0}^{3} xf(x)dx + \int_{3}^{\infty} x \times 0dx \\ &= \int_{0}^{3} xf(x)dx = \int_{0}^{3} \frac{x^3}{9}dx = \left[\frac{x^4}{36}\right]_0^3 \\ &= \frac{9}{4} = 2.25 \end{aligned}$$

We can calculate the variance:

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} x^2f(x)dx - \mu_X^2 \\ &= \int_{-\infty}^{0} x^2f_(x)dx + \int_{0}^{3} x^2f(x)dx + \int_{3}^{\infty} x^2f(x)dx - \mu_X^2 \\ &= \int_{0}^{3} x^2f(x)dx - \mu_X^2 = \int_{0}^{3} \frac{x^5}{9}dx - \mu_X^2 \\ &= 27 - \mu_X^2 = 27 - 2.25 = 24.75 \end{aligned}$$

we can calculate the median, we ignore the range $x > 3$ as the median must be below this.

$$0.5 = \int_{-\infty}^{x} f(y)dy = \int_{-\infty}^{0} f(y)dy + \int_{0}^{x} f(y)dy = \int_{0}^{x} f(y)dy$$
$$0.5 = \int_{0}^{x} \frac{y^2}{9} = \left[\frac{y^3}{27}\right]_0^x = \frac{x^3}{27}$$
$$x = \sqrt[3]{0.5 \times 27} \approx 2.38$$

3.2 Notable Continuous Distributions

Continuous Uniform Distribution **Definition 3.2.1**

A continuous random variable with equal probability of being any value within a range:

For $X \sim U(a, b)$:

PDF	CDF	Expected	Variance
$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & otherwise \end{cases}$	$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$	$\mu = \frac{a+b}{2}$	$\sigma^2 = \frac{(b-a)^2}{12}$

The standard uniform distribution is defined as $X \sim U(0, 1)$:

PDF	CDF	Expected	Variance
$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & otherwise \end{cases}$	$F_X(x) = \begin{cases} 0 & x \leq a \\ x & a < x < b \\ 1 & x \geq b \end{cases}$	$\mu = 1/2$	$\sigma^2 = 1/12$

Other uniform distributions can be mapped linearly to the standard uniform.

Mapping to Standard Uniform **Example Question 3.2.1**

Given $X \sim U(2, 5)$ find the expected, variance and median.

Take $Y \sim U(0, 1)$, $X = 3 \times Y + 2$.

Distribution	Expected	Variance	Median
Y	0.5	$1/12$	0.5
X	3.5	$3/4$	3.5

Exponential Distribution — Definition 3.2.2

Given a rate of events λ, what is the probability of waiting X time for the event to occur.

For $X \sim \mathit{Exponential}(\lambda)$ or $X \sim \mathit{Exp}(\lambda)$ where $\lambda > 0$:

PDF	**CDF**	**Expected**	**Variance**
$f_X(x) = \lambda e^{-\lambda x}$ where $x \geq 0$	$F_X(x) = 1 - e^{-\lambda x}$ where $x \geq 0$	$\mu_X = \frac{1}{\lambda}$	$\sigma^2 = \frac{1}{\lambda^2}$

The distribution has the *Lack of memory property*, namely the time waited already does not affect the next part of the distribution (same shape).

$$P(X > x + t | X > t) = \frac{P(X > x + t \cap X > t)}{P(X > t)} = \frac{P(X > x + t)}{P(X > t)} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} = e^{-\lambda x} = P(X > x)$$

$$P(X > x + t | X > t) = P(X > x)$$

This distribution can be combined with Poisson. Given $X \sim \mathit{Poisson}(\lambda)$ (events occurring in a given time frame), the time between events is modelled by $X \sim \mathit{Exponential}(\lambda)$ (interval time for one event).

There is a variant with θ as the parameter for the distribution where $\theta = \frac{1}{\lambda}$.

Normal Distribution — Definition 3.2.3

Given a mean value (μ) and a variance (σ^2) from the mean the symmetrical distribution is a *Normal Distribution*.

For $X \sim \mathit{Normal}(\mu, \sigma^2)$ or $X \sim N(\mu, \sigma^2)$ where $\sigma > 0$:

PDF	**CDF**
$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt$

The *Standard/Unit Normal Distribution* is $X \sim N(0, 1)$:

PDF	**CDF**
$\phi(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}x^2\right\}$	$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$

We can apply linear functions:

$$X \sim N(\mu, \sigma^2) \to \text{ and } aX + b \sim N(a\mu + b, a^2\sigma^2)$$

Hence we can use the *Standard Normal Distribution*:

$$X \sim N\mu, \sigma^2 \Rightarrow \frac{X - \mu}{\sigma} \sim N(0, 1) \text{ and hence } P(X \leq x) = \Phi(\frac{x - \mu}{\sigma})$$

Lognormal Distribution — Definition 3.2.4

Given $X \sim N(\mu, \sigma^2)$ and $Y = e^X$ we can compute the *PDF* of Y:

$$f_Y(y) = \frac{1}{\sigma y\sqrt{2\pi}} exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right]$$

3.3 Central Limit Theorem

Moment Generating Function **Definition 3.3.1**

The moment generating function M_X for a continuous random variable X is:

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$$

Assuming the calculus within the $E(\dots)$ is valid, the nth moment is given by:

$$E[X^n] = \left.\frac{d^n M_x(t)}{dt^n}\right|_{t=0}$$

If the integral does not exist, the *characteristic function* $\phi_X(t) = M_X(\iota t)$ can be used (ι is imaginary unit).

Expected and Variance **Example Question 3.3.1**

$$\begin{aligned}
E[X] &= \left.\frac{dM_x(t)}{dt}\right|_{t=0} \\
&= \left.\frac{dE[e^{tX}]}{dt}\right|_{t=0} \\
&= \left.\frac{d\int_{-\infty}^{\infty} e^{tx} f_X(x)dx}{dt}\right|_{t=0} \\
&= \left.\int_{-\infty}^{\infty} xe^{tx} f_X(x)dx\right|_{t=0} \\
&= \int_{-\infty}^{\infty} xe^{0x} f_X(x)dx \\
&= \int_{-\infty}^{\infty} xf_X(x)dx
\end{aligned}$$

$$\begin{aligned}
E[X^2] &= \left.\frac{d^2 M_x(t)}{dt^2}\right|_{t=0} \\
&= \left.\frac{d^2 E(e^{tX})}{dt^2}\right|_{t=0} \\
&= \left.\frac{d^2\int_{-\infty}^{\infty} e^{tx} f_X(x)dx}{dt^2}\right|_{t=0} \\
&= \left.\frac{d\int_{-\infty}^{\infty} xe^{tx} f_X(x)dx}{dt}\right|_{t=0} \\
&= \left.\int_{-\infty}^{\infty} x^2 e^{tx} f_X(x)dx\right|_{t=0} \\
&= \int_{-\infty}^{\infty} x^2 e^{0x} f_X(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)dx
\end{aligned}$$

$$Var[X] = E[X^2] - (E[X])^2$$

3.4 Product of Random Variables

Given independent random variables $Z_1, Z_2, \ldots, Z_n$:

$$E[\prod_{i=1}^{n} Z_i] = \prod_{i=1}^{n} E[Z_i]$$

The sum of the random variables is the products of their *Moment Generating Functions.*

$$M_{Z_1+Z_2}(t) = E[e^{t(Z_1+Z_2)}] = E[e^{tZ_1}e^{tZ_2}] = E[e^{tZ_1}]E[e^{tZ_2}] = M_{Z_1}(t)M_{Z_2}(t)$$

$$S_n = \sum_{i=1}^{n} Z_i \Rightarrow M_{S_n}(t) = \prod_{j=1}^{n} MX_j(t)$$

3.5 Central Limit Theorem

Central Limit Theorem **Definition 3.5.1**

Given $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables from any distribution with mean μ and finite variance σ^2.

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

Hence we have a distribution with a known expected and variance, so can form a *Normal Distribution.*

$$\begin{array}{rcc} Y = S_n & E(Y) = n\mu & Var(Y) = n\sigma^2 \\ Y = S_n - n\mu & E(Y) = 0 & Var(Y) = n\sigma^2 \\ Y = \dfrac{S_n - n\mu}{\sqrt{n}\sigma} & E(X) = 0 & Var(X) = 1 \end{array}$$

Y can now be used to approximate a *Standard Normal Distribution.*

$$\lim_{n\to\infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

This implies that for large (but finite n):

$$\overline{X} \approx N(\mu, \frac{\sigma^2}{n}) \text{ and } \sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$$

Where $\overline{X}$ is the average value of the random variables $\dfrac{\sum_{i=0}^{n} X_i}{n}$.

The approximation holds for all distributions (including discrete), and is exact when the random variables are from the same *normal distribution.*

3.5.1 An attempt at CLT proof

Given the random variables $X_1, X_2, \ldots, X_n$ we can standardize and get their sum:

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma} = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}\sigma} \text{ where } Y_i = X_i - \mu$$

The moment generating function of Z_n is the product of the *moment generating functions* of the Y (all identically distributed, so identical *MGFs*).

$$M_{Z_n}(t) = \left(M_Y\left(\frac{t}{\sqrt{n}\sigma}\right)\right) \text{ where } M_y \text{ is the moment generating function for all } Y_i$$

We can then expand the M_Y around 0 using Taylor's Theorem:

$$M_Y(t) = M_Y(0) + M_Y'(0)t + {}^1\!/\!{}_2 M_Y''(0)t^2 + O(t^3)$$

$O(t^3)$ is the error term of our approximation, as this is for higher powers, it has a small effect so can be ignored

The derivatives of the MFG are:

$$M_Y'(0) = E(Y_i) = 0 \text{ due to shift performed earlier and } M_Y''(0) = E(Y_i^2) = \sigma^2 + E(Y_i)^2 == \sigma^2 + 0 = \sigma^2$$

Hence we can derive:

$$M_Y(t) = 1 + \frac{\sigma^2 t^2}{2} + O(t^3)$$

Hence we can scale t, and ignore the error term for simplicity:

$$M_Y\left(\frac{t}{\sqrt{n}\sigma}\right) = 1 + \frac{t^2}{2n}$$

As the error term gets very small, we can use limits to get an approximation for $M_{Z_n}(t)$.

$$lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}(1 + \frac{t^2}{2n} + O(n^{-3/2}))^n = e^{t^2/2}$$

Note that $\lim_{m\to\infty}(1 + \frac{x}{m})^m = e^x$.

Coin Tossing **Example Question 3.5.1**

Consider a set of count tosses, each are Bernoulli discrete random variables (take values 0 or 1).

$$X_1, X_2, X_3, \dots, X_n \text{ where } \mu = p \text{ and } \sigma^2 = p(1-p)$$

The total score of toin tosses can be modelled as a binomail distribution:

$$\sum_{i=1}^{n} X_i \text{ is } X \sim Binomial(n,p) \text{ with } E(X) = np \text{ and } Var(X) = np(1-p)$$

For large n can also model it as a normal distribution:

$$\sum_{i=1}^{n} X_i \text{ is } X \sim N(n\mu, n\sigma^2) \equiv N(np, np(1-p))$$

As the number of events (coin tosses) tends to infinity, the distributions tend to look identical.

Chapter 4

Joint Random Distributions

4.1 CDF

Suppose we have random variables X and Y such that:

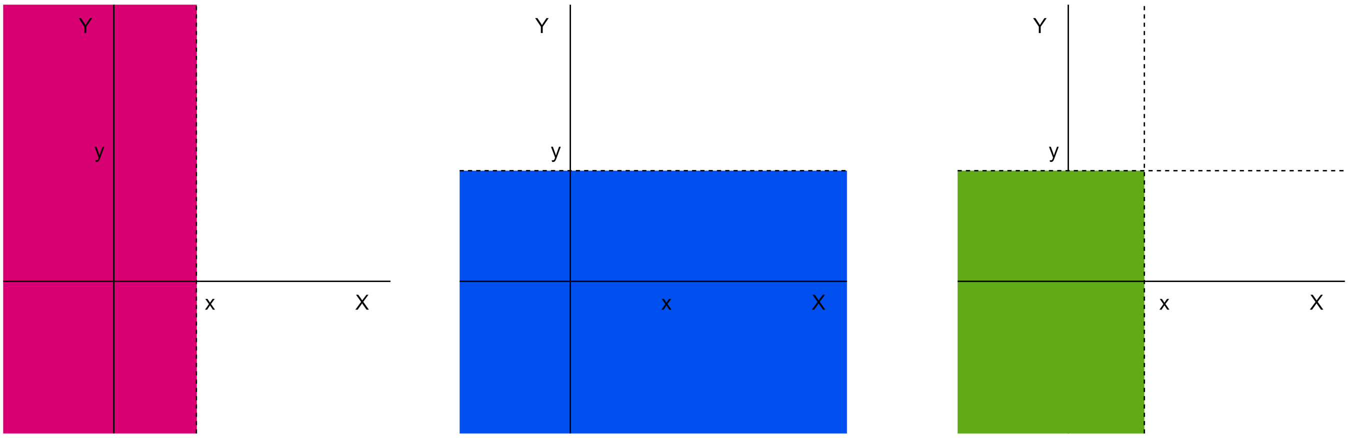
$$X : S_X \to \mathbb{R} \text{ and } Y : S_Y \to \mathbb{R}$$

We can define Z operating on sample space S such that:

$$S = S_1 \times S_2 \quad S = \{(s_X, s_Y) | s_X \in S_X \wedge s_Y \in S_Y\} \quad Z = (X, Y) : S \to \mathbb{R}^2$$

Hence we have a mapping from joint random variable $Z(s)$ onto $(X(s), Y(s))$.

We can consider this using a graph of the sample space:



Marginal of x

Marginal of y

x and y occur

Hence the induced probability function for Z will be:

$$F(x, y) = P_Z(X \leq x, Y \leq y) = P_Z((-\infty, x], (-\infty, y]) = P(S_{XY})$$

Hence we can use the marginals of the joint distribution to get the distribution of the two random variables:

$$F_X(x) = F(x, \infty) \text{ and } F_Y(y) = F(\infty, y)$$

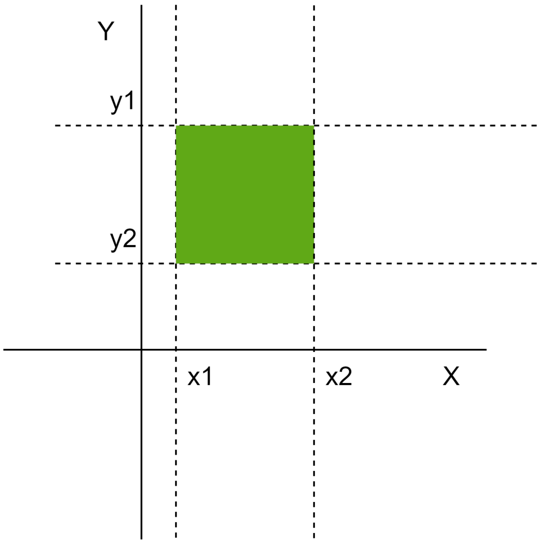
To be a valid *joint cumulative distribution function*:

- $\forall x, y \in \mathbb{R}.\ 0 \leq F(x, y) \leq 1$
- **Monotonicity**

$$\forall x_1, x_2, y_1, y_2 \in \mathbb{R}.\ [x_1 < x_2 \Rightarrow F(x_1, y_1) \leq F(x_2, y_1) \wedge y_1 < y_2 \Rightarrow F(x_1, y_1) \leq F(x_1, y_2)]$$

- $\forall x, y \in \mathbb{R}.\ F(x, -\infty) = F(-\infty, y) = 0$
- $F(\infty, \infty) = 1$

For the probability of intervals we can use the graph mapping concept again:



$$P_Z(x_1 < X \leq x_2, Y \leq y) = F(x_2, y) - F(x_1, y)$$

Hence we can get the interval:

$$P_Z(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$

4.2 PMF

Joint Probability Mass Function **Definition 4.2.1**

$$p(x, y) = P_Z(X = x, Y = y) \text{ where } x, y \in \mathbb{R}$$

We can get the original *pmfs* of the two variables as:

$$p_X(x) = \sum_y p(x, y) \quad \text{and} \quad p_Y(y) = \sum_x p(x, y)$$

To be a valid *pmf*:

- $\forall x, y \in \mathbb{R}.\ 0 \leq p(x, y) \leq 1$
- $\sum_y \sum_x p(x, y) = 1$

4.3 PDF

Fundamental Theorem of Caculus ***Extra Fun!*** **4.3.1**

The fundamental law that integration and differentiation and the inverse of each other (except for constant added in integration c, which does not affect definite integrals).

Joint Probability Density Function — Definition 4.3.1

When the variables being *joined* are continuous we have $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$, in this case:

$$F(x,y) = \int_{a=-\infty}^{y} \int_{b=-\infty}^{x} f(b,a)\ db\ da$$

The sum of the probability density function from $(x,y) \to (-\infty, -\infty)$

Hence by the fundamental theorem of calculus:

$$f(x,y) = \frac{\sigma^2}{\sigma x \sigma y} F(x,y)$$

We can differentiate to go get the PMF from the PDF.

To be valid:

- $\forall x, y \in \mathbb{R}. f(x,y) \geq 0$
- $\int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} f(x,y)\ dx\ dy = 1$

Marginal Density Functions — Definition 4.3.2

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) = \frac{d}{dx} F(x, \infty) \\ &= \frac{d}{dx} \int_{y=-\infty}^{\infty} \int_{s=-\infty}^{x} f(s,y)\ ds\ dy \end{aligned}$$

And likewise for y:

$$f_Y(y) = \frac{d}{dy} \int_{x=-\infty}^{\infty} \int_{s=-\infty}^{y} f(x,s)\ ds\ dx$$

Hence by applying the fundamental theorem of calculus:

$$f_X(x) = \int_{y=-\infty}^{\infty} f(x,y)\ dy$$

$$f_Y(y) = \int_{x=-\infty}^{\infty} f(x,y)\ dx$$

Marginal pdf — Example Question 4.3.1

Given continuous variables $(X,Y) \in \mathbb{R}^2$:

$$f(x,y) = \begin{cases} 1 & |x| + |y| < \frac{1}{\sqrt{2}} \\ 0 & otherwise \end{cases}$$

To determine the marginal pdfs for X and Y:

First notice that: $|x| + |y| < \frac{1}{\sqrt{2}} \Leftrightarrow |y| < \frac{1}{\sqrt{2}} - |x|$.

Hence given an x we can see that for the first case of the probability density function to match, y must be between:

$$\frac{1}{-\sqrt{2}} + |x| < y < \sqrt{2} - |x|$$

$$\begin{aligned} f_X(x) &= \int_{y=-\infty}^{\infty} f(x,y)\ dy \\ &= \int_{y=-\sqrt{2}+|x|}^{\sqrt{2}-|x|} 1\ dy \\ &= [y]_{-\sqrt{2}+|x|}^{\sqrt{2}-|x|} \\ &= \left(\sqrt{2}-|x|\right) - \left(-\sqrt{2}+|x|\right) \\ &= 2\sqrt{2}-2|x| \end{aligned}$$

Similarly for y:

$$f_Y(y) = 2\sqrt{2}-2|y|$$

Multinomial Distribution — Definition 4.3.3

Given:

- sequence of n independent and identical experiments (all same distribution, same parameters).
- r possible outcomes for each experiment.
- Each probability q_i is the probability of outcome i.
- The sum of all probabilities for the outcomes is 1: $\sum_{i=1}^{r} q_i = 1$

We can have a set of random variables where each X_i represents the number of experiments resulting in outcome i.

$$P(X_1 = n_1, X_2 = n_2, \dots, X_r = n_r) = \frac{n!}{n_1! \times n_2! \times \dots \times n_r!} \times q_1^{n_1} \times q_2^{n_2} \times \dots \times q_r^{n_r}$$

We know this as any sequence will have the probability $q_1^{n_1} \times q_2^{n_2} \times \dots \times q_r^{n_r}$ where $n_1 + n_2 + \dots + n_r = n$ (multiplying the probabilities in a sequence).

For a given number of outcomes, there are many different sequences like the above. We can determine the number of sequences as:

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\dots\binom{n-\sum_{i=1}^{r-1} n_i}{n_r} = \frac{n!}{n_1! \times n_2! \times \dots \times n_r!}$$

Party Politics — Example Question 4.3.2

Given 4 different political parties with popularities:

Party	**Polling Percentage**
Ingsoc	40%
Techno Union	20%
Norsefire	15%
Birthday Party	25%

If asking 10 people of what party they prefer, what is the probability that:

- 2 support Ingsoc
- 4 support the Techno Union
- 1 supports Norsefire
- 3 support the Birthday Party

$$P(X_{ingsoc} = 2, X_{techno-union} = 4, X_{norsefire} = 1, X_{birthday} = 3)$$

$$\frac{10!}{2! \times 4! \times 1! \times 3!} \times (0.4)^2 \times (0.2)^4 \times (0.15)^1 \times (0.25)^3$$

$$\frac{189}{25000} = 0.00756 = 0.756\%$$

4.4 Joint Conditional Random Variables

Given random variables X and Y:

$$\text{variables independent } \Leftrightarrow F(x,y) = F_X(x)F_Y(y)$$

(For both continuous and discrete)

More specifically:

$$\begin{array}{rll} \text{For Discrete Variables} & p(x,y) = p_X(x)p_Y(y) & \text{(probability mass function)} \\ \text{For Continuous Variables} & f(x,y) = f_X(x)f_Y(y) & \text{(Probability density function)} \end{array}$$

Diamond at origin **Example Question 4.4.1**

Consider *pdf*:

$$f(x,y) = \begin{cases} 1 & |x| + |y| < \frac{1}{\sqrt{2}} \\ 0 & otherwise \end{cases}$$

By the previous example:

$$f_X(x) = 2\sqrt{2} - 2|x|$$
$$f_Y(y) = 2\sqrt{2} - 2|y|$$

Hence as $f(x,y) \neq f_X(x)f_Y(y)$ and hence X and Y are not independent.

Independent variables **Example Question 4.4.2**

Given two continuous random variables X and Y:

$$f(x,y) = \lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y} \text{ given } x, y > 0$$

We can get the marginal *pdf* by integrating over all of y:

$$\begin{aligned} f(x) &= \int_{y=-\infty}^{\infty} f(x,y)dy \\ &= \int_{y=0}^{\infty} f(x,y)dy \\ &= \lim_{t\to\infty} \int_{y=0}^{t} \lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y} dy \\ &= \lim_{t\to\infty} \int_{y=0}^{t} \lambda_1\lambda_2 e^{-\lambda_1 x} \times e^{-\lambda_2 y} dy \\ &= \lim_{t\to\infty} \left[-\lambda_1 e^{-\lambda_1 x - \lambda_2 y}\right]_{y=0}^{y=t} \\ &= \lim_{t\to\infty} \left(-\lambda_1 e^{-\lambda_1 x - \lambda_2 t}\right) - \left(-\lambda_1 e^{-\lambda_1 x - \lambda_2 0}\right) \\ &= \lim_{t\to\infty} \left(-\lambda_1 e^{-\lambda_1 x - \lambda_2 t}\right) - \left(-\lambda_1 e^{-\lambda_1 x - \lambda_2 0}\right) \\ &= 0 - \left(-\lambda_1 e^{-\lambda_1 x}\right) \\ &= \lambda_1 e^{-\lambda_1 x} \end{aligned}$$

We can do the same for $f_Y(y)$ to get $\lambda_2 e^{-\lambda_2 y}$.

Hence the events are independent as:

$$\lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y} = \lambda_1 e^{-\lambda_1 x} \times \lambda_2 e^{-\lambda_2 y}$$

4.4.1 Conditional PMF

For discrete random variables we can define the joint *pmf* as:

$$p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)} \text{ where } \forall y. p_Y(y) > 0$$

Baye's Theorem **Definition 4.4.1**

Baye's theorem states that on some partition of the sample space S, $P_1, \dots P_k$:

$$P(X) = \sum_{i=1}^{k} P(X|E_i)P(E_i)$$

Given each partition the probability of some X occurring sums to the total probability of X occurring.

Using the conditional joint *pmf* we can also express this theorem (over a single partition) as:

$$p_{X|Y}(x|y) \times p_Y(y) = p_{Y|X}(y|x) \times p_X(x)$$

Conditional PMF Marginal Joint Probabilities **Definition 4.4.2**

$$p(x) = \sum_y p_{X|Y}(x|y)p_Y(y)$$

(Go through every y, summing the probability of x occurring with that y, multiplied by the probability of that y)

4.4.2 Conditional PDF

For continuous random variables we can define the joint *pdf* as:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

$$X \text{ and } Y \text{ independent } \Leftrightarrow \forall x, y \in \mathbb{R}.\ f_{X|Y}(x,y) = f_X(x)$$

And we can now have *bayes theorem* as:

$$f_{X|Y}(x|y) = \frac{f_{Y|X} f_X(x)}{f_Y(y)}$$

Conditional PDF Marginal Joint Probabilities **Definition 4.4.3**

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\ dy$$

and with the cumulative distribution:

$$F_X(x) = \int_{y=-\infty}^{\infty} F_{X|Y}(x|y) f_Y(y)\ dy$$

Independent exponential random variables **Example Question 4.4.3**

Given $X \sim Exp(\lambda)$ and $Y \sim Exp(\mu)$ what is $P(X < Y)$.

$$\begin{aligned}
P(X<Y) &= \int_{x<y} f(x,y)\ dx\ dy \\
&= \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{y} f(x,y)\ dx\ dy \text{ (go over all } y\text{s, for each take the } x\text{s that are less)} \\
&= \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{y} f_X(x)f_Y(y)\ dx\ dy \text{ (} X \text{ and } Y \text{ are independent)} \\
&= \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{y} f_X(x)f_Y(y)\ dx\ dy \text{ (} X \text{ and } Y \text{ are independent)} \\
&= \int_{y=-\infty}^{\infty} F_X(y)\times(\mu e^{-\mu y})\ dx\ dy \text{ (Integrate } f_X \text{ to get } F_X \text{ and then get all below } y\text{)} \\
&= \int_{y=-\infty}^{\infty} (1-e^{-\lambda y})\times(\mu e^{-\mu y})\ dx\ dy \text{ (Substitute definitions)} \\
&= \int_{y=0}^{\infty} (1-e^{-\lambda y})\times(\mu e^{-\mu y})\ dx\ dy \text{ (exponential cut at 0)} \\
&= \lim_{t\to\infty}\int_{y=0}^{t} (1-e^{-\lambda y})\times(\mu e^{-\mu y})\ dx\ dy \\
&= \lim_{t\to\infty}\int_{y=0}^{t} (\mu e^{-\mu y}) - e^{-\lambda y}\times(\mu e^{-\mu y})\ dx\ dy \\
&= \lim_{t\to\infty}\int_{y=0}^{t} (\mu e^{-\mu y}) - \mu e^{(-\lambda-\mu)y}\ dx\ dy \\
&= \lim_{t\to\infty}\left[-e^{-\mu y} + \frac{-\mu}{-\lambda-\mu}e^{(-\lambda-\mu)y}\right]_{y=0}^{y=t} \\
&= \lim_{t\to\infty}\left[-e^{-\mu y} + \frac{\mu}{\lambda+\mu}e^{(-\lambda-\mu)y}\right]_{y=0}^{y=t} \\
&= \lim_{t\to\infty}\left(-e^{-\mu t} + \frac{\mu}{\lambda+\mu}e^{(-\lambda-\mu)t}\right) - \left(-e^{\mu 0} + \frac{\mu}{\lambda+\mu}e^{(-\lambda-\mu)0}\right) \\
&= (0-0) - \left(-1 + \frac{\mu}{\lambda+\mu}\right) \\
&= 1 - \frac{\mu}{\lambda+\mu} = \frac{\lambda}{\lambda+\mu}
\end{aligned}$$

4.5 Expectation and Variance for Joint Random Variables

Joint Expectation **Definition 4.5.1**

Where g is a *bivariat function* on the random variables X and Y:

For *discrete variables*:

$$E(g(X,Y)) = \sum_y \sum_x g(x,y)p(x,y)$$

For *continuous variables*:

$$E(g(X,Y)) = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x,y)f(x,y)\ dx\ dy$$

Hence we have the following:

- **For all** $g(X,Y) = g_1(X) + g_2(Y) \Rightarrow E(g_1(X) + g_2(Y)) = E_X(g_1(X)) + E_Y(g_2(Y))$
- **If X and Y are independent** $E(g_1(X) \times g_2(Y)) = E_X(g_1(X))) \times E_Y(g_2(Y))$
  Hence where $g(X,Y) = X \times Y$ we have $E(XY) = E_X(X) \times E_Y(Y)$

Q

Covariance **Definition 4.5.2**

Covariance measures how two random variables change with respect to one another.

For a single random variable we consider expected value of the difference between the mean and the value, squared.

$$\text{Expectation of } g(X) = (X - \mu_X)^2 = \sigma_X^2$$

For a bivariate we consider the expectation:

$$\text{Expectation of } g(X,Y) = (X - \mu_X)(Y - \mu_Y)$$

We can then defined the covariance as:

$$\begin{aligned} \sigma_{XY} = Cov(X,Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY] - E_X[X] \times E_Y[Y] \\ &= E[XY] - \mu_X \mu_Y \end{aligned}$$

When X and Y are independent so:

$$\sigma_{XY} = Cov(X,Y) = E[XY] - E_X[X] \times E_Y[Y] = E[XY] - E[XY] = 0$$

Correlation **Definition 4.5.3**

Much like covariance, however is invariant to the scale of X and Y.

$$\rho_{XY} = Cor(X,Y) = \frac{\sigma_{XY}}{\sigma_X \times \sigma_Y}$$

If the variables are independent then $\rho_{XY} = \sigma_{XY} = 0$.

4.6 Multivariate Normal Distribution

Multivariate Normal Distribution **Definition 4.6.1**

Given a random vector $X = (X_1, \dots, X_n)$ with means $\mu = (\mu_1, \dots, \mu_n)$ has joint *pdf*:

$$f_X = \frac{1}{\sqrt{(2\pi)^n det \sum}} exp(-{}^1/{}_2(x-\mu)^T \sum^{-1} (x-\mu))$$

Where $\sum$ is the covariance matrix:

$$\sum_{(i,j)} = Cov(X_i, X_j) \;\; \text{where } 1 \leq i, j \leq n$$

The covariance matrix must be *positive-definite* for a *pdf* to exist Note that the random variables do not need to be independent.

Positive Definite real Matrices ***Extra Fun!*** **4.6.1**

$$M \text{ is positive-definite} \Leftrightarrow \forall x \in \mathbb{R}^{\ltimes} \backslash \{0\}.\ x^T M x > 0$$

4.7 Conditional Expectation

Conditional Expectation **Definition 4.7.1**

In general $E(XY) \neq E_X(X)E_Y(Y)$
For discrete random variables the *conditional expectation* of Y given that $X = x$ is:

$$E_{Y|X}(Y|x) = \sum_y y p_{y|X}(y|x)$$

For continuous random variables:

$$E_{Y|X}(Y|x) = \int_{y=-\infty}^{\infty} y f_{Y|X}(y|x)\ dy$$

In both cases the conditional expectation is a function of x and not Y. We are getting the weighted sum over all Ys, for a single value (x) of X.

Expectation of a Conditional Expectation **Definition 4.7.2**

We can define random variable W such that:

$$W = E_{Y|X}(Y|X)$$

W is effectively a function of the random variable $X : S \to \mathbb{R}$ by $W(s) = E_{Y|X}(Y|x)$ where $X(s) = x$.

Using this we can determine that:

$$E_Y(Y) = E_X(E_{Y|X}(Y|X))$$

(Expectation of Y is the same as the expectation function of X, of the expected value of Y given X)

This holds for both discrete and continuous.

$$\int_y \int_x y f_{Y|X}(y|x) f_X(x)\ dx\ dy = \int_y \int_x y f(x,y)\ dx\ dy = \int_y y f_Y(y)\ dy$$

Tower Rule **Definition 4.7.3**

The expectation of a conditional expectation rule extends to chains of expectations:

$$\begin{aligned} E(Y) &= E_{X_1}(E_Y(Y|X_1)) \\ &= E_{X_2}(E_{X_1}(E_Y(Y|X_1, X_2)|X_2)) \\ &= \dots \\ &= E_{X_n}(E_{X_{n-1}}(\dots E_{X_1}(E_Y(Y|X_1, \dots, X_n)|X_2, \dots, X_n) \dots |X_n)) \end{aligned}$$

This is a generalisation of the *partition rule* for conditional expectations.

4.8 Markov Chains

Discrete Time Markov Chain (DTMC) **Definition 4.8.1**

- A series of random variable modelling the state at a time step: $X_0, X_1, X_2, \dots$
- The state space J (all states), where $J = sipp(X_i)$ (contains all states that we can be in at any step)
- We can take a sequence (sample path) through the states $(X_0, X_1, X_2, \dots)$
- We denote the state taken at step n as state J_n

We use an initial probability vector π to determine the start state:

$$\pi_0 = [\dots \text{ probability of starting in state i } \dots]$$

We determine the probability of each next state through the transition probability matrix r:

$$r_{ij} = P(X_{n+1} = j | X_n = i)$$

For a markov chain the probability of being in any next state is **only** dependent on the current state (memoryless, history of previous states does not matter).

$$P(X_{i+1} = J_{n+1} | X_i = J_i) = P(X_{i+1} = J_{n+1} | X_i = J_i) = P(X_{i+1} = J_{n+1} | X_0 = J_0, \dots, X_i = J_i)$$

To get the probability we can use power of the matrix:

$$P(X_n = j | X_0 = i) = (R^n)_{ij}$$

If we have the initial probability vector we can calculate:

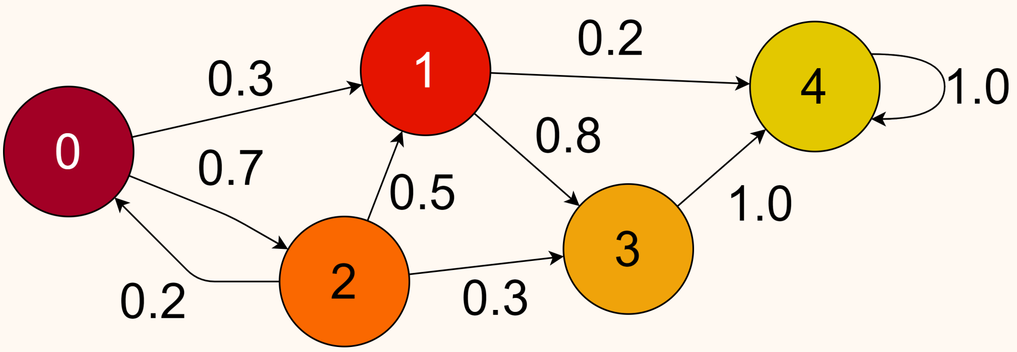
$$\begin{aligned} P(X_n = j) &= \sum_{i \in J} P(X_0 = i) \times P(X_n = j | X_0 = i) \\ &= \sum_{i \in j} \pi_{0_i} (R^n)_{ij} \\ &= (\pi_0 R^n)_{ij} \end{aligned}$$

We can obtain the long term probabilities by using the ∞th step:

$$\lim_{t \to +\infty} \pi_0 R^n = \pi_\infty$$

Note that since $\pi_\infty R = \pi_\infty$ we have eigenvector π_∞ and eigenvalue 1.

Probabilistic Finite State Machine **Example Question 4.8.1**



$J = \{\ 0 \quad 1 \quad 2 \quad 3 \quad 4\ \}$ State Space

$\pi_0 = [\ 1.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0\ ]$ Initial Probability Vector

100% chance we start at 0

To

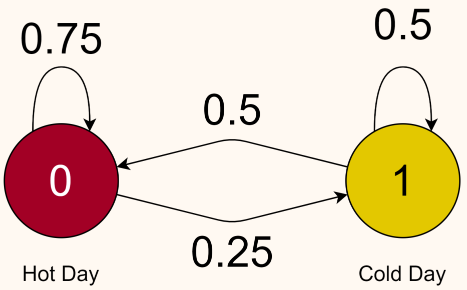
$$\text{From}\begin{bmatrix} \begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ 0 & 0.0 & 0.3 & 0.7 & 0.0 & 0.0 \\ 1 & 0.0 & 0.0 & 0.0 & 0.8 & 0.2 \\ 2 & 0.2 & 0.5 & 0.0 & 0.3 & 0.0 \\ 3 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 4 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{array} \end{bmatrix}$$

Transition Probability Matrix

P(X[i+1] = 1 | X[i] = 2) = 0.5

Can get permentantly stuck at state 4

Modelling Climate — Example Question 4.8.2



Transition Probability Matrix

$$\text{From}\quad \begin{array}{c|cc} & \multicolumn{2}{c}{\text{To}} \\ & 0 & 1 \\ \hline 0 & 0.75 & 0.25 \\ 1 & 0.50 & 0.50 \end{array}$$

$J = \{\ 0 \quad 1\ \}$ State Space

$\pi_0 = [\ 0.8 \quad 0.2\ ]$ Initial Probability Vector

Start probably on a hot day

$\pi_0 = [\ 0.0 \quad 1.0\ ]$

Always start on a cold day

...

Possible sample paths

...

0 1 **0** 1

0 0 **0** 0

1 0 **1** 1

...

↑

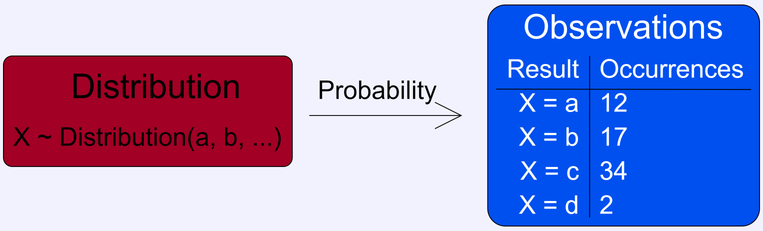
$P(X_2 = 1)$

Chapter 5

Statistics and Estimation

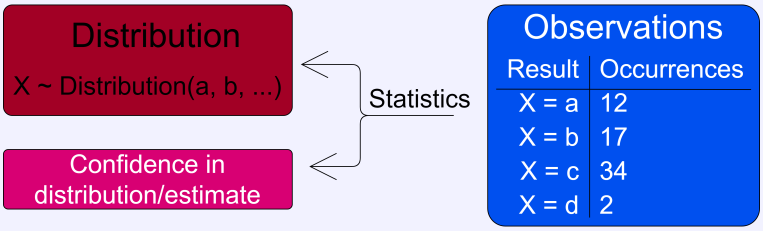
5.1 Statistics Terms

Probability **Definition 5.1.1**



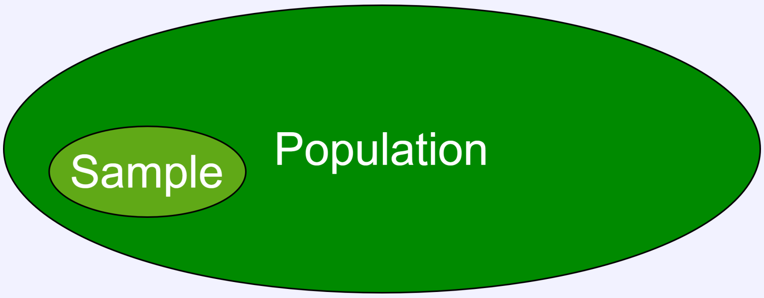
Deducing likelihood, and predicting events based on a known probability distribution.

Statistics **Definition 5.1.2**



Using empirical data/observations from an experiment to determine a probability distribution (and estimate its parameters) that models the observed distribution of results.

Sample — Definition 5.1.3



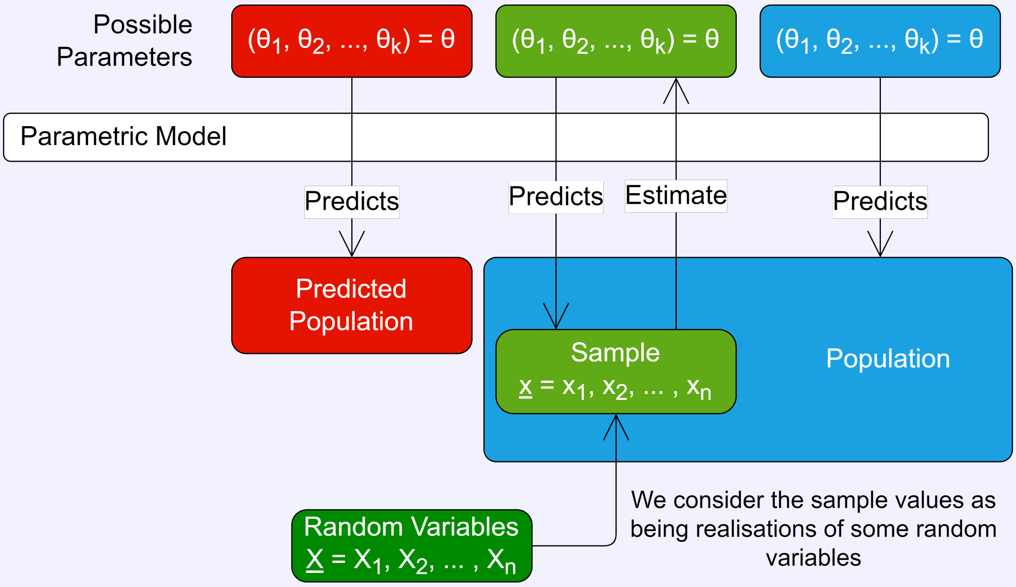
A subset of the population, from which we can use *statistical methods* to make inferences about the characteristics of an entire population.

- In vaccine trials, we can take a random sample as participants, and use there results to infer the possible efficacy of the vaccine over an entire population.
- In manufacturing we may want to test durability, but doing so may destroy the product. Hence we can take a small representative sample, and tests these to gain knowledge about the durability of all products from a given production line, without having to test all to destruction.
- In politics, we can use the political persuasions of a sample to reason about an entire population (such as electorate, or a given group) (polling).

Statistical Models — Definition 5.1.4

Models are a structure (e.g distribution) often developed from a sample that can be used to make inferences about a population.

- Models are usually *parametric*, meaning the models can be described entirely by its parameters.
- Models have a finite set of parameters.



- We can use distributions such as *Normal*, *Poisson*, *Bernoulli* etc. as parametric models.
- If the population is such that the probability of each outcome is $P_{X|\theta}(.|\theta)$ (probability of each is only dependent on parameters) we can assume the random variables $\underline{X}$ are independent and identically distributed.
- $X_1, X_2, \dots, X_n \sim Model(\theta_1, \theta_2, \dots, \theta_k)$ given all are identically & independently distributed.

5.2 Central Limit Theorem for Statistics

Central Limit Theorem **Definition 5.2.1**

Given some distribution random variable X belonging to some distribution. The mean value of a sample of size n from X is:

$$Y \sim N(\mu, \frac{\sigma^2}{n})$$

Where μ is the expected/mean value of X and σ^2 is its variance.

As the sample size increases, the variance in mean between different samples reduces.

At an infinite sample size, we can use the *standard normal distribution*:

$$\lim_{n \to \infty} \left(\frac{Y - \mu}{\frac{\sigma}{\sqrt{n}}} \right) \sim N(0,1)$$

Ages of a class **Example Question 5.2.1**

Given a class of 20 students, we can calculate the mean and variance:

$$\overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i \text{ and } \overline{\sigma}^2 = \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x})^2$$

There is some unknown distribution of students ages in a class.

If sampling is done with replacement (not students removed from the population after being questioned) we can use the central limit theorem to model the mean and variance of this distribution's mean (the mean age of the class) without needing to know the distribution itself.

$$\overline{x} \text{ is distributed according to } N(\mu, \frac{\sigma^2}{20})$$

Meaning the mean age of any group of 20 students will be distributed normally with parameters:

- μ (The average age of all students/ avergae of all possible groups of 20)
- σ^2 (The variance of means, how different two groups of 20 stuident's means may be expected to be).

As we increase sample size, the variance decreases (larger groups of student $\Rightarrow$ means closer together).

We will use this later in tests, e.g to see if a given mean that occurs is so unlikely it is likely our distribution is wrong, or our sampling biased in some way.

5.3 Estimators

Statistic **Definition 5.3.1**

A *statistic* is a function operating on the random variables of a sample:

$$T = T(X_1, X_2, \dots, X_n) = T(\underline{X})$$

As it is a function of random variables, it is itself a random variable. Hence if distribution X's parameters are known, we can use it:

- if T is the sum of ages of a class of 10, and we know the mean age, variance we can calculate porbabilities for T.
- T may be many useful statistics, e.g the lower quartile of a cohort of 100's GCSE results, or the range of distances flown by birds in a flock.

When given some sample $\underline{x} = (x_1, x_2, \dots, x_n)$ we have:

$$t = t(\underline{x}) = t(x_1, x_2, \dots, x_n)$$

Estimator **Definition 5.3.2**

A statistic used to approximate the parameter of the distribution of its arguments.

- Given a sample $\underline{x}$ the value of the estimator $t = t(\underline{x})$ is called an estimate.
- If we can approximately identify the sampling distribution of the statistic ($P_{T|\theta}$) we can find the expectation, variance (and more) related to our statistic.

If the sample size n is large, *central limit theorem* can be used to approximate the distribution $P_{T|\theta}$

$$T = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

And hence we know approximately that:

$$\overline{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$$

For a given unknown distribution we could use several estimators to approximate its parameter.

Using the first/any X_i as the estimator

$$T[X_1, X_2, \dots, X_n] = X_1 \sim P_{X|\theta}$$

Likewise if we use the median with T:

$$T_{median}[X_1, X_2, \dots, X_n] = X_{\left|\frac{n+1}{2}\right|} \sim P_{X|\theta}$$

However this does not work as we do not know the parameters of the distribution X.

Using the mean as an estimator

$$T_{\overline{X}}[X_1, X_2, \dots, X_n] = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

This is a good estimator for the mean of many distributions, while we do not know μ or σ, we do know the type of distribution.

Estimator Bias **Definition 5.3.3**

We define the bias of an estimator T as estimating the parameter θ is:

$$bias(T) = E[T|\theta] - \theta$$

If bias is 0 we call it an unbiased estimator.

For the mean:

$$E(\overline{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} E[X_i]}{n} = \frac{n \times \mu}{n} = \mu$$

For any distribution the sample mean $\overline{x}$ is an unbiased estimate for the population mean μ.

For the variance: If we know the population mean μ we can also use the unbiased estimator:

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

The sample variance is a biased estimator and is defined as:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

We have too few degrees of freedom, that is based on the mean and $x_{1 \to n-1}$ we can determine x_n, hence we apply *bessel's correction* (wikipedia article on source of bias here) to account for what is effectively a missing variance.

After applying bessel's correction, we get the unbiased estimator of *bias-corrected sample variance*:

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

5.3.1 Bessel's Correction Proof

First we attempt to prove that S_μ^2 is an unbiased estimator for variance.

1. We first define S_μ^2.

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

2. We get the expected value of the estimator, to be an unbiased estimator of variance, this should be equal to the variance.

$$\begin{aligned} E[S_\mu^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i^2 - 2X_i\mu + \mu^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(E[X_i^2] - 2E[X_i]\mu + \mu^2\right) \end{aligned}$$

3. We can substitute μ for $E[X_i]$:

$$\begin{aligned}
E[S_\mu^2] &= \frac{1}{n}\sum_{i=1}^{n}\left(E[X_i^2] - 2E[X_i]E[X_i] + (E[x_i])^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E[X_i^2] - (E[x_i])^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Var[X_i]
\end{aligned}$$

4. As all X_i are identically distributed, $Var[X_i] = Var[X] = \sigma^2$.

$$\begin{aligned}
E[S_\mu^2] &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{n \times \sigma^2}{n} \\
&= \sigma^2
\end{aligned}$$

Hence we can see that S_μ^2 is an unbiased estimator of σ^2.

Next we prove the correction:
1. We get the expected of:

$$E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right]$$

2. We can add and subtract μ (keeping the same value)

$$E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] = E\left[\sum_{i=1}^{n}((X_i - \mu) - (\overline{x} - \mu))^2\right]$$

3. Now we can split the expected up (all distributions are independent (the normal for $\overline{x}$ and we assume independence for X_i)).

$$E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] = E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - 2(\overline{x} - \mu)\left(\sum_{i=1}^{n}(X_i - \mu)\right) + \left(\sum_{i=1}^{n}(\overline{x} - \mu)^2\right)\right]$$

4. We can substitute using $\sum_{i=1}^{n}(X_i - \mu) = n \times (\overline{x} - \mu)$.

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - 2(\overline{x} - \mu) \times n \times (\overline{x} - \mu) + \left(\sum_{i=1}^{n}(\overline{x} - \mu)^2\right)\right] \\
&= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - 2n(\overline{x} - \mu)^2 + \left(\sum_{i=1}^{n}(\overline{x} - \mu)^2\right)\right] \\
&= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - 2n(\overline{x} - \mu)^2 + n(\overline{x} - \mu)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - n(\overline{x} - \mu)^2\right]
\end{aligned}$$

5. We can split the expected (independent distributions) substitute in the variance X.

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - n(\overline{x} - \mu)^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n \times E\left[(\overline{x} - \mu)^2\right] \\
&= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - n \times E\left[(\overline{x} - \mu)^2\right]
\end{aligned}$$

5. As $\overline{x}$ is distributed by a normal distribution $N(\mu, \dfrac{\sigma^2}{n})$, the expected of it shifted by μ and squared is the variance.

$$\begin{aligned} E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - n \times \frac{\sigma^2}{n} \\ &= \sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - \sigma^2 \end{aligned}$$

6. We can then use the variance of the distribution of X:

$$\begin{aligned} E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - \sigma^2 \\ &= n\sigma^2 - \sigma^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

7. Hence to get an unbiased estimator, we need to divide this by $(n-1)$ (apply correction).

$$\begin{aligned} E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= (n-1)\sigma^2 \\ \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sigma^2 \\ E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sigma^2 \end{aligned}$$

Hence $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$ is an unbiased estimator of σ^2.

5.4 Efficient Consistent Estimator

We can quantify how *good* estimators are. For example with the *Estimator Bias* (difference between the expected using the estimator and the parameter $bias(T) = E[T|\theta] - \theta$). We also wanto to quantify the *Efficiency of Estimators*.

Estimator Efficiency **Definition 5.4.1**

Given two unbiased estimators $\hat{\Theta}(\underline{X})$ and $\tilde{\Theta}(\underline{X})$ where $\underline{X} = (X_1, \dots, X_n)$ (a sample containing n observations $X \dots$).

We can compare the mean, variances etc to determine which estimator is more efficient (typically lower variance)

$\hat{\Theta}$ is more efficient than $\tilde{\Theta}$ if:

$$\forall \theta Var_{\hat{\Theta}}(\hat{\Theta}|\theta) \leq Var_{\tilde{\Theta}|\theta}(\tilde{\Theta}|\theta) \quad \text{or} \quad \exists \theta Var_{\hat{\Theta}}(\hat{\Theta}|\theta) < Var_{\tilde{\Theta}|\theta}(\tilde{\Theta}|\theta)$$

More efficient means less variance in estimates.

IF an estimator is more efficient than any other possible estimator, it is called *efficient*.

Bias and Efficiency **Example Question 5.4.1**

Given a population with mean μ and variance σ^2. We have a sample:

$$\underline{X} = (X_1, \dots, X_n)$$

We consider two estimators:

1. $\hat{M} = \overline{X}$ (the sample mean)
2. $\tilde{M} = X_1$ (the first observation in the sample)

We can compute the bias as for both:

1. The expected value of the sample mean is the population mean μ, hence $\hat{M}$ is unbiased.
2. The expected value of any observation is μ, so the first observation in the sample is also ubiased.

Next we can consider the variance.

For a single sample we know the variance will be σ^2, hence:

$$Var_{\tilde{M}}(\tilde{M}|\mu \text{ and } \sigma^2) = Var(X_1) = \sigma^2$$

However for the sample mean, we know can use the *Central Limit Theorem* to determine that the variance of the mean of a sample will be divided by the sample size.

$$Var_{\hat{M}}(\hat{M}|\mu \text{ and } \sigma^2) = Var(\overline{X}) = \frac{\sigma^2}{n}$$

Hence for all values of n, the variance of $\hat{M} \leq \tilde{M}$ (at $n = 1$ they are equal), so $\hat{M}$ is the more efficient estimator.

Estimator Consistency **Definition 5.4.2**

A consistent estimator improves as the sample size grows. Formally:

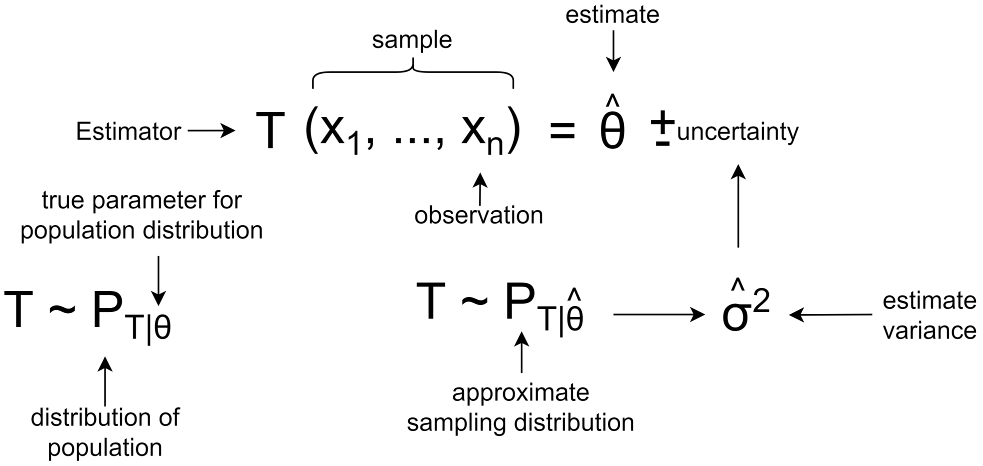
$$\forall \epsilon > 0 \ P(|\hat{\Theta} - \theta|) \to 0 \quad \text{as} \quad n \to \infty$$

If $\hat{\Theta}$ is unbiased, then:

$$\lim_{n \to \infty} Var(\hat{\Theta}) = 0 \Rightarrow \hat{\Theta} \quad \text{is consistent}$$

Note: $\overline{X}$ (sample mean) is a consistent estimator for any population.

5.5 Confidence Intervals



estimate
sample
Estimator
T (x1, ..., xn) = θ̂ ±uncertainty
true parameter for population distribution
observation
T ~ PT|θ
T ~ PT|θ̂
σ̂2
estimate variance
distribution of population
approximate sampling distribution

In order to quantify our degree of uncertainty in an estimate $\hat{\theta}$, when the true value θ is unknown, we use use our estimate as the true value, to compute the distribution $P_{T|\hat{\theta}}$ (the approximate sampling distribution).

5.5.1 Known Variance

Confidence Interval

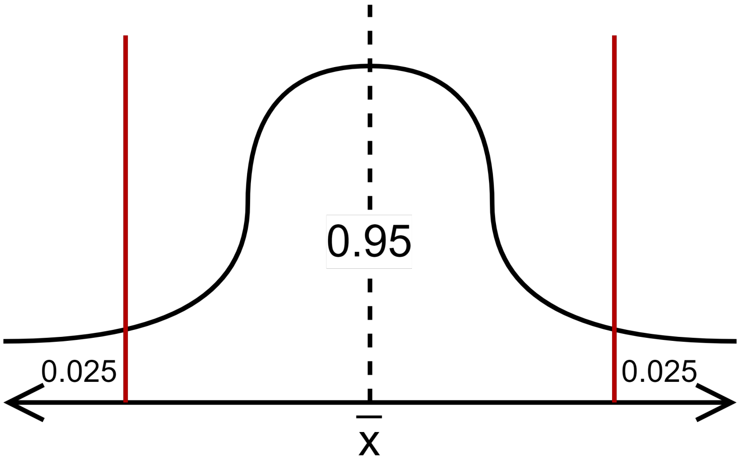
If we know the true variance of the population, then the sample mean would be distributed as:

$$\overline{X} \sim N\left(\overline{x}, \frac{\sigma^2}{n}\right)$$

If μ (population mean) $= \overline{x}$, then we can say that (using the standard normal distribution) there is a 95% probability the observed statistic $\overline{X}$ is in the range:

$$\left[\overline{x} - 1.96\frac{\sigma}{n}, \overline{x} + 1.96\frac{\sigma}{n}\right]$$

(Double ended, 95% confidence interval for μ)



With the Standard Normal Distribution

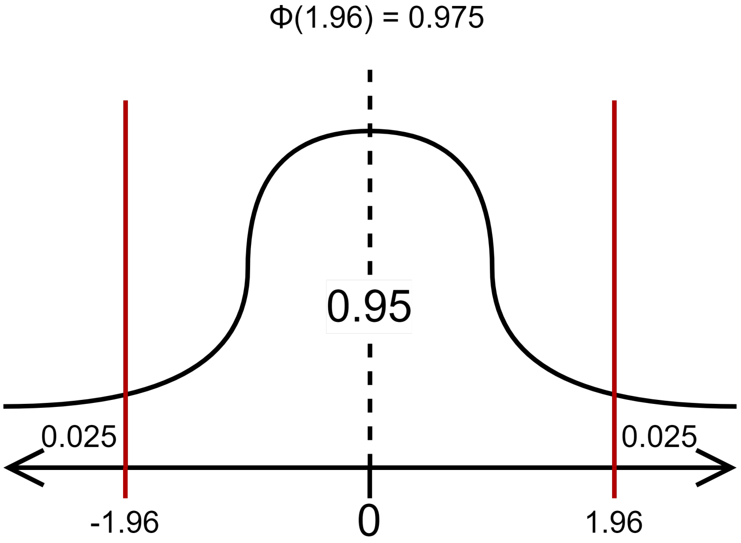
We can define any normal distribution in terms of the standard normal distribution.

$$X \sim N(\mu, \sigma^2) \Leftrightarrow Y = \frac{X - \mu}{\sigma} \Leftrightarrow Y \sim N(0, 1)$$

We can then use tables for the standard normal distribution, using $\Phi(z) = P(X \leq z)$ given $Z \in N(0, 1)$:

Note if you have sample size as part of the variance, $Y = \dfrac{X - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$.

For example in the previous confidence interval, we used the normal distribution to calculate the values.



Given the critical value z for the normal distribution e.g 1.96 for double-ended 95% confidence interval, we have:

Standard Normal	$X \sim N(0,1)$	$[-z, z]$
Normal Distribution	$X \sim N(\mu, \sigma^2)$	$\mu - z\sigma, \mu + z\sigma$
Sample Mean	$\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$	$\left[\mu - z\dfrac{\sigma}{\sqrt{n}}, \mu + z\dfrac{\sigma}{\sqrt{n}}\right]$
Population mean	$\mu \sim N\left(\overline{X}, \dfrac{\sigma^2}{n}\right)$	$\left[\overline{x} - z\dfrac{\sigma}{\sqrt{n}}, \overline{x} + z\dfrac{\sigma}{\sqrt{n}}\right]$

Employees Opinions on the Board **Example Question 5.5.1**

A corporation surveys employees on wether they think the board is doing a good job.

1000 employees are randomly selected, and 732 say the board is doing a good job. Find the 99% confidence interval for the proportion of the employees that think the board is doing a good job. Assume the variance is $\sigma^2 = 0.25$.

First we get the sample mean:

$$\overline{x} = \frac{732}{1000} = 0.732$$

Next we determine the standard deviation:

$$\sigma = \sqrt{0.25} = 0.5$$

We want to get the double-ended 99% interval, so each tail will have size 0.005. By using the standard normal distribution we have $\Phi(2.576) = 0.995$, so $z = 2.576$.

Hence we can calculate the interval as:

$$\begin{aligned}
\mu &= \left[\overline{x} - z\frac{\sigma}{\sqrt{n}}, \overline{x} + z\frac{\sigma}{\sqrt{n}}\right] \\
&= \left[0.732 - 2.576\frac{0.5}{\sqrt{1000}}, 0.732 + 2.576\frac{0.5}{\sqrt{1000}}\right] \\
&= \left[0.732 - 2.576\frac{0.5}{\sqrt{1000}}, 0.732 + 2.576\frac{0.5}{\sqrt{1000}}\right] \\
&\approx 0.732 \pm 0.0407
\end{aligned}$$

5.5.2 Unknown Variance

In a problem where we are trying to fit a normal distribution, but both the mean and variance are unknown.

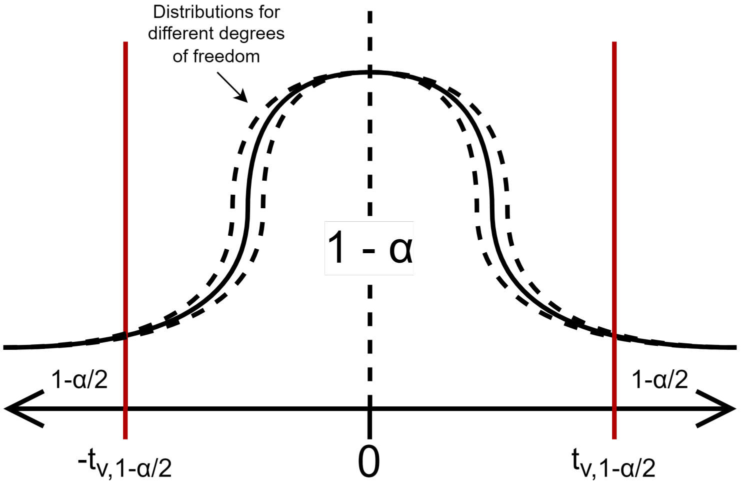
$$\text{Bias Corrected Variance } S_{n-1} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

We use the bias corrected variance of our sample, and as a result must use a different distribution to the normal distribution.

Normal Distribution (σ known)	**Studen't t distribution (σ unknown)**
$\dfrac{\overline{X} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} \sim N(0,1)$	$\dfrac{\overline{X} - \mu}{\left(\dfrac{s_{n-1}}{\sqrt{n}}\right)} \sim t_{n-1}$

In the student's distribution we set degrees of freedom $\nu = n - 1$.

For a double ended confidence $(100 - \alpha)\%$, we compute $t_{\nu=n-1,\ 1-{}^{\alpha}\!/_{2}}$ to find the critical values (the places where the tails start/ the α-quantile of t_ν).



$$\left[\overline{x} - t_{\nu=n-1,\ 1-\alpha/2} \times \frac{s_{n-1}}{\sqrt{n}},\ \overline{x} + t_{\nu=n-1,\ 1-\alpha/2} \times \frac{s_{n-1}}{\sqrt{n}}\right]$$

When using the tables for t values, we use the size we want (e.g 0.975 for 95% double-ended confidence interval), and then use the degrees of freedom $(n-1)$.

Chapter 6

Hypothesis Testing

Hypothesis Test **Definition 6.0.1**

Given two samples, determine if the difference is significant enough to suggest the parameters are different.

- **Null Hypothesis** No statistical relation, there is no evidence for a claim. (H_0)
- **Alternative Hypothesis** There is a statistical relation. (H_1)

We can partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 for the null and alternative hypotheses, which can be expressed as:

$$H_0 \ : \ \theta \in \Theta_0 \ \text{ and } \ H_1 \ : \ \theta \in \Theta_1$$

(We are testing if based on a given sample, based onm the estimated parameter, if it is plausible the sample distribution is from another distribution)

- **Simple Hypothesis** Test that $\theta = \theta_0$
- **Composite Hypothesis** Test that $\theta > \theta_0$ or $\theta < \theta_0$

Typically a test is of the form:

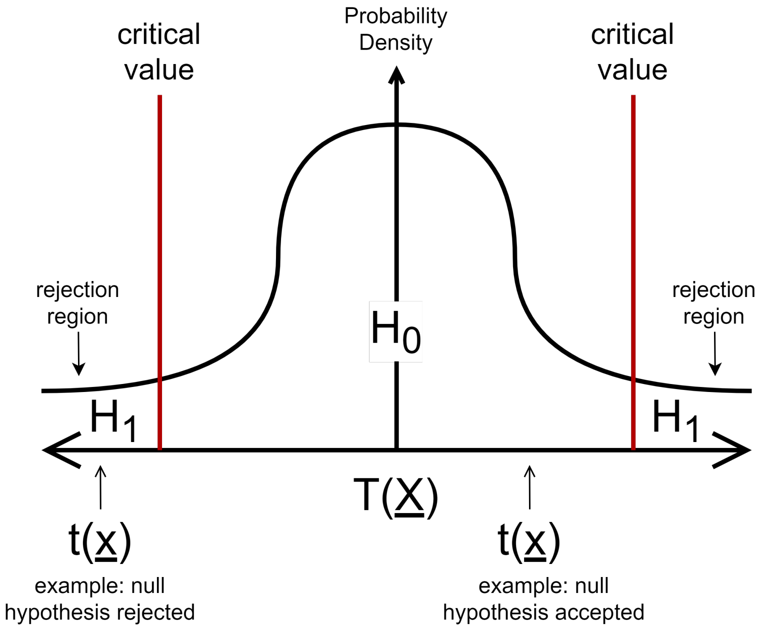
$$H_0 \ : \ \theta = \theta_0 \ \text{ versus } H_1 \ : \ \theta \neq \theta_0$$

Some tests are one-sided, for example:

$$H_0 \ : \ \theta > \theta_0 \ \text{ versus } H_1 \ : \ \theta < \theta_0$$

To test the validity of H_0:

1. Choose a *test statistic* $T(\underline{X})$ to use on the data.
2. Find a distribution P_T under H_0 from the *test statistic*.
3. Determine the rejection region (the region in which a result would invalidate H_0).
4. Calculate the observed *test statistics* $t(\underline{x})$.
5. If $t(\underline{x})$ is in the rejection region, reject H_0 and accept H_1, else retain H_0.



The *significance level/Type 1 Error Rate* $\alpha \in (0, 1)$ of as hypothesis test determines the size of the rejection regions.

- $\alpha \to 0$ Less and less likely to reject H_0, rejection region samller, confidence in our result is lower - easier test.
- $\alpha \to 1$ More and more likely to reject H_0, rejection region larger, confidence higher - stricter test.

The *p-value* of a test is the significance level threshold between rejection/acceptance of H_0 for a given test.

Test Errors **Definition 6.0.2**

- **Type 1** Reject H_0 when it is actually true. $\alpha = P(T \in R|H_0)$
(significance is the probability of incorrectly rejecting the null hypothesis)
- **Type 2** Accepting H_0 when H_1 is true. $\beta = P(T \notin R|H_1)$
Probability a test statistic is not in the rejecting region, when H_1 is true.

Test Power **Definition 6.0.3**

The probability of correctly rejecting the null hypothesis

$$Power = 1 - \beta = 1 - P(T \notin R|H_1) = P(T \in R|H_1)$$

For a given significance level:

$$\alpha = P(T \in R|H_0)$$

A good *test statistic* T and *rejection region* R will have a high *power*, the highest *power* test under H_1 is called the *most powerful.*

Given a control group (placebo) and a test group (given some pharmaceutical), we can test the hypotheis that the drug has an effect on survival rates.

$$H_0 : \text{The drug has no effect - survival rates are the same.}$$
$$H_1 : \text{The drug has an effect - survival rates are different.}$$

6.1 Testing For Population Mean

Sample mean belongs to a normal distribution (*Central Limit Theorem*):

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

We have our two hypotheses:

$$H_0 \ : \ \mu = \mu_0 \quad \text{versus} \quad H_1 \ : \ \mu \neq \mu_0$$

We can derive a new distribution in terms of the standard normal:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Hence for significance α (or confidence interval $1-\alpha$) we can get the rejection/acceptance regions.

$$\Phi(1-\alpha) = threshold \quad \text{results in acceptance region: } [-threshold, threshold]$$

Hence we can calculate z for a given sample, and then determine if it is in the region, if it is then accept H_0, else rejected H_0 and accept H_1.

Weight of Crisp Packets (Known Variance) **Example Question 6.1.1**

A crisp manufacturer sells packets listed as having weight $454g$. From a sample size of 50, we get the mean weight of a bag as $451.22g$.

Assume the variance of bag weights is 70. Is the observed sample consistent with the claim made by the company at the 5% significance.

$$H_0: \ \mu = 454g$$
$$H_1: \ \mu \neq 454g$$

We have the following information:

$$\overline{x} = 451.22g \quad \sigma^2 = 70 \quad n = 50 \quad \alpha = 0.05$$

Hence we can state the hypothesized distribution of the sample mean:

$$\overline{X} \sim N\left(454g, \frac{70}{50}\right)$$

We can get this in terms of the standard normal distribution:

$$Z = \frac{\overline{X} - 454}{\sqrt{35}/5} \sim N(0,1)$$

At the 5% significance, we have 2.5% are each tail. Hence we get our critical value as $z(critical) = 0.975$, where 1.96.

Hence the rejection region is:

$$\frac{\overline{X} - 454}{\sqrt{35}/5} < -1.96$$
$$\frac{\overline{X} - 454}{\sqrt{35}/5} > 1.96$$

Hence in order to accept H_0, $\overline{X}$ must be in the interval:

$$451.6809 < \overline{X} < 456.3191$$

As $\overline{x} = 451.22$ it is in the rejection region, hence at the 95% significance there is sufficient evidence to reject the company's claim.

Weight of Crisp Packets (UnKnown Variance) **Example Question 6.1.2**

A Crisp manufacturer sells packets listed as having weight $454g$. From a sample size of 50, we get the mean weight of a bag as $451.22g$.

Assume the variance of bag weights is 70. Is the observed sample consistent with the claim made by the company at the 5% significance?

$$H_0: \ \mu = 454g$$
$$H_1: \ \mu \neq 454g$$

We have the following information:

$$\overline{x} = 451.22g \quad n = 50 \quad \alpha = 0.05$$

We first calculate the bias corrected sample variance:

$$\begin{aligned} s_{n-1} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \sqrt{70.502} \text{ (Need to calculate from each observation in the sample)} \end{aligned}$$

Hence we can now use the *student's t distribution* with degrees of freedom $n - 1 = 49$.

$$\frac{\overline{x} - \mu_0}{s_{n-1}/\sqrt{n}} \sim t_{49}$$

For $\alpha = 5\%$ we take the tails as 0.025, so use $t_{49,\ 0.975} \approx 2.01$. Hence will reject the regions:

$$\frac{\overline{X} - 454}{\sqrt{70.502}/5\sqrt{2}} < -2.01$$
$$\frac{\overline{X} - 454}{\sqrt{70.502}/5\sqrt{2}} > 2.01$$

Hence to accept H_0, $\overline{X}$ must be:

$$451.6123 < \overline{x} < 456.3868$$

Hence at the 5% significance there is sufficient evidence to reject H_0 and accept H_1.

Optimising Code — Example Question 6.1.3

The previous code had a mean run time of $6s$. Following an optimisation a sample of runs is taken, with sample of size 16, mean $5.8s$ and bias corrected sample standard deviation of $1.2s$. Is the new code faster?

Our test is as follows:

$$H_0 \ : \ \mu \geq 6s \text{ (mean time is same or larger)} \quad \text{versus} \quad H_1 \ : \ \mu < 6s \text{ (mean time is lower)}$$

We have the following information:

$$\overline{x} = 5.8 \quad s_{n-1} = 1.2s \quad n = 16$$

Hence we have the distribution:

$$\frac{\overline{X} - \mu}{s_{n-1}/\sqrt{n}} \sim t_{15}$$

Hence we can use the significance (one ended/top tail) of 5% to find $t_{15,0.95} \approx 1.75$.

Hence will reject the regions:

$$\frac{\overline{X} - 6}{1.2/4} < -1.75$$
$$\frac{\overline{X} - 6}{1.2/4} > 1.75$$

Hence to accept H_0, $\overline{X}$ must be:

$$5.475 < \overline{X} < 6.525$$

Hence as $\overline{x} = 5.8$ this is within the acceptable region, so at the 95% significance we have insufficient evidence to reject H_0.

6.2 Samples from Two Populations

When given two random samples:

$$\underline{X} = (X_1, \dots, X_n) \text{ from } P_X \quad \text{and} \quad \underline{Y} = (Y_1, \dots, Y_n) \text{ from } P_Y$$

We may want to determine the similarity of the distributions of P_X and P_Y.

Typically this involves testing to see if the means of the populations are equal:

$$H_0 \ : \ \mu_X = \mu_Y \quad \text{versus} \quad H_1 \ : \ \mu_X \neq \mu_Y$$

Paired Data **Definition 6.2.1**

A special case when $\underline{X}$ and $\underline{Y}$ are pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ (each X_i and Y_i are possibly dependent on each-other).

For example, where for a person i, X_i is the heart rate before exercise, and Y_i the rate afterwards.

We can consider a sample of the differences, if this has mean 0:

$$Z_i = X_i - Y_i \text{ testing } H_0 \ : \ \mu_Z = 0 \quad \text{versus} \quad H_1 \ : \ \mu_Z \neq 0$$

6.2.1 Known Variance, X and Y are Independent

Given that:

$$\begin{matrix} \underline{X} = (X_1, \dots, X_{n_1}) & X_i \sim N(\mu_X, \sigma_X^2) & \overline{X} \sim N\left(\mu_X, \dfrac{\sigma_X^2}{n_1}\right) \\ \underline{Y} = (Y_1, \dots, Y_{n_2}) & Y_i \sim N(\mu_Y, \sigma_Y^2) & \overline{Y} \sim N\left(\mu_Y, \dfrac{\sigma_Y^2}{n_2}\right) \end{matrix}$$

We can therefore get the distribution of the difference in sample means:

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right)$$

And hence:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_1} + \dfrac{\sigma_Y^2}{n_2}}} \sim N(0,1)$$

As we assume for H_0 that $\mu_x = \mu_Y$ we have:

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n_1} + \dfrac{\sigma_Y^2}{n_2}}} \sim N(0,1)$$

So we can calculate the *test statistic*:

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{\sigma_X^2}{n_1} + \dfrac{\sigma_Y^2}{n_2}}}$$

And use this to determine if H_0 is rejected.

6.2.2 Unknown Variance, X and Y are Independent, Variances Equal

Bias-Corrected Pooled Sample Variance **Definition 6.2.2**

If the variance of two samples is the same, given:

$$\underline{X} = (X_1, \dots, X_{n_1}) \text{ and } \underline{Y} = (Y_1, \dots, Y_{n_2})$$

We can get an unbiased estimator of the variance as:

$$S^2_{N_1+n_2-2}\frac{(n_1-1)S^2_{n_1-1,\ X} + (n_2-1)S^2_{n_2-1,\ Y}}{(n_1-1)+(n_2-1)}$$

Which is equivalent to:

$$S^2_{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_1+n_2-2}$$

If σ_X^2 and σ_Y^2 are unknown, but it is know that $\sigma^2 = \sigma_X^2 = \sigma_Y^2$ we can use an estimator to get an estimate of the variance σ^2 using the samples from the two populations.

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_Y)}{\sigma\sqrt{{}^1/_{n_1} + {}^1/_{n_2}}} \sim N(0,1)$$

Hence if the $H_0 \ : \ \mu_X = \mu_Y$ then:

$$\frac{\overline{X} - \overline{Y}}{\sigma\sqrt{{}^1/_{n_1} + {}^1/_{n_2}}} \sim N(0,1)$$

To get an estimate for the variance we can use the *Bias-Corrected Pooled Sample Variance*

Compiler Comparison **Example Question 6.2.1**

Given two compilers, attempt to determine if compiler 2 produces is faster code (to 5% significance).

Compiler 1	**Compiler 2**
$n_1 = 15$	$n_2 = 15$
$\overline{x} = 114s$	$\overline{y} = 94s$
$s^2_{14} = 310$	$s^2_{14} = 290$
μ_1	μ_2

$$H_0 \ : \ \mu_1 \leq \mu_2 \ \text{ versus } \ H_1 \ : \ \mu_1 > \mu_2$$

We assume that the variances of the population variances are the same for both compilers.

We can get the *Bias-Corrected Pooled Sample Variance*:

$$S_{28} = \frac{14 \times 310 + 14 \times 290}{14 + 14} = 300$$

Hence our *test statistic* is:

$$\frac{\overline{x} - \overline{y}}{\sigma\sqrt{{}^1/_{n_1} + {}^1/_{n_2}}} = \frac{20}{\sqrt{300}\sqrt{\frac{2}{15}}} = \sqrt{10} \approx 3.162$$

We can now use the *student's t distribution* to get the rejection region (one-sided):

$$t_{28,0.95} = 1.701$$

Hence as $3.162 > 1.701$ we have sufficient evidence at the 5% significance to reject H_0 and accept H_1. The second compiler produces faster code.

Welch's t-test *Extra Fun!* **6.2.1**

If the variances are unknwon, and not equal, we can use *Welch's t test*.

The *test statistic* is:

$$\frac{(\overline{x} - \overline{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_{n_1,X}}{n_1} + \frac{S^2_{n_1,Y}}{n_1}}}$$

We then use a t distribution t_ν with the ν degrees of freedom determined by rounding the following to the nearest whole number:

$$\nu = \frac{\left(\left(\frac{S^2_{n_1,\ X}}{n_1}\right) + \left(\frac{S^2_{n_1,\ X}}{n_1}\right)\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{S^2_{n_1,\ X}}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{S^2_{n_2,\ Y}}{n_2}\right)^2}$$

The we proceed as normal, checking the test statistic is within the rejection regions.

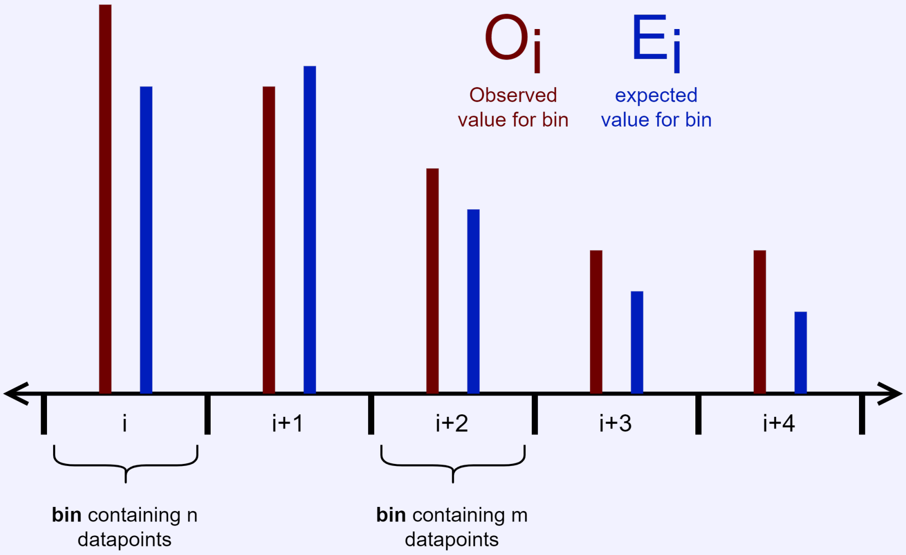
6.3 Chi Squared Testing

6.3.1 Goodness of Fit

Binning **Definition 6.3.1**

Given a distribution, we can partition it into several disjoint *bins*. Essentially we are creating a pesudo-*PMF* (potentially with ranges instead of just discrete values) describing how many datapoints/the frequency we would expect to find from a distribution.

As a result, we can directly compare the expected values E_i (from a distribution we are checking a sample against), with the observations O_i from a sample.



Goodness of Fit/Chi-Square Statistic **Definition 6.3.2**

Denotes the difference between some expected values, and some observed.

For n bins we have:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

6.3.2 Chi-Squared Test for Model Checking

Used to determine if an observed sample matches a given distribution to some significance.

1. Determine expected distribution (can use parameters estimated from the sample).
2. Create a hypotheses based some parameters θ:

$$H_0 \ : \ \theta = \theta_0 \ \text{ versus } \ H_1 \ : \ \theta \neq \theta_0$$

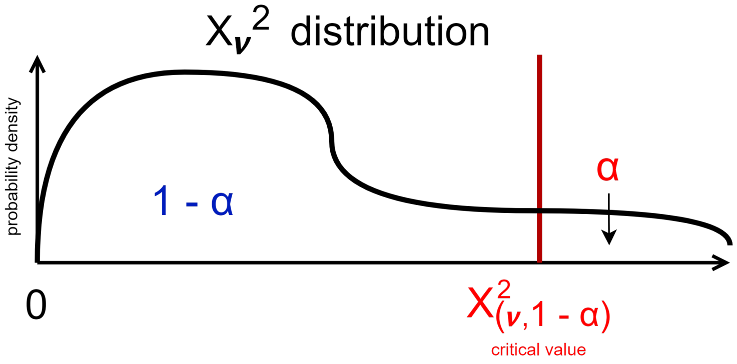
3. Bin the expected distribution (for comparison with the observed).
4. Calculate the *Goodness of Fit/Chi-Square Test Statistic* X^2.
5. Calculate the degrees of freedom as:

$$\nu = (\text{number of possible values } X \text{ can take}) - (\text{number of parameters being estimated}) - 1$$

6. Determine the critical value using the *Chi Squared Distribution* χ^2_ν and the significance α (typically using a table).
7. If $X^2 > \chi^2_{\nu,\ 1-\alpha}$ (test statistic larger than critical value)

Note that:

- All expected values must be larger than 5 for a good test. Hence some bins may have to be merged.
- The number of values X can take is typically the number of bins.



Adverse Drug Effects **Example Question 6.3.1**

A study in the journal of the American Medical Association gives the causes of a sample of 95 adverse drug effects as:

Reason	No. Adverse Effects
Lack of Knowledge	29
Rule Violation	17
Faulty Dose Check	13
Slips	9
Other Cause	27

Test if the true percentages of causes of adverse effects are different at the 5% significance.

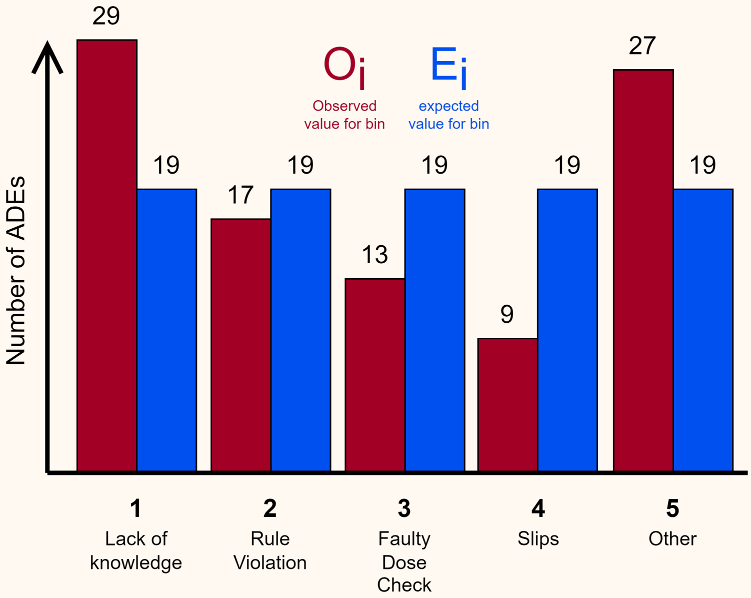
As we are checking the percentages are the same, we effectively have a discrete uniform distribution:

$$X \sim U(1, 5)$$

Hence we can calculate our *null and alternative hypotheses*:

$$H_0 \ : \ X \sim U(1, 5) \ \text{ versus } \ H_1 \ : \ X \nsim U(1, 5)$$

Now we can bin the distribution, (no merging is required as all expected values are larger than 5):



It is now possible to compute goodness of fit.

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(29-19)^2}{19} + \frac{(17-19)^2}{19} + \frac{(13-19)^2}{19} + \frac{(9-19)^2}{19} + \frac{(27-19)^2}{19} \\
&= 16
\end{aligned}
$$

We have $\nu = 4$ as there are 5 possible values, and no parameters were estimated from the data.

Hence we get the critical value from the chi-squared table: $\chi^2_{4,\ 0.95} = 9.49$

As $16 > 9.49$ there is sufficient evidence at the 5% significance level to reject H_0, the percentages differ.

Football Games — Example Question 6.3.2

Given the total number of goals for 2608 football matches, determine if the number of goals scored in a match can be modelled by $X \sim Poisson(3.870)$ at the 5% significance.

Goals Scored (x)	0	1	2	3	4	5	6	7	8	9	≥ 10	
Matches (n_x)	57	203	383	525	532	408	273	139	139	45	27	16

Hence as we already have a distribution, we can create our hypotheses:

$$H_0 \ : \ X \sim Poisson(3.870) \quad \text{versus} \quad H_1 \ : \ X \not\sim Poisson(3.87)$$

We can then use the poisson distribution to calculate the expected for 2608 football matches, for the final (≥ 10) we use the cumulative to get the remaining probability.

Goals	0	1	2	3	4	5	6	7	8	9	≥ 10
O	57	203	383	525	532	408	273	139	45	27	16
E	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1
$\frac{(O-E)^2}{E}$	0.124	0.267	1.461	0.000	1.096	0.534	1.452	0.012	7.723	0.166	0.071

Hence we get our test statistic as: $X^2 = 12.906$.

As we did not estimate any parameters from the sample, the degrees of freedom are $\nu = 11 - 1 = 10$.

The critical value is: $\chi^2_{10,\ 0.95} = 16.91$.

Hence as $12.906 < 16.91$ we there is insufficient evidence as the 5% significance to reject H_0, the goals can be modelled as $Poisson(3.87)$.

6.3.3 Chi-Squared Test for Independence

Contingency Table **Definition 6.3.3**

A table denoting the frequency of each combination of values for X and Y.

		Possible values of y				Marginal
		y_1	y_2	$\dots$	y_l	
Possible x	x_1	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,l}$	$n_{1,\bullet}$
	x_2	$n_{2,1}$	$n_{2,2}$	$\dots$	$n_{2,l}$	$n_{2,\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_k	$n_{k,1}$	$n_{k,2}$	$\dots$	$n_{k,l}$	$n_{k,\bullet}$
Marginal		$n_{\bullet,1}$	$n_{\bullet,2}$	$\dots$	$n_{\bullet,l}$	n

We can use the marginal values to determine the expected value, if the two distributions were independent.

Given a dataset of points $(x,y)_1, (x,y)_2, \dots, (x,y)_n$, we can consider it the joint distribution P_{XY} of the distributions P_X and P_Y.

To test if the distributions P_X and P_Y are independent from the sample (without knowing the actual distributions themselves) we can use a *contingency table.*

For the contingency table entry coordinates $0 < i \leq l$, $0 < j \leq k$:

$$O_{i,j} = n_{i,j} \text{ and } E_{i,j} = \frac{n_{i,\bullet} \times n_{\bullet,j}}{n}$$

Hence we can now compute the X^2 (*Chi Squared test statistic*) using these observed and expected values.

We compute the degrees of freedom as $\nu = (rows - 1) \times (columns - 1)$ (each row and column alone has degrees of freedom $n-1$ as they must sum to the row/column total), and can then do the *Chi-Squared Test* normally.

Fitness and Stress **Example Question 6.3.3**

	Poor Fitness	Average Fitness	Good Fitness	
Stress	206	184	85	475
No Stress	36	28	10	74
	242	212	95	549

Determine at the 5% significance if there is a link between fitness and stress.

For this test the null hypothesis will be that fitness and stress are independent.

$$H_0 \ : \text{ Stress and fitness are independent } \text{ versus } H_1 \ : \text{ Stress and Fitness re not independent}$$

Next we can calculate the expected values:

	Poor Fitness		Average Fitness		Good Fitness		
	O	E	O	E	O	E	
Stress	206	209.4	184	183.4	85	82.2	475
No Stress	36	32.6	28	28.6	10	12.8	74
	242		212		95		549

We can then calculate our test statistic to be $X^2 = 1.133$.

To compute the degrees of freedom $\nu = (2-1) \times (3-1) = 2$.

Hence we can get our critical value $\chi^2_{2,\ 0.95} = 5.99$.

As $5.99 > 1.133$, there is insufficient evidence to reject H_0 at the 5% significance level. Stress and fitness are not linked.

Chapter 7

Maximum Likelihood Estimate

Given some distribution with an unknown parameter θ:

$$X \sim Distribution(\dots\theta\dots)$$

And a sample taken from the distribution $\underline{X}$:

$$\underline{X} = (X_1, X_2, \dots, X_n)$$

We want to know the value of θ for which the likelihood of the sample occurring is highest.

Likelihood Function **Definition 7.0.1**

The likelihood of some observations $x_1, x_2, \dots, X_n$ occurring given some θ are:

$$\begin{aligned} L(\theta) &= P(x_1, x_2, \dots, x_n|\theta) \\ &= \prod_{i=1}^{n} f(x_i|\theta) \end{aligned}$$

This is as f is the *probability mass function*, and as each observation is independent we can multiply their probabilities.

Log Likelihood Function **Definition 7.0.2**

Used more often than likelihood (easier to work with, and converts decimal small values to large negative values - avoids floating point errors)

$$l(\theta) = \ln L(\theta)$$

To do this, we construct the likelihood (or log likelihood) function from the distribution and sample in term of θ.

Then we can differentiate the function to determine the value of θ for the maximum.

This value of θ is the *Maximum Likelihood Estimate* ($\hat{\theta}$).

7.1 Common Maximum Likelihood Estimates

Given a sample $\underline{x} = (x_1, x_2, \dots, x_n)$, we can use formulas for the maximum likelihood.

7.1.1 Exponential Distribution

$$X \sim Exp(\theta) \Rightarrow f(x) = \theta e^{-\theta x}$$

First we determine the *likelihood* in terms of θ.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i) \\ &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &= \theta^n \prod_{i=1}^{n} e^{-\theta x_i} \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \end{aligned}$$

Next we can derive the *log likelihood*

$$\begin{aligned} l(\theta) &= \ln L(\theta) \\ &= \ln \left(\theta^n e^{-\theta \sum_{i=1}^{n} x_i}\right) \\ &= n \ln \theta - \theta \sum_{i=1}^{n} x_i \end{aligned}$$

Next we can differentiate and set equal to zero:

$$\begin{aligned} \frac{dl(\theta)}{d\theta} &= n\frac{1}{\theta} - \sum_{i=1}^{n} x_i = 0 \\ 0 &= \frac{n}{\theta} - \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i &= \frac{n}{\theta} \\ \theta &= \frac{n}{\sum_{i=1}^{n} x_i} \end{aligned}$$

Hence the maximum likelihood estimator is the reciprocal of the mean of the sample.

$$\hat{\theta} = {}^{1}\!/_{\overline{x}}$$

7.1.2 Geometric Distribution

$$X \sim \mathit{Geo}(\theta) \Rightarrow f(x) = \theta(1-\theta)^{x-1}$$

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i) \\ & \prod_{i=1}^{n} \theta(1-\theta)^{x_i - 1} \\ & \theta^n \prod_{i=1}^{n} (1-\theta)^{x_i - 1} \\ & \theta^n (1-\theta)^{\sum_{i=1}^{n}(x_i - 1)} \\ & \theta^n (1-\theta)^{\left(\sum_{i=1}^{n} x_i\right) - n} \end{aligned}$$

Now we find the *log likelihood*.

$$\begin{aligned} l(\theta) &= \ln L(\theta) \\ &= \ln \left(\theta^n (1-\theta)^{\left(\sum_{i=1}^{n} x_i\right) - n}\right) \\ &= n \ln \theta + \left(\left(\sum_{i=1}^{n} x_i\right) - n\right) \ln (1-\theta) \end{aligned}$$

Now we differentiate, and set equal to zero to find $\hat{\theta}$.

$$\begin{aligned}
\frac{dl(\theta)}{d\theta} &= \frac{n}{\theta} + \left(\left(\sum_{i=1}^{n} x_i\right) - n\right)\frac{1}{\theta - 1} = 0 \\
0 &= \frac{n(\theta-1)}{\theta(\theta-1)} + \left(\left(\sum_{i=1}^{n} x_i\right) - n\right)\frac{\theta}{\theta(\theta-1)} \\
0 &= n(\theta - 1) + \left(\left(\sum_{i=1}^{n} x_i\right) - n\right)\theta \\
0 &= n\theta - n + \left(\left(\sum_{i=1}^{n} x_i\right) - n\right)\theta \\
n &= \left(\sum_{i=1}^{n} x_i\right)\theta \\
\frac{n}{\sum_{i=1}^{n} x_i} &= \theta
\end{aligned}$$

Hence the maximum likelihood estimator is the reciprocal of the mean of the sample.

$$\hat{\theta} = {}^{1}\!/_{\bar{x}}$$

7.1.3 Binomial Distribution

$$X \sim Binomial(m, \theta) \Rightarrow f(x) = \binom{m}{x}\theta^{x}(1-\theta)^{m-x}$$

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(x_i) \\
&= \prod_{i=1}^{n}\binom{m}{x_i}\theta^{x_i}(1-\theta)^{m-x_i} \\
&= \prod_{i=1}^{n}\binom{m}{x_i} \times \prod_{i=1}^{n}\theta^{x_i} \times \prod_{i=1}^{n}(1-\theta)^{m-x_i} \\
&= \prod_{i=1}^{n}\binom{m}{x_i} \times \theta^{\sum_{i=1}^{n} x_i} \times (1-\theta)^{\sum_{i=1}^{n} m - x_i} \\
&= \prod_{i=1}^{n}\binom{m}{x_i} \times \theta^{\sum_{i=1}^{n} x_i} \times (1-\theta)^{mn - \sum_{i=1}^{n} x_i}
\end{aligned}$$

Now we find the *log likelihood*.

$$\begin{aligned}
l(\theta) &= \ln L(\theta) \\
&= \ln\left(\prod_{i=1}^{n}\binom{m}{x_i} \times \theta^{\sum_{i=1}^{n} x_i} \times (1-\theta)^{mn-\sum_{i=1}^{n} x_i}\right) \\
&= \ln\prod_{i=1}^{n}\binom{m}{x_i} + \ln\theta^{\sum_{i=1}^{n} x_i} + \ln(1-\theta)^{mn-\sum_{i=1}^{n} x_i} \\
&= \ln\prod_{i=1}^{n}\binom{m}{x_i} + \sum_{i=1}^{n} x_i \ln\theta + \left(mn - \sum_{i=1}^{n} x_i\right)\ln(1-\theta)
\end{aligned}$$

Now we differentiate, and set equal to zero to find $\hat{\theta}$.

$$\begin{aligned}
\frac{dl(\theta)}{d\theta} &= 0 + \sum_{i=1}^{n} x_i \frac{1}{\theta} + \left(mn - \sum_{i=1}^{n} x_i \right) \frac{1}{\theta - 1} = 0 \\
0 &= \sum_{i=1}^{n} x_i \frac{\theta - 1}{\theta(\theta - 1)} + \left(mn - \sum_{i=1}^{n} x_i \right) \frac{\theta}{\theta(\theta - 1)} \\
0 &= \sum_{i=1}^{n} x_i(\theta - 1) + \left(mn - \sum_{i=1}^{n} x_i \right) \theta \\
0 &= \theta \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i + mn\theta - \theta \sum_{i=1}^{n} x_i \\
0 &= - \sum_{i=1}^{n} x_i + mn\theta \\
\frac{\sum_{i=1}^{n} x_i}{mn} &= \theta
\end{aligned}$$

Hence the maximum likelihood estimator is the sample mean divided by the number of trials (for binomial):

$$\hat{\theta} = \frac{\overline{x}}{m}$$

Chapter 8

Posterior

8.1 MLE Sensitivity

There are several shortcomings of *MLE*:

- **Sensitive to Sample Size**
  In a small sample, small fluctuations can change the *MLE* considerably.
- **Does not use any Prior Information**
  Only uses the given sample.
- **Returns a single value**
  Only returns the single and specific value $\hat{\theta}$, not a distribution $P(\theta|\underline{x})$ for some sample $\underline{x}$.

  Hence we cannot know how close other θ are, how strong our estimate is.
- **Cannot Assess**
  Can only assess using confidence intervals, however these are also dependent on the sample.

8.2 Bayes & Posterior

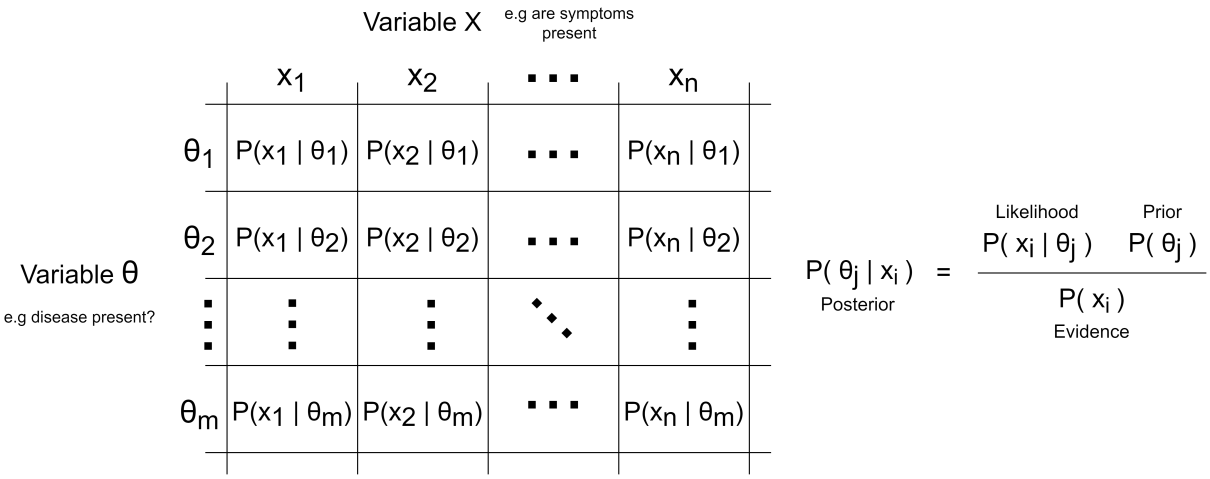
Baye's Theorem **Definition 8.2.1**

Given two events A and B, where $P(B) \neq 0$:

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

Note that we can use the law of total probability to re-express this without knowing $P(B)$:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\overline{A})(1 - P(A))}$$



	x_1	x_2	$\cdots$	x_n
θ_1	$P(x_1 \mid \theta_1)$	$P(x_2 \mid \theta_1)$	$\cdots$	$P(x_n \mid \theta_1)$
θ_2	$P(x_1 \mid \theta_2)$	$P(x_2 \mid \theta_2)$	$\cdots$	$P(x_n \mid \theta_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
θ_m	$P(x_1 \mid \theta_m)$	$P(x_2 \mid \theta_m)$	$\cdots$	$P(x_n \mid \theta_m)$

$$\underset{\text{Posterior}}{P(\theta_j \mid x_i)} = \frac{\overset{\text{Likelihood}}{P(x_i \mid \theta_j)} \quad \overset{\text{Prior}}{P(\theta_j)}}{\underset{\text{Evidence}}{P(x_i)}}$$

By law of total probability:

$$\text{Given } j \in [1,m].\ \sum_{i=1}^{n} P(x_i|\theta_j) = 1 \ \text{ and given } i \in [1,n] \sum j = 1^m P(\theta_j|x_i) = 1$$

When calculating the *MLE* using a sample $\underline{x}$ we calculated:

$$\hat{\theta}_{MLE} = arg\max_{\theta} L(\theta|\underline{x}) = arg\max_{\theta} \left[\prod_{i=1}^{n} P(x_i|\theta)\right]$$

(The θ most likely to give the sample $\underline{x}$)

We can apply this to the distributions X and θ to get a joint distribution:

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Where the *evidence* (X), acts as a normalizer (does not alter the shape of the distribution, just stretches/compresses it to normalize so that the distribution of $\theta|X$ has total probability 1)

$$\int_{-\infty}^{\infty} P(\theta|X)d\theta = 1$$

Hence we can say that the likelihood, and the posterior are directly proportional:

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

8.3 Maximum a Posteriori (MAP) Estimate

Maximum a Posteriori Estimate (MAP Estimate) **Definition 8.3.1**

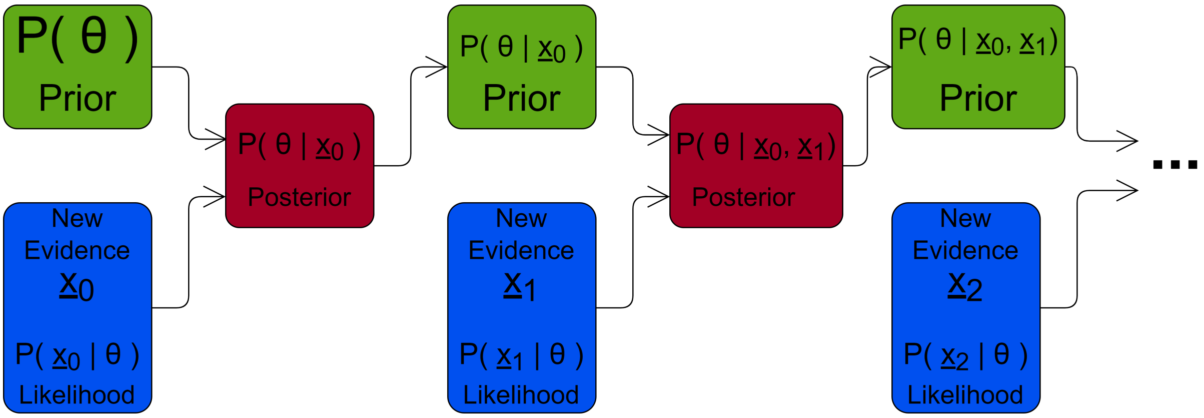
Given some prior information $(P(\theta))$ we can effectively get the *MLE*, but each probability is weighted by the prior information.

$$\begin{aligned}\hat{\theta}_{MAP} &= arg\max_{\theta} \left[\prod_{i=1}^{n} P(\theta|X = x_i)\right] \\ &= arg\max_{\theta} \left[\prod_{i=1}^{n} \frac{P(X = x_i|\theta) \times P(\theta)}{P(X = x_i)}\right] \\ &= arg\max_{\theta} \left[\prod_{i=1}^{n} P(X = x_i|\theta) \times P(\theta)\right]\end{aligned}$$

Using the uniform distribution as $P(\theta)$ yields the *MLE* as each $P(X = x_i|\theta)$ is equally weighted.

8.4 Conjugate Priors

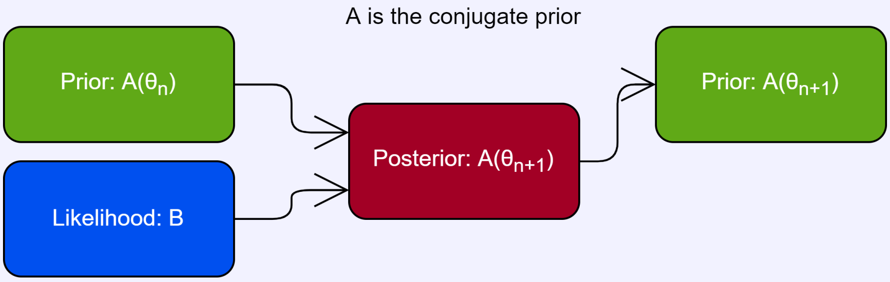
We can continually use the *MAP* to get new prior information, to use with new evidence to refine the *MAP*. This process of continually using the previous estimate and new evidence to refine the estimate is called *Baysian Inference*


P(θ)
Prior
New Evidence $\underline{x}_0$
P($\underline{x}_0$ | θ)
Likelihood
P(θ | $\underline{x}_0$)
Posterior
P(θ | $\underline{x}_0$)
Prior
New Evidence $\underline{x}_1$
P($\underline{x}_1$ | θ)
Likelihood
P(θ | $\underline{x}_0$, $\underline{x}_1$)
Posterior
P(θ | $\underline{x}_0$, $\underline{x}_1$)
Prior
New Evidence $\underline{x}_2$
P($\underline{x}_2$ | θ)
Likelihood
...

$$\text{where } P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{\int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta}$$

Conjugate Prior **Definition 8.4.1**

When continually inferring new prior distributions, if the prior distribution is in the same family of distributions (i.e parameters can be different, but same distribution) as the posterior, then it is a *conjugate prior*.



Likelihood	**Conjugate Prior**
Bernoulli Binomial Geometric	Beta
Poisson Exponential	Gamma
Normal	Normal

Beta Prior Distribution **Definition 8.4.2**

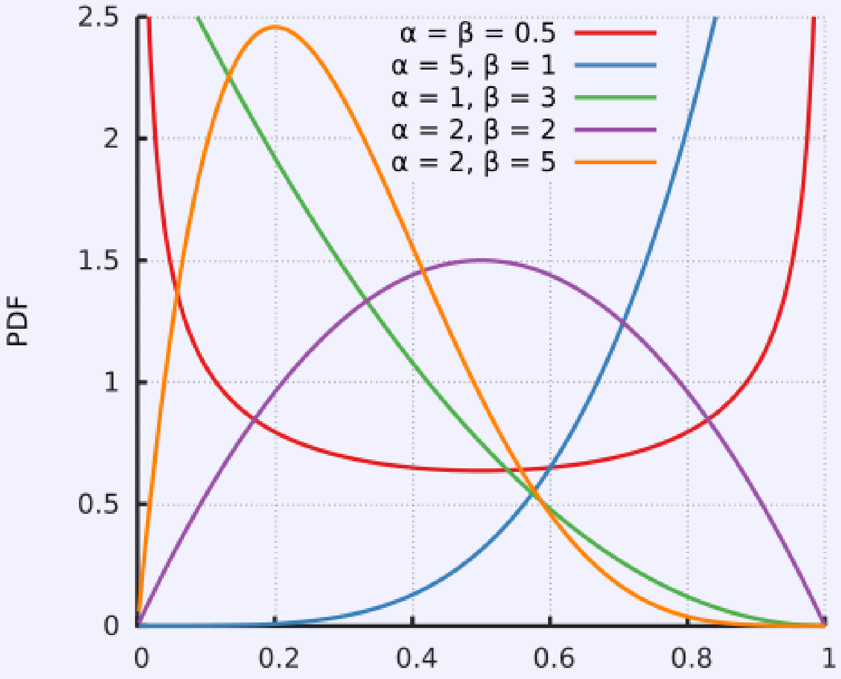
Where $\alpha, \beta > 0$ are *hyper-parameters* that determine the shape of the distribution, the parameter is θ:

$$Beta(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Where the normalising value (ensures total integral sums to 1 so it is a valid *pdf*) is:

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta$$

maximal value/θ_{MAP}	**mean/bayesian estimate θ_B**	**variance**
$argmax_\theta[Beta(\theta; \alpha, \beta)]$	$E[\theta]$	$E[\theta^2] - (E[\theta])^2$
$m_{\alpha,\beta} = \dfrac{\alpha - 1}{\alpha + \beta - 2}$	$\mu_{\alpha,\beta} = \dfrac{\alpha}{\alpha + \beta}$	$\sigma^2_{\alpha,\beta} = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



- When $\alpha = \beta$ it is symmetrical about 0.5
- higher values result in steeper/narrower distribution
- The MAP estimate pulls the estimate towards the prior.
- As $\alpha \to 1$ and $\beta \to 1$ $Beta(\theta; \alpha, \beta) \to U(0, 1)$ and $\hat{\theta}_{MAP} \to \hat{\theta}_{MLE}$.

8.5 Computing Terms

8.5.1 Bernoulli Distribution

Prior Posterior

$$\begin{tikzcd}
\theta \sim Beta(\ \theta;\ a,\ b\ ) \arrow[dr] & & \theta \sim Beta(\ \theta;\ \overline{x}n + a,\ n(1-\overline{x}) + b\ ) \\
& \text{sample } \underline{x} = x_1, \ldots, x_n \text{ where the mean is } \overline{x} \arrow[ur] &
\end{tikzcd}$$

Given some $x_i|\theta \sim Bernoulli(\theta)$ we choose the conjugate pair as $\theta \sim Beta(\theta; \alpha, \beta)$ where $\alpha > 1$ and $\beta > 1$.

We have a sample from the distribution: $\underline{x} = x_1, x_2, \ldots, x_n$

Step 1. Given $\theta \sim Beta(\theta; \alpha, \beta)$, the sample $\underline{x} = x_1, x_2, \ldots, x_n$ and sample mean $\overline{x}$ we need to calculate:

$$P(\theta|\underline{x}) = \frac{P(\underline{x}|\theta)P(\theta)}{P(\underline{x})} = \frac{P(\underline{x}|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(\underline{x}|\theta)P(\theta)\ d\theta}$$

We know that the number of 1s in the sample is $\overline{x}n$.

Step 2. First we calculate $P(\underline{x}|\theta)P(\theta)$ using the bernoulli *PMF*:

$$\begin{aligned} P(\underline{x}|\theta) &= \prod_{i=1}^{n} P(x_i|\theta) \\ &= \theta^{\overline{x}n}(1-\theta)^{n-\overline{x}n} \\ &= \theta^{\overline{x}n}(1-\theta)^{n(1-\overline{x})} \end{aligned}$$

By the pdf of the *Beta* distribution:

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

Where B is the beta distribution normalization.

Hence we can multiply to get $P(\underline{x}|\theta)P(\theta)$:

$$\begin{aligned} P(\underline{x}|\theta)P(\theta) &= \theta^{\overline{x}n}(1-\theta)^{n(1-\overline{x})}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \\ &= \frac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)} \end{aligned}$$

Step 3. We derive $P(\theta|\underline{x})$:

$$\begin{aligned} P(\theta|\underline{x}) &= \frac{P(X|\theta)P(\theta)}{P(\int_{-\infty}^{\infty} P(X|\theta)P(\theta))\ d\theta} \\ &= \frac{\dfrac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)}}{\int_{-\infty}^{\infty} \dfrac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)}\ d\theta} \\ &= \frac{\dfrac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)}}{\dfrac{1}{B(\alpha,\beta)}\int_{-\infty}^{\infty} \theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}\ d\theta} \\ &= \frac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{\int_{-\infty}^{\infty} \theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}\ d\theta} \\ &= P(\theta) \ \text{ given } \ \theta \sim Beta(\theta; \overline{x}n+\alpha, n(1-\overline{x})+\beta) \end{aligned}$$

Hence we have the posterior distribution:

$$\theta \sim Beta(\theta; \overline{x}n+\alpha, n(1-\overline{x})+\beta)$$

New Bayesian Estimate *Extra Fun!* **8.5.1**

The new bayesian estimate is a *convex combination* of the *sample mean* $\overline{x}$ and the prior mean (prior bayesian estimate).

$$\begin{aligned}
\hat{\theta}_B &= \frac{\overline{x}n + \alpha}{\overline{x}n + \alpha + n(1-\overline{x}) + \beta} \\
&= \frac{\overline{x}n + \alpha}{\alpha + n + \beta} \\
&= \left(\underbrace{\overline{x}}_{\hat{\theta}_{MLE}} \times \frac{n}{n+\alpha+\beta}\right) + \left(\underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{old } \hat{\theta}_B = \mu_{\alpha,\beta}} \times \frac{\alpha+\beta}{n+\alpha+\beta}\right)
\end{aligned}$$

8.5.2 Normal Distribution - Single Datapoint Sample

Given some $x|\mu \sim N(\mu, \sigma^2)$ where σ^2 is known and μ is unknown. Using a sample of a single datapoint x.

Step 1. The likelihood can be found using the *Normal Distribution PDF*:

$$\begin{aligned}
P(x|\mu) &= f(x|\mu) \\
&= \frac{1}{\sigma\sqrt{2\pi}} \times exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \text{ where } exp\{n\} = e^n
\end{aligned}$$

Hence we now need to calculate the prior (the previous μ value that we will update with our estimate, using the sample):

$$\mu \sim N(\mu_0, \sigma_0^2)$$

Hence we can now calculate the *posterior distribution.*

Step 2. We calculate the *posterior distribution*

$$\begin{aligned}
P(\mu|x) = f(\mu|x) &= \frac{f(x|\mu)f(\mu)}{f(x)} = \frac{f(x|\mu)f(\mu)}{\int_{-\infty}^{\infty} f(x|\mu)f(\mu)\ d\mu} \\
&\vdots \\
&= (\text{some constant}) \times exp\left\{-\frac{\left(\mu - \dfrac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma^2+\sigma_0^2}\right)^2}{2 \times \dfrac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}}\right\}
\end{aligned}$$

We can express the new variance as:

$$\sigma_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2}\right)$$

With the new posterior density function as:

$$\mu|X \sim N(\mu_1, \sigma_1^2)$$

8.5.3 Normal Distribution - Sample Size n

We extend the previous proof for a sample $\underline{x} = x_1, \dots, x_n$ and distribution $x_i|\mu \sim N(\mu, \sigma^2)$ where σ is known.

Step 1. We calculate the likelihood:

$$\begin{aligned}
P(\underline{x}|\mu) &= f(\underline{x}|\mu) = f(x_1|\mu)f(x_2|\mu)\dots f(x_n|\mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \times \prod_{i=1}^{n} exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \times exp\left\{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sigma^n(2\pi)^{n/2}} \times exp\left\{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}\right\}
\end{aligned}$$

And then the prior probability which is distributed by $\mu \sim N(\mu_0, \sigma_0^2)$.

$$P(\mu) = f(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

Step 2. We can then calculate the posterior using *baye's theorem* .

$$\begin{aligned}
P(\mu|\underline{x}) &= \frac{1}{\sigma_0\sqrt{2\pi}} exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \times \frac{1}{\sigma^n(2\pi)^{n/2}} \times expr\left\{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{(2\pi)^{(n+1/2)}\sigma_0\sigma^n} exp\left\{\frac{-\mu^2+2\mu\mu_0-\mu_0^2}{2\sigma_0^2} - \sum_{i=1}^{n} \frac{x_i^2-2\mu x_i+\mu^2}{2\sigma^2}\right\} \\
&\vdots \\
&\propto exp\left\{-\frac{\left(\mu - \dfrac{\mu_0\sigma^2+\sum_{i=1}^{n}\sigma_0^2 x_i}{\sigma^2+n\sigma_0^2}\right)^2}{2\dfrac{\sigma_0^2\sigma^2}{\sigma^2+n\sigma_0^2}}\right\}
\end{aligned}$$

Hence we have:

$$\mu|\underline{x} \sim N(\mu_1, \sigma_1^2)$$

$$\sigma_1^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2} = \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \frac{\mu_0\sigma^2+\sum_{i=1}^{n}\sigma_0^2 x_i}{\sigma^2+n\sigma_0^2} = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2}+\sum_{i=1}^{n}\frac{x_i}{\sigma^2}\right)$$

8.5.4 Normal Distribution - Sufficient Statistic

Sufficient Statistic **Definition 8.5.1**

A statistic is *sufficient* for a given model (our chosen distribution) and its associated parameter if no other statistic can be calculated from a sample that provides additional information in computing the value/estimate of the unknown parameter.

For a *normal distribution* the sufficient statistic is the sample mean:

$$T(\underline{x}) = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Hence we will use the sample mean in calculating our posterior distribution.

Step 1. We can directly calculate the posterior distribution using the likelihood and prior.

$$\begin{aligned}
P(\mu|\underline{x}) = f(\mu|\underline{x}) &= \frac{f(\mu)f(\underline{x}|\mu)}{\int_{-\infty}^{\infty} f(\mu)f(\underline{x}|\mu)\ d\mu} \\
&\propto \frac{f(\mu)f(T(\underline{x})|\mu)}{\int_{-\infty}^{\infty} f(\mu)f(\underline{x}|\mu)\ d\mu} \\
&\propto f(\mu)f(T(\underline{x})|\mu) \\
&= f(\mu)f(\overline{x}|\mu) \\
&= \frac{1}{\sigma_0\sqrt{2\pi}} exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \times \frac{1}{\sqrt{2\pi\dfrac{\sigma^2}{n}}} exp\left\{-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right\} \\
&\vdots \\
&\propto exp\left\{\frac{-\left(\mu - \dfrac{\mu_0{}^{\sigma^2}\!/\!{}_n + \overline{x}\sigma_0^2}{{}^{\sigma^2}\!/\!{}_n + \sigma_0^2}\right)^2}{2\dfrac{\sigma_0^2{}^{\sigma^2}\!/\!{}_n}{{}^{\sigma^2}\!/\!{}_n + \sigma_0^2}}\right\}
\end{aligned}$$

Hence we have the exponential part of the pdf for a normal distribution.

Step 2. We can now compute the posterior distribution.

$$\mu|\underline{x} \sim N(\mu_1, \sigma_1^2)$$

$$\sigma_1^2 = \frac{\sigma_0^2{}^{\sigma^2}\!/\!{}_n}{{}^{\sigma^2}\!/\!{}_n + \sigma_0^2} = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \frac{\mu_0{}^{\sigma^2}\!/\!{}_n + \overline{x}\sigma_0^2}{{}^{\sigma^2}\!/\!{}_n + \sigma_0^2} = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{\overline{x}n}{\sigma^2}\right)$$

Chapter 9

Credit

Image Credit

Front Cover *"A wave with a sunset in the background in the style of a textbook oil painting"* - OpenAI Dall-E.

Content

Based on the statistics course taught by Professor Giuliano Casale and Professor Chiraag Lala.

These notes were written by Oliver Killane.