Probability and Stochastic Modelling 1

Semester 1, 2022

Dr Alexander Browning

Tarang Janawalkar

Contents

1 Events and Probability

1.1 Events

Definition 1.1 (Event)**.** An event is a set of outcomes in a random experiment commonly denoted by a capital letter. Events can be simple (a single event) or compound (two or more simple events).

Definition 1.2 (Sample space)**.** The set of all possible outcomes of an experiment is known as the sample space for that experiment and is denoted Ω.

Definition 1.3 (Intersection)**.** An intersection between two events A and B describes the set of outcomes that occur in both A and B. The intersection can be represented using the set `AND` operator ($\cap$) — $A \cap B$ (or AB).

Definition 1.4 (Disjoint)**.** Disjoint (mutually exclusive) events are two events that cannot occur simultaneously or have no common outcomes.

Theorem 1.1.1 (Intersection of disjoint events)**.** *The intersection of disjoint events results in the null set ($\varnothing$).*

Lemma 1.1.1.1. *Disjoint events are* ***dependent*** *events as the occurrence of one means the other cannot occur.*

Definition 1.5 (Union)**.** A union of two events A and B describes the set of outcomes in either A or B. The union is represented using the set `OR` operator ($\cup$) — $A \cup B$.

Definition 1.6 (Complement)**.** The complement of an event E is the set of all other outcomes in Ω. The complement of E is denoted $\overline{E}$.

Theorem 1.1.2 (Intersection of complement set)**.**

$$A\overline{A} = \varnothing$$

Theorem 1.1.3 (Union of complement set)**.**

$$A \cup \overline{A} = \Omega$$

Definition 1.7 (Subset)**.** A is a (non-strict) subset of B if all elements in A are also in B. This can be denoted as $A \subset B$.

Theorem 1.1.4. *All events E are subsets of Ω.*

Theorem 1.1.5. *Given $A \subset B$*

$$AB = A \qquad \textit{and} \qquad A \cup B = B$$

Corollary 1.1.5.1. *Given $\varnothing \subset E$*

$$\varnothing E = \varnothing \qquad \textit{and} \qquad \varnothing \cup E = E$$

Theorem 1.1.6 (Associative Identities)**.**

$$A\,(BC) = (AB)\,C$$
$$A \cup (B \cup C) = (A \cup B) \cup C$$

Theorem 1.1.7 (Distributive Identities)**.**

$$A\,(B \cup C) = AB \cup AC$$
$$A \cup BC = (A \cup B)\,(A \cup C)$$

1.2 Probability

Definition 1.8 (Probability)**.** Probability is a measure of the likeliness of an event occurring. The probability of an event E is denoted $\Pr(E)$ (sometimes $\mathrm{P}(E)$).

$$0 \leq \Pr(E) \leq 1$$

where a probability of 0 never happens, and 1 always happens.

Theorem 1.2.1 (Probability of Ω)**.**

$$\Pr(\Omega) = 1$$

Theorem 1.2.2 (Complement rule)**.** *The probability of the complement of E is given by*

$$\Pr\left(\overline{E}\right) = 1 - \Pr(E)$$

Theorem 1.2.3 (Multiplication rule for independent events)**.** *The probability of the intersection between two independent events A and B is given by*

$$\Pr(AB) = \Pr(A)\Pr(B)$$

Theorem 1.2.4 (Addition rule for independent events)**.** *The probability of the union between two independent events A and B is given by*

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(AB).$$

If A and B are disjoint, then $\Pr(AB) = 0$, so that $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

Corollary 1.2.4.1 (Addition rule for 3 events)**.** *The addition rule for 3 events is as follows*

$$\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B) + \Pr(C) - \Pr(AB) - \Pr(AC) - \Pr(BC) + \Pr(ABC).$$

Proof. If we write $D = A \cup B$ and apply the addition rule twice, we have

$$\begin{aligned}
\Pr(A \cup B \cup C) &= \Pr(D \cup C) \\
&= \Pr(D) + \Pr(C) - \Pr(DC) \\
&= \Pr(A \cup B) + \Pr(C) - \Pr((A \cup B)C) \\
&= \Pr(A) + \Pr(B) - \Pr(AB) + \Pr(C) - \Pr(AC \cup BC) \\
&= \Pr(A) + \Pr(B) - \Pr(AB) + \Pr(C) - (\Pr(AC) + \Pr(BC) - \Pr(ACBC)) \\
&= \Pr(A) + \Pr(B) + \Pr(C) - \Pr(AB) - \Pr(AC) - \Pr(BC) + \Pr(ABC)
\end{aligned}$$

□

Theorem 1.2.5 (De Morgan's laws)**.** *Recall De Morgan's Laws:*

$$\begin{aligned}
\overline{A \cup B} &= \overline{A}\ \overline{B} \\
\overline{AB} &= \overline{A} \cup \overline{B}.
\end{aligned}$$

Taking the negation of both sides and applying the complement rule yields

$$\begin{aligned}
\Pr(A \cup B) &= 1 - \Pr\left(\overline{A}\ \overline{B}\right) \\
\Pr(AB) &= 1 - \Pr\left(\overline{A} \cup \overline{B}\right)
\end{aligned}$$

1.3 Circuits

A signal can pass through a circuit if there is a functional path from start to finish.
We can define a circuit where each component i functions with probability p, and is independent of other components.
Then W_i to be the event in which the associated component i functions, we can determine the event S in which the system functions, and probability $\Pr(S)$ that the system functions.
As the probability that any component functions is p, in other words

$$\Pr(W_i) = p,$$

$\Pr(S)$ will be a function of p defined $f : [0,\ 1] \to [0,\ 1]$.

2 Independence

Definition 2.1 (Conditional probability)**.** When discussing multiple events, it is possible that the occurrence of one event changes the probability that another will occur. This can be denoted using a vertical bar, and is read as "the probability of event A given B":

$$\Pr(A \mid B) = \frac{\Pr(AB)}{\Pr(B)}.$$

Definition 2.2 (Multiplication rule)**.** For events A and B, the general multiplication rule states that

$$\Pr(AB) = \Pr(A \mid B)\Pr(B)$$

Theorem 2.0.1 (Independent events)**.** *If A and B are independent events then*

$$\Pr(A \mid B) = \Pr(A)$$
$$\Pr(B \mid A) = \Pr(B)$$

Theorem 2.0.2 (Complement of independent events)**.** *If A and B are independent, all complement pairs are also independent. Given*

$$\Pr(A \mid B) = \Pr(A)$$
$$\Pr(B \mid A) = \Pr(B)$$

the following statements are also true

$$\Pr\left(A \mid \overline{B}\right) = \Pr(A) \qquad \Pr\left(B \mid \overline{A}\right) = \Pr(B)$$
$$\Pr\left(\overline{A} \mid B\right) = \Pr\left(\overline{A}\right) \qquad \Pr\left(\overline{B} \mid A\right) = \Pr\left(\overline{B}\right)$$
$$\Pr\left(\overline{A} \mid \overline{B}\right) = \Pr\left(\overline{A}\right) \qquad \Pr\left(\overline{B} \mid \overline{A}\right) = \Pr\left(\overline{B}\right)$$

2.1 Probability Rules with Conditional

All probability rules hold when conditioning on some event C.

Theorem 2.1.1 (Complement rule with condition)**.**

$$\Pr\left(\overline{A} \,|\, C\right) = 1 - \Pr\left(A \,|\, C\right)$$

Theorem 2.1.2 (Addition rule with condition)**.**

$$\Pr\left(A \cup B \,|\, C\right) = \Pr\left(A \,|\, C\right) + \Pr\left(B \,|\, C\right) - \Pr\left(AB \,|\, C\right)$$

Theorem 2.1.3 (Multiplication rule with condition)**.**

$$\Pr\left(AB \,|\, C\right) = \Pr\left(A \,|\, BC\right) \Pr\left(B \,|\, C\right)$$

In the above examples, all probabilities are conditional on the sample space, hence we are effectively changing the sample space.

2.2 Conditional Independence

Definition 2.3 (Conditional independence)**.** Suppose events A and B are not independent, i.e.,

$$\Pr\left(A \,|\, B\right) \neq \Pr\left(A\right)$$

but they become independent when conditioned with another event C, i.e.,

$$\Pr\left(A \,|\, BC\right) = \Pr\left(A \,|\, C\right)$$

Here we say that A and B are **conditionally independent** given C. Furthermore

$$\Pr\left(AB \,|\, C\right) = \Pr\left(A \,|\, C\right) \Pr\left(B \,|\, C\right)$$

Conversely, events A and B may be conditionally dependent but unconditionally independent, i.e.,

$$\begin{aligned} \Pr\left(A \,|\, B\right) &= \Pr\left(A\right) \\ \Pr\left(A \,|\, BC\right) &\neq \Pr\left(A \,|\, C\right) \\ \Pr\left(AB \,|\, C\right) &= \Pr\left(A \,|\, BC\right) \Pr\left(B \,|\, C\right) \end{aligned}$$

Theorem 2.2.1. *Given events A, B, and C. Pairwise independence does not imply mutual independence. I.e.,*

$$\begin{cases} \Pr\left(AB\right) = \Pr\left(A\right) \Pr\left(B\right) \\ \Pr\left(AC\right) = \Pr\left(A\right) \Pr\left(C\right) \\ \Pr\left(BC\right) = \Pr\left(B\right) \Pr\left(C\right) \end{cases}$$

does not imply

$$\Pr\left(ABC\right) = \Pr\left(A\right) \Pr\left(B\right) \Pr\left(C\right).$$

In summary, independence should not be assumed unless explicitly stated.

2.3 Disjoint Events

Theorem 2.3.1 (Probability of disjoint events)**.** *The probability of disjoint events A and B is given by*

$$\Pr(AB) = 0$$
$$\Pr(\varnothing) = 0.$$

Disjoint events are highly dependent events, since the occurrence of one means the other cannot occur. This implies

$$\Pr(A \mid B) = 0$$

2.4 Subsets

Theorem 2.4.1 (Probability of subsets)**.** *If $A \subset B$ then $\Pr(A) \leq \Pr(B)$. We also know that $\Pr(AB) = \Pr(A)$ and $\Pr(A \cup B) = \Pr(B)$.*
Here, if A happens, then B definitely happens.

$$\Pr(B \mid A) = 1$$

Given $\Pr(AB) = \Pr(A)$

$$\Pr(A \mid B) = \frac{\Pr(A)}{\Pr(B)}$$

These events are also highly dependent.

3 Total Probability

Definition 3.1 (Marginal probability)**.** Marginal probability is the probability of an event irrespective of the outcome of another variable.

Theorem 3.0.1 (Total probability for complements)**.** *By writing the event A as $AB \cup A\overline{B}$, and noting that AB and $A\overline{B}$ are disjoint, the marginal probability of A is given by*

$$\Pr(A) = \Pr(AB) + \Pr\left(A\overline{B}\right).$$

By applying the multiplication rule to each joint probability:

$$\Pr(A) = \Pr(A \mid B)\Pr(B) + \Pr\left(A \mid \overline{B}\right)\Pr\left(\overline{B}\right)$$

Theorem 3.0.2 (Law of total probability)**.** *The previous theorem partitioned Ω into disjoint events B and $\overline{B}$.*
By partitioning Ω into a collection of disjoint events $B_1,\ B_2,\ \ldots,\ B_n$, such that $\bigcup_{i=1}^{n} B_i = \Omega$, we have

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i)$$

Theorem 3.0.3 (Bayes' Theorem)**.** *Given the probability for A given B, the probability of the reverse direction is given by*

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

4 Combinatorics

Definition 4.1 (Number of outcomes)**.** Let $|A|$ denote the number of outcomes in an event A.

Theorem 4.0.1 (Addition principle)**.** *Given a sample space S with k disjoint events $\{S_1, \ldots, S_k\}$, where the ith event has n_i possible outcomes, the number of possible samples from any event is given by*

$$|\bigcup_{i=0}^{k} S_i| = \sum_{i=1}^{k} n_i$$

Theorem 4.0.2 (Multiplication principle)**.** *Given a sample space S with k events $\{S_1, \ldots, S_k\}$, where the ith event has n_i possible outcomes, the number of possible samples from every event is given by*

$$|\bigcap_{i=0}^{k} S_i| = \prod_{i=1}^{k} n_i$$

Theorem 4.0.3 (Counting probability)**.** *Given a sample space S with equally likely outcomes, the probability of an event $S_i \subset S$ is given by*

$$\Pr(S_i) = \frac{|S_i|}{|S|}$$

4.1 Ordered Sampling with Replacement

When ordering is important and repetition is allowed, the total number of ways to choose k objects from a set with n elements is

$$n^k$$

4.2 Ordered Sampling without Replacement

When ordering is important and repetition is not allowed, the total number of ways to arrange k objects from a set of n elements is known as a k-permutation of n-elements denoted ${}^nP_k$

$$\begin{aligned} {}^nP_k &= n \times (n-1) \times \cdots \times (n-k+1) \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

for $0 \leq k \leq n$.

Definition 4.2 (Permutation of n elements)**.** An n-permutation of n elements is the permutation of those elements. In this case, $k = n$, so that

$$\begin{aligned} {}^nP_n &= n \times (n-1) \times \cdots \times (n-n+1) \\ &= n! \end{aligned}$$

4.3 Unordered Sampling without Replacement

When ordering is not important and repetition is not allowed, the total number of ways to choose k objects from a set of n elements is known as a k-combination of n-elements denoted ${}^nC_k$ or $\binom{n}{k}$

$$\begin{aligned} {}^nC_k &= \frac{{}^nP_k}{k!} \\ &= \frac{n!}{k!\,(n-k)!} \end{aligned}$$

for $0 \leq k \leq n$. We divide by $k!$ because any k-element subset of n-elements can be ordered in $k!$ ways.

4.4 Unordered Sampling with Replacement

When ordering is not important and repetition is allowed, the total number of ways to choose k objects from a set with n elements is

$$\binom{n+k-1}{k}$$

5 Random Variables and Distributions

6 Random Variables

Definition 6.1 (Random variable)**.** A random variable X is a measurable variable whose value holds some uncertainty.

An event is when a random variable assumes a certain value or range of values.

Definition 6.2 (Discrete random variables)**.** A discrete random variable takes discrete values.

Definition 6.3 (Continuous random variables)**.** A continuous random variable can take any real value.

6.1 Probability Distributions

Definition 6.4 (Probability distribution)**.** The probability distribution of a random variable X is a function that links all outcomes $x \in \Omega$ to the probability that they will occur $\Pr(X = x)$.

Definition 6.5 (Probability mass function)**.** The probability distribution of a discrete random variable X is described by a Probability Mass Function (PMF) p_x.

$$\Pr(X = x) = p_x$$

p_x is a valid PMF provided,

$$\forall x \in \Omega : \Pr(X = x) \geq 0 \qquad \text{and} \qquad \sum_{x \in \Omega} \Pr(X = x) = 1.$$

Definition 6.6 (Probability density function)**.** The probability distribution of a continuous random variable X is described by a Probability Density Function (PDF) $f(x)$.
The probability that X is exactly equal to a specific value is always 0. Therefore we compute probabilities over intervals:

$$\Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x$$

$f(x)$ is a valid PDF provided,

$$\forall x \in \Omega : f(x) \geq 0 \qquad \text{and} \qquad \int_{\Omega} f(x)\,\mathrm{d}x = 1.$$

Definition 6.7 (Cumulative distribution function)**.** The Cumulative Distribution Function (CDF) computes the probability that the random variable is less than or equal to a particular realisation x. For $U = \{k \in \Omega : k \leq x\}$

$$F(x) = \Pr(X \leq x) = \begin{cases} \sum_{u \in U} p_u & \text{for discrete random variables} \\ \int_U f(u)\,\mathrm{d}u & \text{for continuous random variables.} \end{cases}$$

$F(x)$ is a valid CDF if:

1. F is monotonically increasing and continuous
2. $\lim_{x \to -\infty} F(x) = 0$
3. $\lim_{x \to \infty} F(x) = 1$

We can recover the PDF given the CDF, by using the Fundamental Theorem of Calculus.

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x} \int_{-\infty}^{x} f(u)\,\mathrm{d}u = f(x)$$

Definition 6.8 (Complementary CDF)**.** For a continuous random variable X the complement function,

$$\Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F(x)$$

is called the complementary CDF, or the survival function.

6.2 Quantiles

Definition 6.9 (p-Quantile)**.** For a continuous random variable, the p-quantile, x, is defined such that

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u = p.$$

Definition 6.10 (Median)**.** The median, m, is a special p-quartile defined as the value such that

$$\int_{-\infty}^{m} f(u)\,\mathrm{d}u = \int_{m}^{\infty} f(u)\,\mathrm{d}u = \frac{1}{2}.$$

Definition 6.11 (Lower and upper quartile)**.** Likewise the lower quartile and upper quartiles are two values q_1 and q_2 such that

$$\int_{-\infty}^{q_1} f(u)\,\mathrm{d}u = \frac{1}{4}$$

and

$$\int_{-\infty}^{q_2} f(u)\,\mathrm{d}u = \frac{3}{4}.$$

Definition 6.12 (Quantile function)**.** The quantile function is the inverse of the CDF and can be used to find the x that a certain p provides. I.e.,

$$x = F^{-1}(p) = Q(p)$$

6.3 Summary Statistics

Definition 6.13 (Expectation)**.** The expectation $\mathrm{E}(X)$, or $\mathbb{E}(X)$ of a random variable X is its expected value given an infinite number of observations.
The expectation is also known as the mean of the X, denoted μ.

$$\mathrm{E}(X) = \begin{cases} \sum_{x\in\Omega} xp_x & \text{for discrete variables} \\ \int_{\Omega} xf(x)\,\mathrm{d}x & \text{for continuous variables} \end{cases}$$

Theorem 6.3.1. *Using integration by parts, it can be proved that*

$$\mathrm{E}(X) = -\int_{-\infty}^{0} F(x)\,\mathrm{d}x + \int_{0}^{\infty} (1 - F(x))\,\mathrm{d}x$$

Definition 6.14 (Variance)**.** The variance $\mathrm{Var}(X)$, or $\mathbb{V}(X)$ of a random variable X is a measure of spread of the distribution (defined as the average squared distance of each value from the mean). Variance is also denoted as σ^2.

$$\begin{aligned} \mathrm{Var}(X) &= \begin{cases} \sum_{x\in\Omega} (x-\mu)^2 p_x & \text{for discrete variables} \\ \int_{\Omega} (x-\mu)^2 f(x)\,\mathrm{d}x & \text{for continuous variables} \end{cases} \\ &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \end{aligned}$$

Definition 6.15 (Standard deviation)**.** The standard deviation is defined as

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

6.3.1 Transformations

For a simple linear function of a random variable

$$\begin{aligned} \mathrm{E}(aX \pm b) &= a\,\mathrm{E}(X) \pm b \\ \mathrm{Var}(aX \pm b) &= a^2\,\mathrm{Var}(X) \end{aligned}$$

7 Special Discrete Distributions

7.1 Discrete Uniform Distribution

A discrete uniform distribution describes the probability distribution of a single trial in a set of equally likely elements.
A discrete random variable X with a discrete uniform distribution is denoted

$$X \sim \text{Uniform}\,(a,\, b)$$

with

$$\begin{aligned}\Pr\,(X = x) &= \frac{1}{b-a+1} \\ \Pr\,(X \leq x) &= \frac{x-a+1}{b-a+1}\end{aligned}$$

for outcomes $x \in \{a,\, a+1,\, \ldots,\, b-1,\, b\}$. We can also summarise the following:

$$\begin{aligned}\text{E}\,(X) &= \frac{a+b}{2} \\ \text{Var}\,(X) &= \frac{(b-a+1)^2 - 1}{12}\end{aligned}$$

7.2 Bernoulli Distribution

A Bernoulli (or binary) distribution describes the probability distribution of a Boolean-valued outcome, i.e., success (1) or failure (0).
A discrete random variable X with a Bernoulli distribution is denoted

$$X \sim \text{Bernoulli}\,(p)$$

with

$$\begin{aligned}\Pr\,(X = x) &= \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases} \\ &= p^x\,(1-p)^{1-x} \\ \Pr\,(X \leq x) &= \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & k \geq 1 \end{cases}\end{aligned}$$

for a probability $p \in [0, 1]$ and outcomes $x \in \{0,\, 1\}$. We can also summarise the following:

$$\begin{aligned}\text{E}\,(X) &= p \\ \text{Var}\,(X) &= p\,(1-p)\end{aligned}$$

where $(1-p)$ is sometimes denoted as q.

7.3 Binomial Distribution

A binomial distribution describes the probability distribution of the number of successes for n independent trials with the same probability of success p.
A discrete random variable X with a binomial distribution is denoted

$$X \sim \mathrm{B}\left(n,\, p\right)$$

with

$$\Pr\left(X = x\right) = \binom{n}{x} p^x \left(1-p\right)^{n-x}$$
$$\Pr\left(X \leq x\right) = \sum_{u=0}^{x} \binom{n}{u} p^u \left(1-p\right)^{n-u}$$

for number of successes $x \in \{0,\, 1,\, \ldots,\, n\}$.
Here each individual trial is a Bernoulli trial, so that X can be written as the sum of n *independent and identically distributed* (iid) Bernoulli random variables, $Y_1,\, Y_2,\, \ldots,\, Y_n$.

$$X = Y_1 + Y_2 + \cdots + Y_n, \qquad Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}\left(p\right) : \forall i \in \{1,\, 2,\, \ldots,\, n\}\,.$$

We can then summarise the following:

$$\mathrm{E}\left(X\right) = np$$
$$\mathrm{Var}\left(X\right) = np\left(1-p\right)$$

Proof. Given n trials, the probability of x successes will be p^x. Similarly the probability of $n-x$ failures will be $\left(1-p\right)^{n-x}$.

We then consider the number of ways to choose x successes out of n trials, i.e., $\binom{n}{x}$.

The intersection of these three events gives the PMF for a binomial distribution. □

7.4 Geometric Distribution

A geometric distribution describes the probability distribution of the number of trials up to and including the first success, where each trial is independent and has the same probability of success p.
A discrete random variable N with a geometric distribution is denoted

$$N \sim \mathrm{Geom}\left(p\right)$$

with

$$\Pr\left(N = n\right) = \left(1-p\right)^{n-1} p$$
$$\Pr\left(N \leq n\right) = 1 - \left(1-p\right)^{n}$$

for number of trials $n \geq 1$.
We can also summarise the following:

$$\mathrm{E}\left(N\right) = \frac{1}{p}$$
$$\mathrm{Var}\left(N\right) = \frac{1-p}{p^2}$$

7.4.1 Alternate Geometric Definition

We can alternatively consider the number of failures until a success, Y:

$$Y = N - 1$$

Therefore the PMF and CDF for Y are:

$$\Pr(Y = y) = (1-p)^y p$$
$$\Pr(Y \leq y) = 1 - (1-p)^{y+1}$$

for number of failures $y \geq 0$. This gives the following summary statistics using transformation rules:

$$\mathrm{E}(Y) = \frac{1-p}{p}$$
$$\mathrm{Var}(Y) = \frac{1-p}{p^2}$$

7.5 Negative Binomial Distribution

A negative binomial distribution describes the probability distribution of the number of trials until $k \geq 1$ successes, where each trial is independent and has the same probability of success p.
A discrete random variable N with a negative binomial distribution is denoted

$$N \sim \mathrm{NB}(k,\ p)$$

with

$$\Pr(N = n) = \binom{n-1}{k-1} (1-p)^{n-k} p^k$$
$$\Pr(N \leq n) = \sum_{u=k}^{n} \binom{u-1}{k-1} (1-p)^{u-k} p^k$$

for number of trials $n \geq k$. Here each individual trial is a Geometric trial, so that N can be written as the sum of k *independent and identically distributed* (iid) Geometric random variables, $Y_1,\ Y_2,\ \ldots,\ Y_k$.

$$N = Y_1 + Y_2 + \cdots + Y_k, \qquad Y_i \overset{\text{iid}}{\sim} \mathrm{Geom}(p) : \forall i \in \{1,\ 2,\ \ldots,\ k\}.$$

We can then summarise the following:

$$\mathrm{E}(N) = \frac{k}{p}$$
$$\mathrm{Var}(N) = \frac{k(1-p)}{p^2}$$

Proof. Given n trials, the probability of k successes will be p^k. Similarly the probability of $n-k$ failures will be $(1-p)^{n-k}$.
We then consider the number of ways to arrange $k-1$ successes for $n-1$ trials, because the last trial must be a success, i.e., $\binom{n-1}{k-1}$.
The intersection of these three events gives the PMF for a negative binomial distribution. □

7.5.1 Alternate Negative Binomial Definition

We can alternatively consider the number of failures Y until k successes:

$$Y = N - k$$

The PMF and CDF for Y are given by:

$$\Pr(Y = y) = \binom{y+k-1}{k-1}(1-p)^y p^k$$
$$\Pr(Y \leq y) = \sum_{u=0}^{y} \binom{u+k-1}{k-1}(1-p)^u p^k$$

for number of failures $y \geq 0$. This gives the following summary statistics using transformation rules:

$$\mathrm{E}(Y) = \frac{k(1-p)}{p}$$
$$\mathrm{Var}(Y) = \frac{k(1-p)}{p^2}$$

7.6 Poisson Distribution

A Poisson distribution describes the probability distribution of the number of events N which occur over a fixed interval of time λ.
A discrete random variable N with a Poisson distribution is denoted

$$N \sim \mathrm{Pois}(\lambda)$$

with

$$\Pr(N = n) = \frac{\lambda^n e^{-\lambda}}{n!}$$
$$\Pr(N \leq n) = e^{-\lambda} \sum_{u=0}^{n} \frac{\lambda^u}{u!}$$

for number of events $n \geq 0$. We can also summarise the following:

$$\mathrm{E}(N) = \lambda$$
$$\mathrm{Var}(N) = \lambda$$

7.7 Modelling Count Data

If we want to utilise these distributions to model data, we can use the following observations:

- Poisson (mean = variance)
- Binomial (underdispersed, mean > variance)
- Geometric/Negative Binomial (overdispersed, mean < variance)

8 Special Continuous Distributions

8.1 Continuous Uniform Distribution

A continuous uniform distribution describes the probability distribution of an outcome within some interval, where the probability of an outcome in one interval is the same as all other intervals of the same length.
A continuous random variable X with a continuous uniform distribution is denoted

$$X \sim \mathrm{U}\left(a,\, b\right)$$

with

$$f\left(x\right) = \frac{1}{b-a}$$
$$F\left(x\right) = \frac{x-a}{b-a}$$

for outcomes $a < x < b$. We can also summarise the following:

$$\mathrm{E}\left(X\right) = \frac{a+b}{2}$$
$$\mathrm{Var}\left(X\right) = \frac{\left(b-a\right)^2}{12}$$
$$m = \frac{a+b}{2}$$

8.2 Exponential Distribution

An exponential distribution describes the probability distribution of the time between events with rate η.
A continuous random variable T with an exponential distribution is denoted

$$T \sim \mathrm{Exp}\left(\eta\right)$$

with

$$f\left(t\right) = \eta e^{-\eta t}$$
$$F\left(t\right) = 1 - e^{-\eta t}$$

for time $t > 0$. We can also summarise the following:

$$\mathrm{E}\left(X\right) = \frac{1}{\eta}$$
$$\mathrm{Var}\left(X\right) = \frac{1}{12}$$
$$m = \frac{\ln\left(2\right)}{\eta}$$

Proof. By considering an event taking longer than t seconds, we can represent this as nothing happening over the interval $[0, t]$. Using $T \sim \text{Exp}(\eta)$ and $N \sim \text{Pois}(\eta t)$, we have

$$\Pr(T > t) = \Pr(N = 0) = e^{-\eta t}$$

where $\lambda = \eta t$. The CDF for the exponential distribution is then

$$\begin{aligned}\Pr(T < t) &= 1 - \Pr(T > t) \\ &= 1 - e^{-\eta t}.\end{aligned}$$

□

8.3 Memoryless Property

In an exponential distribution with $T \sim \text{Exp}(\eta)$, the distribution of the waiting time $t + s$ until a certain event does not depend on how much time t has already passed.

$$\Pr(T > s + t \,|\, T > t) = \Pr(T > s).$$

The same property also applies in an Geometric distribution with $N \sim \text{Geom}(p)$.

8.4 Normal Distribution

The normal distribution is used to represent many random situations, in particular, measurements and their errors. This distribution arises in many statistical problems and can be used to approximate other distributions under certain conditions.
A continuous random variable X with a normal distribution is denoted

$$X \sim \text{N}(\mu,\, \sigma^2)$$

with

$$\begin{aligned}f(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ F(t) &= \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)\end{aligned}$$

for $x \in \mathbb{R}$ where $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, \text{d}t$ is the error function. We can also summarise the following:

$$\begin{aligned}\text{E}(X) &= \mu \\ \text{Var}(X) &= \sigma^2\end{aligned}$$

Given the complexity of the analytic expressions for the PDF and CDF of the normal distribution, we often use software to numerically determine probabilities associated with normal distributions.

8.5 Standard Normal Distribution

Given $X \sim \text{N}(\mu,\, \sigma^2)$, consider the transformation

$$Z = \frac{X - \mu}{\sigma}$$

so that $Z \sim \text{N}(0,\, 1)$. This distribution is called the standard normal distribution. This allows us to deal with the standard normal distribution regardless of μ and σ.

9 Central Limit Theorem

The central limit theorem states that the sum of independent and identically distributed random variables, when properly standardised, can be approximated by a normal distribution, as the number of elements increases.

9.1 Approximating the Average of Random Variables

Given a set of independent and identically distributed random variables $X_1, \ldots, X_n$ from the distribution X, if $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then we can define $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ so that

$$\mathrm{E}\left(\overline{X}\right) = \mu$$
$$\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

By standardising $\overline{X}$, we can define

$$Z = \lim_{n\to\infty} \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

so that $Z \to \mathrm{N}(0,\, 1)$ as $n \to \infty$.

9.2 Approximating the Sum of Random Variables

Given a set of independent and identically distributed random variables $X_1, \ldots, X_n$ from the distribution X, if $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then we can define $\overline{Y} = \sum_{i=1}^{n} X_i$ so that

$$\mathrm{E}(Y) = n\mu$$
$$\mathrm{Var}(Y) = n\sigma^2$$

Then for large n

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

9.3 Approximating the Binomial Distribution

9.3.1 Normal Distribution

Given a binomial distribution $X \sim \mathrm{B}(n,\, p)$, we can write X as the sum of n independent and identically distributed Bernoulli random variables $X_1, \ldots, X_n$, so that $X_i \sim \text{Bernoulli}(p)$.
Thus by the central limit theorem, we can use a normal approximation for X, provided that n is large.

$$X \approx Y \sim \mathrm{N}(np,\, np(1-p))$$

In general, this approximation is sufficient when $np > 5$ and $n(1-p) > 5$.

9.3.2 Poisson Distribution

When $np < 5$ we can use the Poisson distribution to approximate X with the mean np:

$$X \approx Y \sim \text{Pois}\,(np).$$

When $n\,(1-p) < 5$ we can consider the number of failures $W = n - X$, so that,

$$W \approx Y \sim \text{Pois}\,(n\,(1-p)).$$

9.3.3 Continuity Correction

Given an approximation Y (either Normal or Poisson) for the binomial distribution $X \sim \text{B}\,(n,\, p)$ the equality

$$\Pr\,(X \leq x) = \Pr\,(X < x + 1)$$

must hold for any x. Therefore by adding $\frac{1}{2}$ we apply a continuity correction to the approximate probability:

$$\Pr\left(Y \leq x + \frac{1}{2}\right).$$

9.4 Approximating a Poisson Distribution

Given a set of independent Poisson distributions $X_1, \ldots, X_n$ where $X_i \sim \text{Pois}\,(\lambda)$ so that $\text{E}\,(X_i) = \lambda$ and $\text{Var}\,(X_i) = \lambda$ for all i.
If we consider $X = \sum_{i=1}^{n} X_i$ then

$$\begin{aligned} \text{E}\,(X) &= n\lambda \\ \text{Var}\,(X) &= n\lambda \end{aligned}$$

so that by the central limit theorem, we can use the approximation

$$X \approx Y \sim \text{N}\,(n\lambda,\, n\lambda).$$

In general, this approximation is sufficient when $n\lambda > 10$, and when an accurate approximation is desired, $n\lambda > 20$.

10 Bivariate Distributions

Definition 10.1 (Bivariate probability mass function)**.** The distribution over the joint space of two discrete random variables X and Y is given by a bivariate probability mass function:

$$\Pr\,(X = x,\, Y = y) = p_{x,\,y}$$

for all pairs of $x \in \Omega_1$ and $y \in \Omega_2$. This function must satisfy

$$\forall x \in \Omega_1 : \forall y \in \Omega_2 : \Pr\,(X = x,\, Y = y) \geq 0 \qquad \text{and} \qquad \sum_{y \in \Omega_2} \sum_{x \in \Omega_1} \Pr\,(X = x,\, Y = y) = 1.$$

The joint probability mass function can be shown using a table:

$X = x \backslash Y = y$	y_1	$\cdots$	y_n
x_1	$\Pr(X = x_1,\ Y = y_1)$	$\cdots$	$\Pr(X = x_1,\ Y = y_n)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_n	$\Pr(X = x_n,\ Y = y_1)$	$\cdots$	$\Pr(X = x_n,\ Y = y_n)$

Definition 10.2 (Bivariate probability density function)**.** The distribution over the joint space of two continuous random variables X and Y is given by a bivariate probability density function $f(x,\ y)$ over the intervals $x \in \Omega_1$ and $y \in \Omega_2$.

$$\Pr(x_1 \leq X \leq x_2,\ y_1 \leq Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x,\ y)\,\mathrm{d}y\,\mathrm{d}x$$

This function must satisfy

$$\forall x \in \Omega_1 : \forall y \in \Omega_2 : f(x,\ y) \geq 0 \qquad \text{and} \qquad \int_{x \in \Omega_1} \int_{y \in \Omega_2} f(x,\ y)\,\mathrm{d}y\,\mathrm{d}x = 1.$$

When considering the sum of these two variables, we must consider the appropriate bounds. For $x_1 + y_1 < a < x_2 + y_2$, if:

- $a - y_2 \leq x_1$ and $a - y_1 \leq x_2$:

$$\Pr(X + Y > a) = \int_{x_1}^{a-y_1} \int_{a-x}^{y_2} f(x,\ y)\,\mathrm{d}y\,\mathrm{d}x + \int_{a-y_1}^{x_2} f(x)\,\mathrm{d}x$$

- $a - y_2 \leq x_1$ and $a - y_1 > x_2$:

$$\Pr(X + Y > a) = \int_{x_1}^{x_2} \int_{a-x}^{y_2} f(x,\ y)\,\mathrm{d}y\,\mathrm{d}x$$

- $a - y_2 > x_1$ and $a - y_1 \leq x_2$:

$$\Pr(X + Y > a) = \int_{a-y_2}^{a-y_1} \int_{a-x}^{y_2} f(x,\ y)\,\mathrm{d}y\,\mathrm{d}x + \int_{a-y_1}^{x_2} f(x)\,\mathrm{d}x$$

- $a - y_2 > x_1$ and $a - y_1 > x_2$:

$$\Pr(X + Y > a) = \int_{a-y_2}^{x_2} \int_{a-x}^{y_2} f(x,\ y)\,\mathrm{d}y\,\mathrm{d}x$$

10.1 Marginal Probability

The marginal probability function can be obtained by calculating the probability function of each random variable. Once the function has been determined, we must specify the range of values that variable can take.

Definition 10.3 (Marginal probability mass function)**.**

$$\Pr(X = x) = p_x = \sum_{y \in \Omega_2} \Pr(X = x,\, Y = y)$$
$$\Pr(Y = y) = p_y = \sum_{x \in \Omega_1} \Pr(X = x,\, Y = y)$$

Definition 10.4 (Marginal probability density function)**.**

$$\Pr(X = x) = f(x) = \int_{y_1}^{y_2} f(x,\, y)\, \mathrm{d}y$$
$$\Pr(Y = y) = f(y) = \int_{x_1}^{x_2} f(x,\, y)\, \mathrm{d}x$$

10.2 Conditional Probability

Using the joint probability and marginal probability, we can determine the conditional probability function. Once the function has been determined, we must specify the range of values that variable can take.

Definition 10.5 (Conditional probability mass function)**.**

$$\Pr(X = x \,|\, Y = y) = \frac{\Pr(X = x,\, Y = y)}{\Pr(Y = y)}.$$

It follows that

$$\sum_{x \in \Omega_1} \Pr(X = x \,|\, Y = y) = 1$$

Definition 10.6 (Conditional probability density function)**.**

$$f(x \,|\, y) = \frac{f(x,\, y)}{f(y)}$$

It follows that

$$\int_{x_1}^{x_2} f(x \,|\, y)\, \mathrm{d}x = 1$$

10.3 Independence

Two discrete random variables X and Y are independent if

$$\Pr(X = x \,|\, Y = y) = \Pr(X = x)$$

for all pairs of x and y. From this we can show that

$$\Pr(X = x,\, Y = y) = \Pr(X = x) \Pr(Y = y)$$

for all pairs of x and y. If these random variables are not independent then,

$$\Pr(X = x,\, Y = y) = \Pr(X = x \,|\, Y = y) \Pr(Y = y)$$

Two continuous random variables X and Y are independent if we can express $f(x,\, y)$ as

$$f(x,\, y) \propto g(x)\, h(y)$$

and if the joint range of X and Y do not depend on each other. This leads to

$$f(x \,|\, y) = f(x)\,.$$

10.4 Law of Total Expectation

Given the conditional distribution of $X \,|\, Y = y$, we can compute its expectation and variance. For discrete random variables, the conditional expectation is

$$\mathrm{E}(X \,|\, Y = y) = \sum_{x \in \Omega_1} x \Pr(X = x \,|\, Y = y)$$

For continuous random variables, the conditional expectation is

$$\mathrm{E}(X \,|\, Y = y) = \int_{x_1}^{x_2} x f(x \,|\, y)\, \mathrm{d}x$$

The conditional variance is given by

$$\mathrm{Var}(X \,|\, Y = y) = \mathrm{E}(X^2 \,|\, Y = y) - \mathrm{E}(X \,|\, Y = y)^2$$

When X and Y are independent,

$$\begin{aligned} \mathrm{E}(X \,|\, Y = y) &= \mathrm{E}(X) \\ \mathrm{Var}(X \,|\, Y = y) &= \mathrm{Var}(X) \end{aligned}$$

By treating $\mathrm{E}(X \,|\, Y)$ as a random variable of Y, then we can calculate its expected value such that

$$\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X \,|\, Y)).$$

This is known as the law of total expectation.

10.5 Expectation

The following property holds for both dependent and independent random variables X and Y

$$\mathrm{E}(X \pm Y) = \mathrm{E}(X) \pm \mathrm{E}(Y)$$

If X and Y are independent then

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$$

10.6 Variance of Independent Random Variables

If X and Y are independent then

$$\begin{aligned} \mathrm{Var}(X \pm Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ \mathrm{Var}(XY) &= \mathrm{Var}(X)\,\mathrm{Var}(Y) + \mathrm{E}(X)^2\,\mathrm{Var}(Y) + \mathrm{E}(Y)^2\,\mathrm{Var}(X) \end{aligned}$$

10.7 Covariance

Definition 10.7 (Covariance)**.** Covariance is a measure of the dependence between two random variables, it can be determined using

$$\begin{aligned}\operatorname{Cov}(X,\,Y) &= \operatorname{E}\left((X-\operatorname{E}(X))\,(Y-\operatorname{E}(Y))\right)\\ &= \operatorname{E}(XY)-\operatorname{E}(X)\operatorname{E}(Y)\end{aligned}$$

The covariance of X and Y is:

Positive if an increase in one variable is more likely to result in an increase in the other variable.

Negative if an increase in one variable is more likely to result in a decrease in the other variable.

Zero if X and Y are independent. Note that the converse is not true.

The linear transformation of two random variables have the following covariance

$$\operatorname{Cov}(aX+b,\,cY+d) = ac\operatorname{Cov}(X,\,Y)$$

for constants a, b, c, and d.

Definition 10.8 (Joint expectation)**.** The joint expectation of two discrete random variables is

$$\operatorname{E}(XY) = \sum_{x\in\Omega_1}\sum_{y\in\Omega_2} xy\Pr(X=x,\,Y=y)$$

and for continuous random variables

$$\operatorname{E}(XY) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} xyf(x,\,y)\,\mathrm{d}y\,\mathrm{d}x.$$

10.8 Variance of Dependent Random Variables

If X and Y are dependent then

$$\operatorname{Var}(X\pm Y) = \operatorname{Var}(X)+\operatorname{Var}(Y)\pm 2\operatorname{Cov}(X,\,Y)$$

similarly for the sum of three dependent random variables

$$\operatorname{Var}(X+Y+Z) = \operatorname{Var}(X)+\operatorname{Var}(Y)+\operatorname{Var}(Z)+2\operatorname{Cov}(X,\,Y)+2\operatorname{Cov}(X,\,Z)+2\operatorname{Cov}(Y,\,Z)$$

10.9 Correlation

The covariance of two random variables describes the direction of a relationship, however it does not quantify the strength of such a relationship. The correlation explains both the direction and strength of a linear relationship between two random variables.
The correlation of two random variables X and Y is denoted $\rho(X,\,Y)$

$$\rho(X,\,Y) = \frac{\operatorname{Cov}(X,\,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

where $-1 \le \rho(X,\,Y) \le 1$.
These value can be interpretted as follows:

- $\rho(X,\, Y) > 0$ iff X and Y have a positive linear relationship.
- $\rho(X,\, Y) < 0$ iff X and Y have a negative linear relationship.
- $\rho(X,\, Y) = 0$ if X and Y are independent. Note that the converse is not true.
- $\rho(X,\, Y) = 1$ iff X and Y have a perfect linear relationship with positive slope.
- $\rho(X,\, Y) = -1$ iff X and Y have a perfect linear relationship with negative slope.

Note that the slope of a perfect linear relationship cannot be obtained from the correlation.

11 Markov Chains

A Markov chain is a discrete time and state stochastic process that describes how a state evolves over time. In this process, the set of all states is discrete and disjoint and states change probabilistically so that a step may not result in a changed state. At each step, the next state depends only on the current state of the random variable.
States are denoted by the random variable X_t at time step t.

Definition 11.1 (Markov property)**.** The state X_t is conditionally independent of X_{t-2}, X_{t-3}, ..., X_0 given X_{t-1}.

$$\Pr(X_t = x_t \,|\, X_{t-1} = x_{t-1},\, X_{t-2} = x_{t-2},\, \ldots,\, X_0 = x_0) = \Pr(X_t - x_t \,|\, X_{t-1} = x_{t-1})$$

11.1 Homogeneous Markov Chain

A Markov chain is homogeneous when

$$\Pr(X_{t+n} = j \,|\, X_t = i) = \Pr(X_n = j \,|\, X_0 = i) = p_{ij}^{(n)}$$

that is, the n-step conditional probabilities do not depend on the time step t.

11.2 Transition Probability Matrix

A homogeneous Markov chain is characterised by the transition probability matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$, where m is the number of states. $\mathbf{P}$ must fulfil the following properties:

- $p_{i,j} = \Pr(X_t = j \,|\, X_{t-1} = i)$
- $p_{i,j} \geq 0 : \forall i,\, j$
- $\sum_{j=1}^{m} p_{i,j} = 1 : \forall j$

$\mathbf{P}$ has the following form

$$\mathbf{P} = \begin{matrix} \\ x_t \end{matrix} \overset{x_{t+1}}{\begin{bmatrix} & \\ & \end{bmatrix}}$$

The n-step transition probability is given by $\mathbf{P}^n$.

11.3 Unconditional State Probabilities

The unconditional probability of being in state j at time n is given by

$$\Pr(X_n = j) = p_j^{(n)}$$

Given multiple states, let $\boldsymbol{s}^{(n)}$ denote the vector of all states $p_j^{(n)}$ at time n. Then

$$\boldsymbol{s}^{(n)^\top} = \boldsymbol{s}^{(n-1)^\top}\mathbf{P} \qquad \text{and} \qquad \boldsymbol{s}^{(n)^\top} = \boldsymbol{s}^{(0)^\top}\mathbf{P}^n$$

11.4 Stationary Distribution

At steady-state, the probability of being in a particular state does not change from one step to the next. This implies

$$\boldsymbol{s}^{(n+1)} = \boldsymbol{s}^{(n)} \implies \boldsymbol{s}^{(n)^\top} = \boldsymbol{s}^{(n)^\top}\mathbf{P}$$

The stationary distribution $\boldsymbol{\pi}$ satisfies $\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top\mathbf{P}$. To determine $\boldsymbol{\pi}$, we must use the equation $\sum_{i=1}^{m} \pi_i = 1$.

11.5 Limiting Distribution

Under certain conditions, as $n \to \infty$, each row of $\mathbf{P}^n$ will be equal to $\boldsymbol{\pi}^\top$, so that each state moves to the next step with the same probability. This is known as the limiting distribution. Here $\boldsymbol{\pi}$ provides the long run probabilities of being in each state and the process forgets where it starts.
A sufficient condition for the above is if $\mathbf{P}^n$ has positive entries for some finite n.
Note that a stationary distribution does not imply that a limiting distribution exists.