Discrete probability

Definitions

For some statistical experiment being performed:

- The set of all possible outcomes is called the **sample space**, denoted Ω.
- A subset $E \subseteq \Omega$ is an **event**.

Elementary principles

Events

Let E and F be events from some sample space Ω.

- $E \cup F$ is the event that either (or both) E or F happens
- EF (or $E \cap F$) is the event that both E and F happen
- E^c (or $\bar{E}$) is the event that E does **not** happen ($\Omega - E$)

De Morgan's Law

For events $\{E_i\}_{i=1}^n$:

$$\left(\bigcup_{i=1}^n E_i\right)^c = \bigcap_{i=1}^n E_i^c$$

$$\left(\bigcap_{i=1}^n E_i\right)^c = \bigcup_{i=1}^n E_i^c$$

Axioms

For each event $E \subseteq \Omega$, we assign a probability $\mathbb{P}\left(E\right)$ satisfying:

- $0 \leq \mathbb{P}\left(E\right) \leq 1$
- $\mathbb{P}\left(\Omega\right) = 1$
- For any sequence $\{E_i\}_{i=1}^\infty$ of **mutually exclusive events**:

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \mathbb{P}\left(E_j\right)$$

Where **mutual exclusivity** means ($E_j \cap E_k = \emptyset \quad \forall j \neq k$)

- $\mathbb{P}\left(E^c\right) = 1 - \mathbb{P}\left(E\right)$

Inclusion-exclusion principle

For a finite sequence of arbitrary events $\{E_i\}_{i=1}^n$ where $E_i \subseteq \Omega \quad \forall i$:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{j=1}^{n} \mathbb{P}\left(E_j \cap E_k\right) + \sum_{j<k<l}^{n} \mathbb{P}\left(E_j \cap E_k \cap E_l\right) - \cdots + (-1)^{n+1} \cdot \mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right)$$

Example

> - For $n = 2$
>
> $$\mathbb{P}\left(E_1 \cup E_2\right) = \mathbb{P}\left(E_1\right) + \mathbb{P}\left(E_2\right) - \mathbb{P}\left(E_1 \cap E_2\right)$$
>
> - For $n = 3$
>
> $$\begin{aligned} \mathbb{P}\left(E_1 \cup E_2 \cup E_3\right) = \; & \mathbb{P}\left(E_1\right) + \mathbb{P}\left(E_2\right) + \mathbb{P}\left(E_3\right) \\ & - \mathbb{P}\left(E_1 \cap E_2\right) - \mathbb{P}\left(E_1 \cap E_3\right) - \mathbb{P}\left(E_2 \cap E_3\right) \\ & + \mathbb{P}\left(E_1 \cap E_2 \cap E_3\right) \end{aligned}$$

Random variables

A **random variable** is a function that maps each outcome of the sample space to some numerical value.

Given a sample space Ω, a random variable X with values in some set $\mathcal{R}$ is a function:

$$X : \Omega \mapsto \mathcal{R}$$

> Where $\mathcal{R}$ is typically $\mathbb{N}_0$ or $\mathbb{N}$ in discrete probability and $\mathbb{R}$ in continuous probability.

Discrete random variables

- The random variable X is a **discrete random variable** when its range is finite (or countably infinite).

Continuous random variables

- The random variable X is a **continuous random variable** when its range is uncountably infinite.

Notation

Random variables often make it easier to ask questions such as:

> How likely is it that the value of X is equal to 2?

This is the same as the probability of the event $\{x \in \Omega \mid X(x) = 2\}$, which is often denoted as $\mathbb{P}\left(X = 2\right)$ and read "*the probability of the random variable X taking on the value 2*".

Example

> Let our statistical experiment be the toss of a fair coin. We will perform this experiment 3 times, giving us:
>
> $$\Omega = \{H, T\}^3 = \{(H,H,H), (H,H,T), (H,T,H), \dots, (T,T,T)\}$$
>
> Let X be the random variable denoting the number of heads after 3 coin flips.
>
> - $\mathbb{P}\left(X = 0\right) = \mathbb{P}\left(\{(T,T,T)\}\right) = \frac{1}{8}$
> - $\mathbb{P}\left(X = 1\right) = \mathbb{P}\left(\{(T,T,H), (T,H,T), (H,T,T)\}\right) = \frac{3}{8}$
> - $\mathbb{P}\left(X = 2\right) = \mathbb{P}\left(\{(T,H,H), (H,T,H), (H,H,T)\}\right) = \frac{3}{8}$
> - $\mathbb{P}\left(X = 3\right) = \mathbb{P}\left(\{(H,H,H)\}\right) = \frac{1}{8}$

Note that for the collection $\{\mathbb{P}\left(X = k\right)\}_{k=0}^{3}$, we have:

$$0 \leq \mathbb{P}\left(X = k\right) \leq 1$$
$$\sum_{k=0}^{3} \mathbb{P}\left(X = k\right) = 1$$

As we will see later, this represents a probablity distribution, and these are properties that all probability distributions must have.

Stirling's approximation

Stirling's approximation is an approximation for the factorial operation. It is an accurate estimation, even for smaller values of n.

The approximation is:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Where the $\sim$ sign means that the two quantities are asymptotic. This means that their ratio tends to 1 as n tends to ∞.

Alternatively, there is a version of Stirling's formula with bounds valid for all positive integers n, rather than asymptotics:

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n} \leq n! \leq en^{n+\frac{1}{2}}e^{-n}$$

Distributions

A **probability distribution** is a mathematical function that maps each outcome of a statistical experiment to its probability of occurrence.

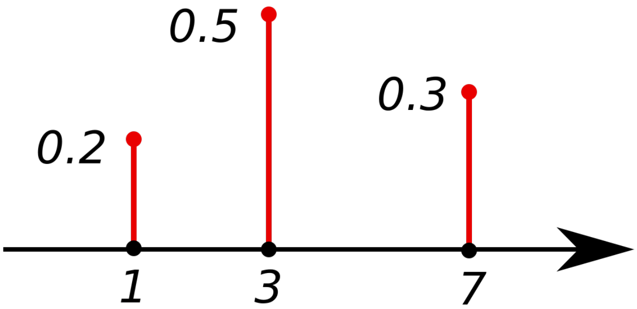
Probability mass function

A **probability mass function** is a function that gives the probability that a discrete random variable is exactly equal to some value. It defines a discrete probability distribution.

Suppose that $X : \Omega \mapsto \mathcal{R}$ is a discrete random variable. Then the probability mass function $f_X : \mathcal{R} \mapsto [0, 1]$ for X is defined as:

$$f_X(x) = \mathbb{P}\left(X = x\right) = \mathbb{P}\left(\{s \in \Omega \mid X(s) = x\}\right)$$

Example



This is the probability mass function of a discrete probability distribution.

In this case, we have a random variable $X : \Omega \mapsto \mathbb{N}$ and a probability mass function $f_X : \mathbb{N} \mapsto [0, 1]$.

Consider the following probabilities as examples:

> - $\mathbb{P}\left(X = 1\right) = \mathbb{P}\left(\{s \in \Omega \mid X(s) = 1\}\right) = \mathbb{P}\left(\{1\}\right) = 0.2$
> - $\mathbb{P}\left(X = 3\right) = \mathbb{P}\left(\{s \in \Omega \mid X(s) = 3\}\right) = \mathbb{P}\left(\{3\}\right) = 0.5$
> - $\mathbb{P}\left(X = 7\right) = \mathbb{P}\left(\{s \in \Omega \mid X(s) = 7\}\right) = \mathbb{P}\left(\{7\}\right) = 0.3$
> - $$\begin{aligned} \mathbb{P}\left(X \geq 1\right) = \mathbb{P}\left(\bigcup_{s \in \Omega} \{s\}\right) &= \sum_{s \in \Omega} \mathbb{P}\left(\{s\}\right) \\ &= \mathbb{P}\left(\{1, 3, 7\}\right) = \mathbb{P}\left(\{1\}\right) + \mathbb{P}\left(\{3\}\right) + \mathbb{P}\left(\{7\}\right) \\ &= 0.2 + 0.5 + 0.7 \\ &= 1 \end{aligned}$$

Conditions

For any probability distribution (with some random variable $X : \Omega \mapsto \mathcal{R}$), its probability mass function must satisfy both of the following conditions:

- $0 \leq \mathbb{P}\left(X = k\right) \leq 1 \qquad \forall k \in \mathcal{R}$
- $\sum_{k \in \mathcal{R}} \mathbb{P}\left(X = k\right) = 1$

Cumulative distribution function

The **cumulative distribution function** of a random variable X evaluated at x is the probability that X will take a value less than or equal to x.

If X is a discrete random variable that maps to values $\{x_i\}_{i=1}^{n}$, then the cumulative distribution function F_X is defined as:

$$F_X(x) = \mathbb{P}\left(X \leq x\right) = \sum_{x_i \leq x} \mathbb{P}\left(X = x_i\right) = \sum_{x_i \leq x} \mathbb{P}\left(\{x_i\}\right)$$

Complementary cumulative distribution function

Sometimes, it is useful to study the opposite question — how often the random variable is **above** a particular value. This is called the **complementary cumulative distribution function** or simply the **tail distribution**, and is denoted $\overline{F_X}(x)$, and is defined as:

$$\begin{aligned}\overline{F_X}(x) &= \mathbb{P}(X > x) \\ &= 1 - F_X(x)\end{aligned}$$

Uniform distribution

A random variable is **uniformly distributed** if every possible outcome is equally likely to be observed. In other words, for some statistical experiment, suppose there are n different outcomes. Then the probability of each outcome is $\frac{1}{n}$.

Therefore, the probability mass function for a uniformly distributed discrete random variable $X : \Omega \mapsto \mathbb{N}_0$ for n possible outcomes would be:

$$\mathbb{P}(X = k) = \frac{1}{n}$$

Parameter	Meaning
$n \in \mathbb{N}$	Number of possible outcomes

Binomial distribution

The **binomial distribution** with parameters n and p is the discrete probability distribution of the number of successes (k) in a sequence of n **Bernoulli trials**.

The probability mass function for a binomially distributed discrete random variable $X : \Omega \mapsto \mathbb{N}_0$ for n Bernoulli trials (each with probability of success p) would be:

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

Parameter	Meaning
$n \in \mathbb{N}_0$	Number of trials
$p \in [0, 1]$	Probability of success in each trial

Quantity (or function)	Formula
Mean (expected value)	$\mathbb{E}[X] = np$
Variance	$\mathrm{Var}[X] = np(1-p)$
Moment-generating function	$M_X(t) = (1 - p + pe^t)^n$

Poisson distribution

The **Poisson distribution** is a discrete probability distribution that expresses the probability of a given number of events occuring in a fixed interval of time or space if these events occur with a known constant rate and independently of time since the last event.

The probability mass function for a Poisson distributed discrete random variable $X : \Omega \mapsto \mathcal{R}$ with some constant rate λ would be:

$$\mathbb{P}\left(X = k\right) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Parameter	Meaning
$\lambda \in \mathbb{R} : \lambda > 0$	Rate

Quantity (or function)	Formula
Mean (expected value)	$\mathbb{E}\left[X\right] = \lambda$
Variance	$\mathrm{Var}\left[X\right] = \lambda$
Moment-generating function	$M_X(t) = e^{\lambda(e^t - 1)}$

Negative binomial distribution

The **negative binomial distribution** is a discrete probability distribution of the number of trials in a sequence of independent and identically distributed Bernoulli trials before a specified number of successes occurs.

The probability mass function for a negative binomially distributed discrete random variable $X : \Omega \mapsto \mathbb{N}_0$ with $n \in \mathbb{N}_0 : n \geq k$ trials given k successes, would be:

$$\mathbb{P}\left(X = n\right) = \binom{n-1}{k-1} p^k \left(1 - p\right)^{n-k}$$

Parameter	Meaning
$k \in \mathbb{N}$ (but can be extended to $\mathbb{R}$)	Number of successes until the experiment is stopped
$p \in [0, 1]$	Success probability in each experiment

Quantity (or function)	Formula
Mean (expected value)	$\mathbb{E}\left[X\right] = \frac{pk}{1-p}$
Variance	$\mathrm{Var}\left[X\right] = \frac{pk}{(1-p)^2}$

Moment-generating function	$M_X(t) = \left(\frac{1-p}{1-pe^t}\right)^r$

Different forms of the distribution

X counts	PMF	Formula	Support
n trials, given k successes	$\mathbb{P}\left(X = n\right)$	$\binom{n-1}{k-1} p^k \left(1-p\right)^{n-k}$	$n \geq k$
r failures, given k successes	$\mathbb{P}\left(X = r\right)$	$\binom{k+r-1}{r} p^k \left(1-p\right)^r$	$r \in \mathbb{N}_0$

Geometric distribution

The **geometric distribution** is a special case of the negative binomial distribution, with the parameter $r = 1$.

The geometric distribution gives the probability that the first occurence of success requires k independent Bernoulli trials, each with success probability p.

The probability mass function for a geometrically distributed discrete random variable $X : \Omega \mapsto \mathbb{N}_0$ with the first success being the k^{th} trial, would be:

$$\mathbb{P}\left(X = k\right) = \left(1-p\right)^{k-1} p$$

Parameter	Meaning
$p \in [0, 1]$	Success probability in each experiment

Quantity (or function)	Formula
Mean (expected value)	$\mathbb{E}\left[X\right] = \frac{1}{p}$
Variance	$\mathrm{Var}\left[X\right] = \frac{1-p}{p^2}$
Moment-generating function	$M_X(t) = \frac{pe^t}{1-(1-p)e^t}$

When to use?

- The phenomenon being modelled is a sequence of independent trials
- There are only two possible outcomes for each trial (success/failure)
- The probability of success, p, is the same for every trial

Hypergeometric distribution

The **hypergeometric distribution** is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, **without replacement**, from a finite population of size N that contains exactly K objects with that feature, where each draw is either a success or failure.

The probability mass function for a hypergeometrically distributed discrete random variable $X : \Omega \mapsto \mathbb{N}_0$ with k **successes**, would be:

$$\mathbb{P}\left(X = k\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Parameter	Meaning
$N \in \mathbb{N}$	Population size
$K \in \{0, 1, 2, \ldots, N\}$	Number of objects with a specific feature
$n \in \{0, 1, 2, \ldots, N\}$	Number of draws

Quantity (or function)	Formula
Mean (expected value)	$\mathbb{E}\left[X\right] = n\frac{K}{N}$

Joint probability

Previously, we introduced $\mathbb{P}\left(A \cap B\right)$ as the probability of the intersection of the events A and B.

If instead, we let these events be described by the random variables:

- A = $X \text{ at value } x$
- $B = Y \text{ at value y}$

Then we can write:

$$\mathbb{P}\left(A \cap B\right) = \mathbb{P}\left(X = x \cap Y = y\right) = \mathbb{P}\left(X = x, Y = y\right)$$

Typically we write $\mathbb{P}\left(X = x, Y = y\right)$, and this is referred to as the **joint probability** of $X = x$ and $Y = y$.

Joint probability distribution

If X and Y are discrete random variables, the function given by $f(x, y) = \mathbb{P}\left(X = x, Y = y\right)$ for each pair of values $(x, y) \in \text{Image}\left(X\right) \times \text{Image}\left(Y\right)$, is called the **joint probability distribution** of X and Y.

Joint cumulative distribution function

If X and Y are discrete random variables, the definition of the **joint cumulative distribution function** of X and Y is given by:

$$F(x, y) = \mathbb{P}\left(X \leq x, Y \leq y\right) = \sum_{s \leq x} \sum_{t \leq y} f(s, t)$$

where $f(s, t)$ is the joint probability distribution of X and Y at (s, t).

Independence of random variables

Consider two discrete random variables X and Y. We say that X and Y are independent if:

$$\mathbb{P}\left(X = x, Y = y\right) = \mathbb{P}\left(X = x\right) \cdot \mathbb{P}\left(Y = y\right) \qquad \forall x, y$$

The definition of independence can be extended to n random variables:

Consider n discrete random variables $\{X_i\}_{i=1}^n$. We say that $\{X_i\}_{i=1}^n$ are **mutually independent** if:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} (X_i = x_i)\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) \qquad \forall x \in \{x_i\}_{i=1}^n$$

Conditional probability

Conditional probability is a measure of the probability of an event, given that some other event has occurred.

If the event of interest is A and the event B is known to have occurred, the conditional probability of A given B is written as:

$$\mathbb{P}\left(A|B\right)$$

Conditioning of an event

Given two events A and B, the conditional probability of A given B is defined as:

$$\mathbb{P}\left(A|B\right) = \frac{\mathbb{P}\left(A \cap B\right)}{\mathbb{P}\left(B\right)} \qquad \text{where } \mathbb{P}\left(B\right) > 0$$

This may be visualised as restricting the sample space to B.

Axiomatic definition

Sometimes the definition of conditional probability is treated as an **axiom of probability**:

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

This is simply a rearrangement of the equation previously shown.

Independent events

Events A and B are said to be **statistically independent** if their joint probability equals the product of the probability of each event:

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

Consequences

- By substituting this into the definition of conditional probability, we get:
  $$\mathbb{P}(A|B) = \mathbb{P}(A)$$
  Intuitively this makes sense, as if A and B are independent, then the fact that event B has already occured should not influence the probability of event A occuring.

General case

Pairwise independence

A finite set of events $\{E_i\}_{i=1}^{n}$ is **pairwise independent** if every pair of events is independent — that is, **iff**:

$$\mathbb{P}(E_m \cap E_k) = \mathbb{P}(E_m) \cdot \mathbb{P}(E_k) \qquad \forall k, m : (1 \leq k \leq n) \wedge (1 \leq m \leq n) \wedge (k \neq m)$$

Mutual independence

A finite set of events is **mutually independent** if every event is independent of any intersection of the other events — that is, **iff** for every k-element subset of $\{E_i\}_{i=1}^{n}$:

$$\mathbb{P}\left(\bigcap_{i=1}^{k} E_i\right) = \prod_{i=1}^{k} \mathbb{P}(E_i)$$

Law of total probability

The **law of total probability** is the proposition that if $\{B_i\}_{i=1}^{n}$ is a finite **partition** of a sample space (in other words, a set of pairwise disjoint events whose union is the entire sample space), then for any event A of the same **probability space**:

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

Bayes' theorem

Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Derivation

Bayes' theorem shows that:

$$\mathbb{P}(A|B) \propto \mathbb{P}(B|A)\,\mathbb{P}(A) \quad \text{and} \quad \mathbb{P}\left(\bar{A}|B\right) \propto \mathbb{P}\left(B|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)$$

In other words, there exists some constant $c \in \mathbb{R}$ such that:

$$\mathbb{P}(A|B) = c \cdot \mathbb{P}(A)\,\mathbb{P}(B|A) \quad \text{and} \quad \mathbb{P}\left(\bar{A}|B\right) = c \cdot \mathbb{P}\left(B|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)$$

If we add these two formulas, we deduce that:

$$\begin{aligned}1 &= c\cdot\mathbb{P}\left(A\right)\mathbb{P}\left(B|A\right)+c\cdot\mathbb{P}\left(B|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)\\ 1 &= c\left[\mathbb{P}\left(A\right)\mathbb{P}\left(B|A\right)+\mathbb{P}\left(B|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)\right]\end{aligned}$$

Therefore, the constant c can be expressed as:

$$\begin{aligned}c &= \frac{1}{\mathbb{P}\left(A\right)\mathbb{P}\left(B|A\right)+\mathbb{P}\left(B|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)}\\ &= \frac{1}{\mathbb{P}\left(B\right)} \qquad \text{(By the law of total probability)}\end{aligned}$$

Definition

Bayes' theorem is then mathematically defined as:

$$\mathbb{P}\left(A|B\right)=\frac{\mathbb{P}\left(A\right)\mathbb{P}\left(B|A\right)}{\mathbb{P}\left(B\right)}=\frac{\mathbb{P}\left(A\right)\mathbb{P}\left(B|A\right)}{\mathbb{P}\left(A\right)\mathbb{P}\left(B|A\right)+\mathbb{P}\left(B|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)}$$

Or alternatively:

$$\mathbb{P}\left(\bar{A}|B\right)=\frac{\mathbb{P}\left(\bar{A}\right)\mathbb{P}\left(B|\bar{A}\right)}{\mathbb{P}\left(B\right)}=\frac{\mathbb{P}\left(\bar{A}\right)\mathbb{P}\left(B|\bar{A}\right)}{\mathbb{P}\left(A\right)\mathbb{P}\left(B|A\right)+\mathbb{P}\left(B|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)}$$

Chain rule

The **chain rule** (or **multiplication rule**) permits the calculation of any member of the **joint distribution** of a set of random variables using only conditional probabilities.

Consider an indexed collection of events $\{E_i\}_{i=1}^{n}$, then we can apply the definition of conditional probability to calculate the joint probability:

$$\mathbb{P}\left(E_n,E_{n-1},\ldots,E_1\right)=\mathbb{P}\left(E_n|E_{n-1},E_{n-2},\ldots,E_1\right)\cdot\mathbb{P}\left(E_{n-1},E_{n-2},\ldots,E_1\right)$$

Repeating this process with each final term creates the product:

$$\mathbb{P}\left(\bigcap_{i=1}^{n}E_i\right)=\prod_{i=1}^{n}\mathbb{P}\left(E_i\middle|\bigcap_{j=1}^{k-1}E_j\right)$$

Example

> With four variables, the chain rule produces this product of conditional probabilities:
>
> $$\mathbb{P}\left(E_4,E_3,E_2,E_1\right)=\mathbb{P}\left(E_4|E_3,E_2,E_1\right)\cdot\mathbb{P}\left(E_3|E_2,E_1\right)\cdot\mathbb{P}\left(E_2|E_1\right)\cdot\mathbb{P}\left(E_1\right)$$

Mutual independence

Two events are **mutually independent** (or **disjoint**) if they cannot both occur. In other words, events A and B are mutually independent **iff** $\mathbb{P}\left(A\cap B\right)=0$.

This has a consequence to the inclusion-exclusion principle. If A and B are mutually independent, then:

$$\begin{aligned}\mathbb{P}\left(A \cup B\right) &= \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right) - \cancelto{0}{\mathbb{P}\left(A \cap B\right)} \\ &= \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right)\end{aligned}$$

Example

> If our statistical experiment is the toss of a fair coin and:
>
> - A is the event that a heads was tossed
> - B is the event that a tails was tossed
>
> Then $\mathbb{P}\left(A\right) = \mathbb{P}\left(B\right) = \frac{1}{2}$, but $A \cap B = \emptyset$ since a coin cannot show heads and tails simultaneously (unless it is some kind of coin that exists in quantum superposition).
>
> Therefore $\mathbb{P}\left(A \cap B\right) = 0$.

Other properties

- $\mathbb{P}\left(A|\bar{B}\right) = 1 - \mathbb{P}\left(\bar{A}|\bar{B}\right)$
- $\mathbb{P}\left(\bar{A}|B\right) = 1 - \mathbb{P}\left(A|B\right)$

Expectation

The **expectation** of a random variable is the probability-weighted average of all possible values.

The expectation of a random variable $X \in \{x_i\}_{i=1}^{k}$ is:

$$\mathbb{E}\left[X\right] = \sum_{i=1}^{k} x_i \cdot \mathbb{P}\left(X = x_i\right)$$

Where the notation $X \in \{x_i\}_{i=1}^{k}$ means that X takes on the values $\{x_i\}_{i=1}^{k}$ (its image consists of the values in this set).

Example

> If $X \in \{1, 2, 3, \ldots, n\}$, where X is a uniformly distributed random variable (with each outcome having a probability of $\frac{1}{n}$), then its expectation is given by:
>
> $$\begin{aligned}\mathbb{E}\left[X\right] &= \sum_{k=1}^{n} k \cdot \mathbb{P}\left(X = k\right) \\ &= \sum_{k=1}^{n} k \cdot \frac{1}{n} \\ &= \frac{1}{n} \sum_{k=1}^{n} k \\ &= \frac{1}{n} \cdot \frac{n(n+1)}{2} \\ &= \frac{n+1}{2}\end{aligned}$$

Moments

Let X be a random variable and $n \in \mathbb{N}$, then $\mathbb{E}\left[X^n\right]$ is called the **n^{th} moment** of X.

In general, if $g(X)$ is a function of X (e.g. X^2, $\ln(X)$), then $g(X)$ is also a random variable.

If $g(X) \in \{x_j\}_{j=1}^{k}$, its expectation is given by:

$$\mathbb{E}\left[g(X)\right] = \sum_{j=1}^{k} g(x_j) \cdot \mathbb{P}\left(X = x_j\right)$$

Therefore, if we let $g(X) = X^n$ where $n \in \mathbb{N}$ and $g(X)$ takes on the values $\{x_j\}_{j=1}^{k}$, then an expression for the n^{th} moment of X would be:

$$\mathbb{E}\left[X^n\right] = \sum_{j=1}^{k} x_j^n \cdot \mathbb{P}\left(X = x_j\right)$$

Properties

Linearity

Linearity of expectation is the property that the expectation of the sum of random variables is equal to the sum of their individual expected values, regardless of whether they are independent.

More formally, for random variables $\{X_i\}_{i=1}^{n}$ and constants $\{c_i\}_{i=1}^{n}$:

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

Example

> $\mathbb{E}\left[X + 2Y\right] = \mathbb{E}\left[X\right] + 2 \cdot \mathbb{E}\left[Y\right]$

Proof of linearity of expectation

Proving the theorem for discrete random variables X and Y, by the basic definition of expectation:

$$\begin{aligned}
\mathbb{E}\left[X + Y\right] &= \sum_{x}\sum_{y}[(x+y) \cdot \mathbb{P}\left(X = x, Y = y\right)] \\
&= \sum_{x}\sum_{y}[x \cdot \mathbb{P}\left(X = x, Y = y\right)] + [y \cdot \mathbb{P}\left(X = x, Y = y\right)] \\
&= \sum_{x} x \underbrace{\sum_{y} \mathbb{P}\left(X = x, Y = y\right)}_{\mathbb{P}(X=x)} + \sum_{y} y \underbrace{\sum_{x} \mathbb{P}\left(X = x, Y = y\right)}_{\mathbb{P}(Y=y)} \\
&= \sum_{x} x \cdot \mathbb{P}\left(X = x\right) + \sum_{y} y \cdot \mathbb{P}\left(Y = y\right) \\
&= \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]
\end{aligned}$$

This result can be extended for n variables using induction.

Other properties

- If $X = c$ where c is some constant, then $\mathbb{E}\left[X\right] = c$.
    - In particular, for any random variable X:

      $\mathbb{E}\left[\mathbb{E}\left[X\right]\right] = \mathbb{E}\left[X\right]$

      This is due to the fact that the expectation of a random variable is simply a constant.
- $\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$ for independent X and Y.

Variance and standard deviation

The **variance** of a random variable X, denoted by $\text{Var}\left[X\right]$, is defined as:

$$\begin{aligned}\text{Var}\left[X\right] &= \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

The variance is a measure of how far a set of numbers are spread out from their mean.

The **standard deviation** of a random variable X, denoted by $\text{SD}\left(X\right)$, is defined as:

$$\text{SD}\left(X\right) = \sqrt{\text{Var}\left[X\right]}$$

Properties

- $\text{Var}\left[aX\right] = a^2\text{Var}\left[X\right] \implies \text{SD}\left(aX\right) = |a|\text{SD}\left(X\right)$
- $\text{Var}\left[aX + b\right] = a^2\text{Var}\left[X\right]$ (Adding a constant to a random variable does not change its variance)
- $\text{Var}\left[X + Y\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right] + 2\text{Cov}\left[X, Y\right]$
    - In the case that X and Y are independent, $\text{Cov}\left[X, Y\right] = 0$, thus:

      $\text{Var}\left[X + Y\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right]$
    - This can also extend to three or more random variables.

      $$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\text{Var}\left[X_i\right] + 2\sum\sum_{i<j}\text{Cov}\left[X_i, X_j\right]$$

      Note that if $\text{Cov}\left[X_i, X_j\right] = 0 \quad \forall i, j : i \neq j$, then:

      $$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\text{Var}\left[X_i\right]$$

      > Example:
      >
      > $\text{Var}\left[X + Y + Z\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right] + \text{Var}\left[Z\right] + 2\text{Cov}\left[X, Y\right] + 2\text{Cov}\left[X, Y\right] + 2\text{Cov}\left[Y, Z\right]$
      >
      > Where all covariances will equal zero if X, Y and Z are mutually independent.

Covariance

The **covariance** of two random variables X and Y, denoted by $\text{Cov}\left[X, Y\right]$, is defined as:

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

Properties

- $\mathrm{Cov}\left[aX, bY\right] = ab\mathrm{Cov}\left[X, Y\right]$
- If $\mathrm{Cov}\left[X, Y\right] = 0$, this does not necessarily mean that X and Y are independent random variables.

  However, if X and Y are independent, then $\mathrm{Cov}\left[X, Y\right] = 0$.

  The contrapositive may also be useful:

  If $\mathrm{Cov}\left[X, Y\right] \neq 0$, then X and Y are dependent.