1 Counting

Basic Principle of Counting

If in sub-experiment i there are n_i possible outcomes for $i = 1, \dots, k$, then there are $n_1 n_2 \dots n_k$ possible outcomes for the experiment.

Definition

The number of ways to distribute n elements into k ordered piles of sizes $n_1, n_2, \dots, n_k$ is given by the **multinomial coefficient**

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1! n_2! \dots n_k!}$$

The Binomial Theorem

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Theorem

The number of ways to choose k elements out of n elements is given by the **binomial coefficient**

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

Definition

The number of **permutations** is the number of ways of distributing n elements into k ordered piles of sizes $n_1, \dots, n_k$:

$$P_n^k = \frac{n!}{n_1! n_2! \dots n_k!}$$

Corollary

There are $n!$ possible ways to order n (distinguishable) items.

$$P_n = n!$$

Theorem

The number of **combinations** is the number of ways of choosing k from n items such that the order of selection does not matter:

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

Definition

The number of **variations** is the number of ways of choosing k out of n items such that the order of selection *does* matter:

- without repetition:

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

- with repetition:

$$n^k$$

2 Axioms of Probability

2.1 Probability Spaces

Definition

A statistical experiment is a random procedure whose **outcome** is uncertain.

The set of all possible outcomes is called the **sample space** Ω. It is "collectively exhaustive".

Subsets of the sample space are called **events**. An event occurs if the outcome of the statistical experiment is an element of the event.

Definition

If $E, F \subset \Omega$ are events, then the following too are events:

1. their union $E \cup F$
2. their intersection $E \cap F$
3. their complement(s) $E^C = \Omega \setminus E$

They are **disjoint (or mutually exclusive)** if $EF = \emptyset$.

De Morgan's Laws

$$(\cup A_i)^C = \cap A_i^C, \qquad (\cap A_i)^C = \cup A_i^C$$

Definition

A **probability space** (Ω, P) incorporates a sample space Ω with collectively exhaustive and mutually exclusive events $E \in \Omega$ and their probabilities P:

$$0 \le P(\omega) \le 1 \quad \forall \quad \omega \in \Omega, \qquad P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1$$

The probability of an event $E_i \subseteq \Omega$ is the sum of the probabilities of the possible outcomes in E_i:

$$P(E_i) \triangleq \sum_{\omega \in E_i} P(\omega) \quad \forall \quad E_i \in \Omega$$

Theorem

A probability space with m different possible outcomes, i.e. with a sample space of size $|\Omega| = m$, has 2^m **possible distinct events**, incl.

- the empty set (which represents an impossible situation, and has probability zero),
- the whole sample space itself (which has $P(\Omega) = 1$),
- all combinations of events in between (all with probability $> 0, < 1$).

2.2 Probability Functions

Definition

A **probability function** P assigns to each event $A \in \Omega$ a value $P(A) \in [0, 1]$, called the probability of the event, such that:

1. the total probability is equal to one:

$$P(\Omega) = \frac{|\Omega|}{|\Omega|} = 1$$

2. the sum rule / inclusion-exclusion formula applies:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$
$$\left(= \frac{|E| + |F| - |E \cap F|}{|\Omega|} \right)$$

Theorem

Given any probability function P, for any event E, its **complement**

$$P(E^C) = 1 - P(E)$$

Definition

A **Laplace experiment** is a statistical experiment with finite sample space Ω such that all outcomes $\omega \in \Omega$ are equally likely. The probability of any given event in a Laplace experiment is

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

2.3 Conditional Probability

Definition

The **conditional probability** of an event A, given an event B, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) \neq 0$$

Corollary

The multiplication rule: For any two events A and B, for which $P(B) \neq 0$,

$$P(A \cap B) = P(B)P(A|B)$$

Example: Random variables W, I are not independent:

```
>>> prob_W_I = np.array([[1/2, 0], [0, 1/6], [0, 1/3]])
>>> prob_W_I
array([[0.5       , 0.        ],
       [0.        , 0.16666667],
       [0.        , 0.33333333]])
>>> prob_W = prob_W_I.sum(axis=1)
>>> prob_I = prob_W_I.sum(axis=0)
>>> prob_W
array([0.5       , 0.16666667, 0.33333333])
>>> prob_I
array([0.5, 0.5])
>>> np.outer(prob_W, prob_I)
array([[0.25      , 0.25      ],
       [0.08333333, 0.08333333],
       [0.16666667, 0.16666667]])
```

Definition

Two events A and B are **independent**, $A \perp B$, if $P(A|B) = P(A)$, $P(B|A) = P(B)$, $P(A \cap B) = P(B)P(A)$. *The outcome of one event is independent of the outcome of the other event(s).*

Multiple events A, B, C may be **mutually (or marginally)**, $A \perp B \perp C$, or **pairwise independent**, $A \perp B$ and / or $A \perp C$ and / or $B \perp C$.

Example: With a fair coin, let's say that we just tossed it five times and tails turned up all five times. Is it more likely now that we'll see heads? No, because the previous outcome doesn't tell us anything about the outcome of a new toss.

Example: Random variables X, Y are independent:

```
>>> prob_X_Y = np.array([[1/4, 1/4], [1/12, 1/12], [1/6
>>> prob_X_Y
array([[0.25      , 0.25      ],
       [0.08333333, 0.08333333],
       [0.16666667, 0.16666667]])
>>> prob_X = prob_X_Y.sum(axis=1)
>>> prob_Y = prob_X_Y.sum(axis=0)
>>> prob_X
array([0.5       , 0.16666667, 0.33333333])
>>> prob_Y
array([0.5, 0.5])
>>> np.outer(prob_X, prob_Y)
array([[0.25      , 0.25      ],
       [0.08333333, 0.08333333],
       [0.16666667, 0.16666667]])
```

2.4 Law of Total Probability and Bayes' Theorem

The Law of Total Probability

Suppose $B_1, \ldots, B_m$ are disjoint events such that

$$B_1 \cup \cdots \cup B_m = \Omega$$

Then for any event A,

$$P(A) = \sum_i P(A|B_i)P(B_i)$$
$$(= P(A|B_1)P(B_1) + \cdots + P(A|B_m)P(B_m))$$

Corollary

In the special case where $m = 2$,

$$P(A) = P(A|B)P(B) + P(A|B^C)P(B^C)$$

Bayes' Theorem

For any two events A and B (with non-zero probabilities,

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

3 Discrete Random Variables

Definition

Let Ω be a sample space. A **discrete random variable** is a function $X : \Omega \to \mathbb{R}$ that takes on a finite number of values $x_1, x_2, \dots, x_n$.

Definition

The **(probability) mass (function)** of a discrete random variable X is $p(X) = P(X = x)$ $= P(X \leq x) - \lim_{y \to x^-} P(X \leq y) = F(x) - \lim_{y \to x^-} F(y)$

Definition

The **(cumulative) distribution (function)** of a discrete random variable X is the non-decreasing, right-continuous function

$$F(x) = F_X(x) = P(X \leq x) \quad s.t.$$

$$\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$$

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

3.1 Expectation and Variance

The expectation is comparable to the centre of mass of an object - the point where it would balance.

The variance and standard deviation measure the spread of a probability distribution around the expectation. Specifically, the standard deviation is the distance from the mean, with 68%, 95% and 99.7% of data points within one, two and three standard deviations from the expectation, respectively.

Definition

Let X be a discrete random variable with probability mass function p_X and possible values $x_1, x_2, \dots$. The **expectation** of X is the weighted average of the possible values of X:

$$E[X] = \sum_i x_i p(x_i) = \sum_i x_i P(X = x_i)$$

Example: Consider the random variable X for which $P(X = -1) = P(X = 1) = p, P(X = 0) = 1 - 2p, 0 < p < \frac{1}{2}$.

$$E[X] = (-1) \cdot p + 1 \cdot p + 0 \cdot (1 - 2p) = 0$$
$$\text{Var}(X) = (-1)^2 \cdot p + 1^2 \cdot p + 0^2 \cdot (1 - 2p) - (E[X])^2 = 2p$$

Theorem

Let f be a function and X a discrete random variable. Then

$$E[g(X)] = \sum_i g(x_i) P(X = x_i)$$

Note: In general, $E[g(X)] \neq g(E[X])$.

Jensen's Inequaltiy

If $E[g(X)]$ and $g(E[X])$ are finite, and $g(x)$ is a **convex** function on R_X, i.e. $f''(x) \geq 0$, then

$$E[g(X)] \geq g(E[X])$$

If $E[g(X)]$ and $g(E[X])$ are finite, and $g(x)$ is a **concave** function on R_X, i.e. $f''(x) \leq 0$, then

$$E[g(X)] \leq g(E[X])$$

Example:

Let $g(x) = x^a \quad \forall \quad a \in \mathbb{R}$. Then $f''(x) = a(a-1)x^{a-2} > 0 \iff a < 0$ or $a > 1$. Therefore,

$$\begin{cases} E[X^a] \geq (E[X])^a & \text{if } a < 0 \text{ or } a > 1 \\ E[X^a] = (E[X])^a & \text{if } a = 0 \text{ or } a = 1 \\ E[X^a] \leq (E[X])^a & \text{if } 0 < a < 1 \end{cases}$$

For the standard deviation, consider the expression $X - E[X]$. It is also a random variable, so one could take its expectation, however, by definition it will always occur that $E[X - E[X]] = 0$. Therefore, a more practical definition is:

Definition

The **variance** of a r.v. X, $\text{Var}(X) = \text{Var}(-X)$, is

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 \geq 0$$

The **standard deviation** of a random variable X is

$$\text{SD}(X) = \sqrt{\text{Var}(X)} \geq 0$$

Change-of-Unit Formulae

Linearity of expectation:

$$E[rX + s] = rE[X] + s$$

$$E[rX + sY] = rE[X] + sE[Y] \qquad E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$$

$$\text{Var}(rX + s) = r^2 \text{Var}(X) \qquad \text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\text{Var}(rX + sY) = r^s \text{Var}(X) + 2\text{Cov}(X, Y) + s^2 \text{Var}(Y)$$

$$\text{Var}(rX + sY) = r^s \text{Var}(X) + s^2 \text{Var}(Y) \iff X \perp Y$$

Proofs:

$$\begin{aligned} E[rX + s] &= \sum_i (rx_i + s) P(X = x_i) \\ &= \sum_i rx_i P(X = x_i) + \sum_i sP(X = x_i) \\ &= r \sum_i x_i P(X = x_i) + s \sum_i P(X = x_i) \\ &= rE[X] + s \end{aligned}$$

$$\begin{aligned} \text{Var}(rX + s) &= E[(rX + s)^2] - (E[rX + s])^2 \\ &= E[r^2 X^2 + 2rsX + s^2] - (rE[X] + s)^2 \\ &= (r^2 E[X^2] + 2rsE[X] + s^2) \\ &\quad - (r^2 (E[X])^2 + 2rsE[X] + s^2) \\ &= r^2 (E[X^2] - (E[X])^2) \\ &= r^2 \text{Var}(X) \end{aligned}$$

3.2 Discrete Distributions

Definition

A random variable X has a **Bernoulli distribution** with parameter p if X can take the values $k = 0, 1$ with probabilities

$$X \sim \text{Ber}(p) : P(X = k) = \begin{cases} 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$$

The geometric distribution describes how many *independent* Bernoulli experiments are needed until the first success.

Definition

A random variable X has a **geometric distribution** with parameter p if X takes on the values $k = 1, 2, 3, \dots$ with probability

$$X \sim \text{Geom}(p) : P(X = k) = p(1 - p)^{k-1}$$

Example: The geometric distribution is memory-less:

$$\begin{aligned} P(X = 11 | X > 7) &= \frac{P(X = 11 \cap X > 7)}{P(X > 7)} \\ &= \frac{P(X = 11)}{P(X > 7)} \\ &= \frac{p(1-p)^{11-1}}{(1-p)^7} \\ &= p(1-p)^3 \end{aligned}$$

The binomial distribution models the number of successes in a sequence of length n.

Definition

A random variable X follows a **binomial distribution** with parameters n, p if X can take on the values $k = 0, 1, 2, \dots, n$ with probabilities

$$X \sim \text{Bin}(n, p) : P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Corollary

$$X \sim \text{Bin}(n, p) : P(X = k + 1) = \frac{p}{1 - p} \cdot \frac{n - k}{k + 1} \cdot P(X = k)$$

Theorem

Given random variables $X_1, X_2, \dots, X_n$, if

1. all X_i are mutually independent, and
2. all $X_i \sim \text{Ber}(p)$ for the same value of p
3. the sample space is fixed, i.e. there is a fixed number of trials n

then the sum of of all X_is is a random variable with binomial distribution:

$$\left(\sum_{i=1}^{n} X_i \right) \sim \text{Bin}(n, p)$$

The **negative binomial distribution** models the number of independent $\text{Ber}(p)$ trials until the r^{th} success: $\sum_{i=1}^{r} \text{Geom}(p)$.

The **hypergeometric distribution** models the number of red balls drawn when we draw n balls out of m red balls and $N - m$ blue balls.

The Poisson distribution models the number of random occurrences of a specific event over a fixed interval (time, space).

Definition

A random variable follows a **Poisson distribution** with parameters λ if X can take on the values $k = 0, 1, 2, \dots$ with probabilities

$$X \sim \text{Pois}(\lambda) : P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Theorem

The Poisson distribution requires the following assumptions:

1. Occurrences take place instantaneously and individually.
2. Occurrences take place independently.
3. The interval under consideration is finite and constant.
4. The average occurrence rate is constant.

Derivation:

Divide a given interval into n sub-intervals. The occurrence of an event in each sub-interval is either 0 or 1:

$$X_m \sim \text{Ber}\left(\frac{\lambda}{n}\right), m = 1, \dots, n$$

By independence, the total number of occurrences in the discrete interval is

$$Y_n = X_1 + \dots + X_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$$

That is, $Y_n \sim \text{Bin}(n, \frac{\lambda}{n})$ has probability mass function

$$P(Y_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

The Poisson distribution is the $\lim_{n \to \infty} \text{Bin}(n, p), \lambda = np$ (and, **for large n and small p, can be used as an approximation for the binomial distribution** $\text{Bin}(n, p)$):

$$\begin{aligned} \lim_{n \to \infty} P(Y_n = k) &= \lim_{n \to \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

Distribution	PMF	$E[X]$	$\text{Var}(X)$
$\text{Ber}(p)$	$\begin{cases} p & x = 1 \\ 1 - p & p = 0 \end{cases}$	p	$p(1-p)$
$\text{Geom}(p)$	$p(1-p)^{(k-1)}$	$1/p$	$(1-p)/p^2$
Bin(n, p)	$\binom{n}{k} p^k (1-p)^{(n-k)}$	np	$np(1-p)$
$\text{Pois}(\lambda)$	$\lambda^k e^{-\lambda}/k!$	λ	λ
$\text{NBin}(r, p)$	$\binom{n-1}{r-1} (1-p)^{n-r} p^r$	r/p	$r(1-p)/p^2$
$\text{HG}(N, m, n)$	$\binom{m}{k} \binom{N-m}{n-k} / \binom{N}{n}$	nm/M	

4 Continuous Random Variables

Theorem

A **probability density function** (pdf) is a function $f : \mathbb{R} \to \mathbb{R}$ such that for any $B \subset \mathbb{R}$,

$$P(X \in B) = \int_B f_X(x)\,dx \quad s.t.$$

$$f_X(x) \geq 0, \qquad \int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

$$P(X = a) = \int_a^a f_X(x)\,dx = 0 \quad \text{for all } a \in \mathbb{R}$$

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

Theorem

For a continuous random variable X with pdf f_X, **cumulative distribution function** (CDF) is the non-decreasing, right-continuous function

$$F(x) = F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x)\,dx \quad s.t.$$

$$\lim_{a \to -\infty} F_X(a) = 0, \qquad \lim_{a \to \infty} F_X(a) = 1, \qquad F(\infty) = 1$$

$$\frac{d}{dx}F(x) = f(x), \text{ where } F(x) \text{ is differentiable}$$

The PDF represents the instantaneous probability density at a point, while the CDF represents the accumulated probability up to that point. This relationship is a direct consequence of the Fundamental Theorem of Calculus.

4.1 Expectation and Variance

Definition

Let X be a continuous random variable with probability density function f_X. Then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx, \quad \text{Var}(X) = \int (x - E[X])^2 f_X(x)\,dx$$

Definition

Let X be a continuous random variable with probability density function f_X. Then:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

Jensen's inequality and the change-of-unit formula apply just like in the discrete case.

4.2 The Exponential Distribution

Definition

A random variable X follows an **exponential distribution** with parameter $\lambda > 0$ if X takes values in $[0, \infty)$ with probability

$$X \sim \text{Exp}(\lambda): \quad f(x) = \lambda e^{-\lambda x}, \qquad F(x) = 1 - e^{-\lambda x}$$

The exponential distribution models the waiting time between events that are modelled by a Poisson distribution.

Theorem

The exponential distribution requires the same **assumptions** to hold as the Poisson distribution:

1. Occurrences take place instantaneously and individually.
2. Occurrences take place independently.
3. The interval is finite and constant.
4. The average occurrence rate is constant.

The exponential distribution also shares the geometric distribution's property of memorylessness:

- Geom(p) counts discrete "failures" until "successes".
- Exp(λ) measures continuous time between "events".

Theorem

Memorylessness: If $X \sim$ Exp then

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

Derivation:

Assume we expect six customers per hour, or one customer every 10 minutes:

$$N \sim \text{Poisson}(6), N_{\frac{1}{6}} \sim \text{Poisson}(1)$$

In general, the "number of events in t time units" N_t is thus:

$$N_t \sim \text{Poisson}(\lambda t)$$

The probability of waiting at least t units between events is the probability of zero events occurring in t units, as modelled by the Poisson distribution:

$$P(X > t) = P(N_t = 0) = e^{-\lambda t}$$

Therefore, the cumulative distribution function of the exponential distribution, the cumulative waiting time, is

$$F(t) = P(X \leq t) = 1 - P(X > t) = 1 - e^{-\lambda t}$$

Differentiation gives the probability density function.

The expectation can be found by found by integration by parts:

$$E[X] = \int_0^{\infty} x \lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}$$

The **Gamma distribution** is the sum of exponentials: $\Gamma(n, \lambda) = \sum_{i=1}^{n} \text{Exp}(\lambda)$

4.3 The Pareto Distribution

Definition

A random variable X follows a **Pareto distribution** with parameter $\alpha > 0$ if X can take on the values in the interval $[1, \infty)$ with probability

$$X \sim \text{Pareto}(\alpha) : f(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & \text{if } x \geq 1 \\ 0 & \text{if } x < 1 \end{cases}$$

$$F(x) = \int f(x)\,dx = 1 - \left(\frac{1}{x}\right)^{\alpha}$$

TODO: The 0.10^{th} quantile is such that $F(x) = 0.10$ (for all distributions).

Theorem

If $X \sim \text{Pareto}(\alpha)$ and $Y = \log(X)$, then $Y \sim \text{Exp}(\alpha)$.

Proof:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\log(X) \leq y) \\ &= P(X \leq e^y) \\ &= F_X(e^y) \\ &= 1 - (\frac{1}{e^y})^{\alpha} \\ &= 1 - e^{-\alpha y} \end{aligned}$$

4.4 The Normal Distribution

Definition

A random variable X follows a **Normal distribution** with parameters μ and σ^2 if X has the following density function for all x:

$$X \sim \mathcal{N}(\mu, \sigma^2) : f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

The integral of this Gaussian function cannot be written using elementary functions. However, the standard normal distribution can be used to calculate probabilities!

Theorem

Linear transformations:
If $X \sim \mathcal{N}(\mu, \sigma^2)$ and r, s are real numbers, then

$$rX + s \sim \mathcal{N}(r\mu + s, r^2\sigma^2)$$

Definition

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

is called the **standard normal** distribution.

Corollary

Any normally distributed random variable can transformed to a standard normal distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Theorem

$$\Phi(-x) = 1 - \Phi(x) \iff P(X \leq -x) = P(X > x)$$

$$F(x) = P(X \leq x) = P\Big(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\Big) = \Phi\Big(\frac{x - \mu}{\sigma}\Big)$$

Distribution	PDF	$E[X]$	$\text{Var}(X)$
$\text{Uni}(a, b)$	$\frac{1}{b-a}, \quad a < x < b$	$\frac{1}{2}(a + b)$	$\frac{1}{12}(b - a)^2$
$\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}, \quad x > 0$	$1/\lambda$	$1/\lambda^2$
$\text{Gamma}(n, \lambda)$	$\frac{(\lambda x)^{n-1}}{\lambda e^{\lambda x}(n-1)!}, \quad x > 0$	n/λ	n/λ^2
$\text{Pareto}(\alpha)$	$\begin{cases} \frac{\alpha}{x^{\alpha+1}} & \text{if } x \geq 1 \\ 0 & \text{if } x < 1 \end{cases}$	$\begin{cases} \infty & \text{for } \alpha \leq 1 \\ \frac{\alpha}{\alpha-1} & \text{for } \alpha > 1 \end{cases}$	$\begin{cases} \infty & \text{for } \alpha \leq 2 \\ \frac{\alpha}{(\alpha-1)^2(\alpha-2)} & \text{for } \alpha > 2 \end{cases}$
$\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Standard normal	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	0	1

DeMoivre-Laplace Theorem

$S_n \sim \text{Bin}(n, p)$ converges to a standard normal

$$\lim_{n \to \infty} P\Big(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\Big) = \Phi(b) - \Phi(a)$$

Thus, there are two **approximations to the binomial distribution**:

- The Poisson distribution is a good approximation when n is large and p is small.
- The standard normal distribution is a good approximation when $np(1 - p)$ is large.

5 Multivariate Random Variables

5.1 Joint Distributions

Definition

Let X and Y be two discrete random variables. The **joint probability mass function** p of X and Y is the function $p : \mathbb{R}^2 \to [0,1]$ defined by

$$p_{X,Y}(x,y) = P(X = x, Y = y) \text{ for all } x, y$$

Their **joint cumulative distribution function** is the function $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ defined by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \text{ for all } x, y$$

Definition

Let $p_{X,Y}(x,y)$ be the joint probability mass function of (X,Y). Then the **marginal probability mass functions** of X and Y are given by

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x,y)$$

$$p_Y(y) = P(Y = y) = \sum_x p_{X,Y}(x,y)$$

This generalises to multiple random variables, e.g. $p_{X,Y}(x,y) = \sum_z p_{X,Y,Z}(x,y,z)$, etc.

NB: The inverse is not true: In general, *the joint distribution cannot be computed based on the knowledge of marginal distributions.* Intuitively, we would be missing key information about the dependencies between X and Y.

Example: Consider random variables (S,T) and (U,V), where S and U have the same distribution, and T and V have the same distribution. However, (S,T) and (U,V) do not necessarily have the same joint distribution:

S / T	1	0
sunny	1 / 2	0
rainy	0	1 / 6
snowy	0	1 / 3

U / V	1	0
sunny	1 / 4	1 / 4
rainy	1 / 12	1 / 12
snowy	1 / 6	1 / 6

Definition

Let X and Y be two discrete random variables with joint probability mass function $p_{X,Y}(x,y)$. The **expectation** of the joint random vector (X,Y) is the vector

$$E[(X,Y)] = \sum_{x,y} (x,y) p_{X,Y}(x,y)$$

Theorem

Let g be a function of two variables. Then

$$E[g(X,Y)] = \sum_{x,y} g(x,y) p_{X,Y}(x,y)$$

Definition

Let X and Y be two continuous random variables. The **joint probability density function** $f_{X,Y}(x,y)$ must satisfy the probability axioms:

1. It is a non-negative function.
2. The volume below its graph is 1.

Thus, the **joint cumulative distribution function** is

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x,y)\, dy\, dx$$

Definition

Let $f_{X,Y}(x,y)$ be the joint probability density function of (X,Y). Then the **marginal probability density functions** of X and Y are given by

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

$$f_Y(y) = P(Y = y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx$$

The **marginal cumulative distribution functions** of X and Y are thus derived as follows:

$$P(X \leq x) = P(X \leq x, Y < \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx$$

$$= \int_{-\infty}^{x} f_X(x)\, dx = F_X(x)$$

Definition

For two continuous jointly-distributed X, Y,

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)\, dy\, dx$$

Theorem

If X and Y are two continuous random variables with joint probability density function $f_{X,Y}$, then

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

Theorem

The distribution of the **sum of two random variables** is

$$F_{X+Y}(a) = \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\, dy$$

and its density is

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\, dy$$

In particular, the sums of random variables following the following discrete and continuous distributions are defined as:

$$\sum_{i=1}^{n} \mathrm{Ber}(p) = \mathrm{Bin}(n,p) \qquad \sum_{i=1}^{n} \mathrm{Geom}(p) = \mathrm{NegBin}(n,p) \qquad \sum_{i=1}^{n} \mathrm{Poisson}(\lambda_i) = \mathrm{Poisson}\Big(\sum_{i=1}^{n} \lambda_i\Big) = \mathrm{Poisson}\Big(\lambda_1 + \cdots + \lambda_n\Big)$$

$$\sum_{i=1}^{n} \mathrm{Exp}(\lambda) = \Gamma(n,\lambda) \qquad \sum_{i=1}^{n} \mathcal{N}(\mu_i, \sigma_i^2) = \mathcal{N}\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\Big) = \mathcal{N}\Big(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2\Big)$$

5.2 Covariance and Correlation

Definition

Let X and Y be two random variables. The **covariance** of X and Y is given by the expectation of the product of the deviations X, Y from the mean

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[(X-E[X])(Y-E[Y])]\\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]]\\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]\\ &= E[XY] - E[X]E[Y]\end{aligned}$$

$$(\text{where } E[XY] = \int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\, dx\, dy)$$

Corollary

The variance can thus be interpreted as the covariance of a random variable with itself:

$$\begin{aligned}\mathrm{Cov}(X,X) &= E[X^2] - (E[X])^2\\ &= \mathrm{Var}(X)\end{aligned}$$

Theorem

The **change-of-unit formula** for the covariance is

$$\begin{aligned}\mathrm{Cov}(rX+s, tY+u) &= E[rtXY + ruX + stY + su]\\ &\quad - (rE[X]+s)(tE[Y]+u)\\ &= rtE[XY] + ruE[X] + stE[Y] + su\\ &\quad - (rtE[X]E[Y] + ruE[X]\\ &\quad\quad + stE[Y] + su)\\ &= rt\mathrm{Cov}(X,Y)\end{aligned}$$

Definition

Let X and Y be two random variables. The **correlation**

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

such that

1. $-1 \le \rho(X,Y) \le 1$ (i.e. it does not depend on units)
2. $\rho(rX+s, tY+u) = \rho(X,Y)$ if $rt > 0$

5.3 Independent Random Variables

Definition

Let X and Y be two discrete random variables. X and Y are **independent** random variables if their joint probability mass function $p_{X,Y}$ is the product of the two marginal mass functions p_X and p_Y:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

Definition

Let X and Y be two continuous random variables. X and Y are **independent** random variables if their joint probability density function $f_{X,Y}$ is the product of the two marginal density functions f_X and f_Y:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

This has important implications for conditional probabilities:

Theorem

If X and Y are independent random variables, **events** concerning X and Y are also independent.

Theorem

If X and Y are independent random variables, then they are **uncorrelated**:

$$E[XY] = E[X]E[Y] \text{ and } \mathrm{Cov}(X,Y) = 0 = \rho(X,Y)$$

NB: The converse is not necessarily true; if two random variables are uncorrelated, they are not necessarily independent.

5.4 The Bivariate Normal Distribution

Definition

Let X and Y be two continuous random variables with joint probability density function

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 + (\frac{y-\mu_Y}{\sigma_Y})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{x-\mu_X}{\sigma_X})\right)}$$

where $\mu_X, \mu_Y \in \mathbb{R}, \sigma_X$ and σ_Y are positive, and $-1 < \rho < 1$. Then (X,Y) follow a **bivariate normal distribution**.

The **marginal distributions** are the standard normal ones. ρ models the effect of two random variables on each other:

$$\rho(X,Y) = \frac{E[XY] - E[X]E[Y]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\, dx\, dy - \mu_X\mu_Y}{\sigma_X\sigma_Y}$$

Theorem

While this is not generally true; for a bivariate normal distribution it can be shown that if $\rho = 0$, then X and Y are independent:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{1}{2}\left((\frac{x-\mu_X}{\sigma_X})^2 + (\frac{y-\mu_Y}{\sigma_Y})^2\right)} = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2}(\frac{x-\mu_X}{\sigma_X})^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2}(\frac{y-\mu_Y}{\sigma_Y})^2} = f_X(x)f_Y(y)$$

6 Conditional Distributions

6.1 Discrete Conditional Distributions

Definition

Let X and Y be two discrete random variables. Then $X|Y = y$ is a **conditional random variable**.

Its **discrete conditional pmf** is

$$P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

or

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

for all x, provided $P(Y = y) \neq 0$.

Theorem

Let X and Y be two discrete random variables such that

$$P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

Then the marginal distribution can be calculated from just one conditional and one joint probability:

$$P(Y = y) = \frac{P(X = x \cap Y = y)}{P(X = x|Y = y)}$$

for any x.

Furthermore, Bayes' Rule can be derived from this definition:

$$P(X = x \cap Y = y) = P(Y = y)P(X = x|Y = y) = P(X = x)P(Y = y|X = x)$$

$$P(Y = y|X = x) = P(X = x|Y = y)\frac{P(Y = y)}{P(X = x)}$$

6.2 Continuous Conditional Distributions

Definition

Let X and Y be two continuous random variables with joint density function $f_{X,Y}(x, y)$. Let y be such that $f_Y(y) > 0$. Then the **conditional pdf** $f_{X|Y}$ is

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

and if X, Y are independent:

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

Furthermore, as with any regular density function:

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\, dx$$

$$F_{X|Y}(a|y) = P(X \leq a|Y = y) = \int_{-\infty}^{a} f_{X|Y}(x|y)\, dx$$

6.3 Conditional Expectation and Variance

Definition

Let X and Y be two continuous random variables and let y be such that $f_Y(y) > 0$. Then

$$E[X|(Y = y)] = \int x f_{X|Y}(x|y)\, dx$$

$$\text{Var}[X|(Y = y)] = \int (x - E[X|(Y = y)])^2 f_{X|Y}(x|y)\, dx$$

Theorem

Let $g(x, y)$ be a function.

$$E[g(X, Y)|(Y = y)] = \int f(x, y) f_{X|Y}(x|y)\, dx$$

Theorem

$$X|(Y = y) \sim N\Big(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y), (1 - \rho^2)\sigma_X^2\Big)$$

7 Limits

Markov Inequality

For any random variable X and positive constant $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Proof: Since

$$E[X] = \int_0^\infty x\, f(x)\, dx \geq \int_a^\infty x\, f(x)\, dx,$$

it follows that

$$P(X \geq a) = \int_a^\infty f(x)\, dx \leq \frac{1}{a}\int_a^\infty x\, f(x)\, dx = \frac{E[X]}{a}$$

7.1 The Law of Large Numbers (LLN)

Convergence in probability is about the convergence of a given estimator $\bar{\theta}_n$ itself.

Definition

If $X_1, X_2, \ldots, X_n$ are independent and all have the same distribution, then they constitute a **random sample**.

The larger the sample size, the lower the variance.

Theorem

Consider a random sample $X_1, X_2, \ldots, X_n$ with $E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$.

$$\begin{aligned} E[\bar{X}_n] &= E[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)] \\ &= \frac{1}{n}E[X_1 + X_2 + \cdots + X_n] \\ &= \frac{1}{n}n\mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\bar{X}_n) &= \mathrm{Var}\Big(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\Big) \\ &= \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n^2}n\sigma^2 \qquad \text{(by independence)} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

By Chebyshev's inequality:

$$P(|\bar{X}_n - E[\bar{X}_n]| \geq a) \leq \frac{\mathrm{Var}(\bar{X}_n)}{a^2} = \frac{\mathrm{Var}(X_i)}{a^2 n}$$

LLN

If $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with $E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$ for all i, then for any $a > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq a) = 0$$

Counter-example: If the variance is infinite, e.g. as for some parameters of a Pareto distribution, the mean never converges, and the Law of Large Numbers does not hold (because the variance is not finite).

Chebyshev's Inequality

The probability that a random variable X is more than a distance $a > 0$ from its expectation is

$$P(|X - E[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

$$\text{If } a = k\sigma, \quad P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof:

$$\begin{aligned} \mathrm{Var}(X) &= E[(X-\mu)^2] = \int_{-\infty}^\infty (x-\mu)^2 f(x)\, dx \\ &\geq \int_{|x-\mu|\geq a} (x-\mu)^2 f(x)\, dx \geq \int_{|x-\mu|\geq a} a^2 f(x)\, dx \\ &= a^2 P(|X - \mu| \geq a) \end{aligned}$$

7.1 The Central Limit Theorem (CLT)

Convergence in distribution is about the convergence of distributions associated with a given estimator $\bar{\theta}_n$.

CLT for Sums of RVs

If $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with $E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$ for all i, then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ as $n \xrightarrow{d} \infty$

Under these conditions, the **sample mean** $\bar{X}_n$ is approximately normally distributed with parameters $\mu, \frac{\sigma^2}{n}$:

$$\bar{X}_n \approx N\big(\mu, \frac{\sigma^2}{n}\big) \iff \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1)$$

CLT

Therefore, the **sum of the samples is approximately normally distributed** with parameters $n\mu, n\sigma^2$:

$$\sum_{i=1}^n X_i \approx \mathcal{N}(n\mu, n\sigma^2)$$

As a **rule of thumb**: $n \geq 30$.

7.2 Approximation to the Binomial

Theorem

A binomially distributed random variable $Y \sim \mathrm{Bin}(n, p)$ is the sum of n independent Bernoulli random variables and, by the CLT, can be approximated by a normal distribution:

$$Y \approx \mathcal{N}(np, np(1-p))$$

As a **rule of thumb**: both the expected numbers of successes and failures, $np \geq 5, n(1-p) \geq 5$.

Proof:

$$Y \sim \mathrm{Bin}(n,p) \iff Y = \sum_{i=1}^n X_i \text{ s.t. } X_i \sim \mathrm{Ber}(p)$$

$$E[X_i] = p, \quad \mathrm{Var}(X_i) = p(1-p)$$

$$E[Y] = np, \quad \mathrm{Var}(Y) = np(1-p), \quad Y \approx \mathcal{N}(np, np(1-p))$$

8 Stochastic Processes

A stochastic process is a collection of random variables describing the state of the system at a particular point in time.

8.1 Bernoulli Processes

Bernoulli processes deal with a sequence of independent trials, where each trial has two possible outcomes (often denoted success and failure). These trials are indexed by a natural number (often representing the order in which they occur).

Definition — Theorem

Description: Text

$$\mathbf{v} \cdot \mathbf{w} = 0 \iff \alpha = \frac{\pi}{2} \iff \mathbf{v} \perp \mathbf{w}$$

8.2 Poisson Processes

Poisson processes model the arrival of events over continuous time. The random variable represents the number of arrivals in a specific time interval, until a time t.

Definition — Theorem

A Poisson process is a collection of random variables $\{N(t) : N(t) \sim \text{Poisson}(\lambda t), t \geq 0\}$ such that

1. $N(0) = 0$
2. $N(s) \leq N(t)$ for $s < t$
3. For all t and for "small" h,

$$P\big(N(t+h) = n + m | N(t) = n\big) = \begin{cases} 1 - \lambda h + o(h), & m = 0 \\ \lambda h + o(h), & m = 1 \\ o(h), & m \geq 2 \end{cases}$$

   where $o(h)$ means "anything much smaller than h", i.e. $\lim_{h \to 0} o(h)/h = 0$

4. State transitions are independent, i.e.

$$\big(N(t_4) - N(t_3)\big) \perp \big(N(t_2) - N(t_1)\big) \text{ for } t_1 < t_2 < t_3 < t_4$$

9 Markov Chains

A stochastic process is a sequence of random variables that evolve over time or space. The key idea is that the outcome at any given point depends on chance (i.e., is random).

A Markov chain adds an additional layer to this concept. It has the Markov property, which states that the probability of transitioning to the next state depends only on the current state, and not on any previous states. In simpler terms, the future depends only on the present, and the past is irrelevant. This ”memoryless” property is what distinguishes Markov chains from more general stochastic processes.

So, all Markov chains are stochastic processes, but not all stochastic processes are Markov chains.

9.1 Markov Chains

Definition

A **Markov chain** is a sequence of random variables $X_0, X_1, X_2, \dots$ taking values in a countable state space such that

$$\begin{aligned} &P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\ &\quad = P(X_{n+1} = j | X_n = i) \end{aligned}$$

for all integers $n > 0$ and states $j, i, i_0, \dots, i_{n-1}$.

$P_{i,j} := P(X_{n+1} = j | X_n = i)$ is the **transition probability** from state i to state j, for which the probability axioms hold:

1. $P_{i,j} \geq 0$
2. $\sum_j P_{i,j} = 1$

Note: It follows that

$$\begin{aligned} &P(X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_1 = i_1 | X_0 = i_0) \\ &\quad = P_{i_0,i_1} P_{i_1,i_2} \dots P_{i_{n-1},i_n} \end{aligned}$$

Definition

Markov chain **transition matrices** are square matrices with all non-negative entries whose rows sum to one.

The Chapman-Kolmogorov Theorem

Let $P^{(n)}_{i,j} = P(X_{m+n} = j | X_m = i)$ denote the n-step transition probability from state i to state j. Then for any $r < n$:

$$P^{(n)}_{i,j} = \sum_k P^{(r)}_{i,k} P^{(n-r)}_{k,j} \text{ , or equivalently, } P^{(n)} = P^{(r)} P^{(n-r)}$$

Corollary

By induction, $P^{(n)} = P^n$ for all $n \geq 1$.

Theorem

Let $X_0, X_1, X_2, \dots$ be a Markov chain with transition matrix P. Let P be such that there exists an integer n with $P^n_{i,j} > 0$ for all states i, j. Then there exists a unique distribution π_j such that the transition probabilities are convergent:

$$\lim_{n \to \infty} P^n_{i,j} = \pi_j$$

A convergent Markov chain is called **ergodic**, and its unique solution is the row vector π such that

$$\pi = \pi P \iff \pi^T = P^T \pi^T, \text{ and } \sum_j \pi_j = 1$$

When looking at the long-term behaviour of a Markov chain, the initial value does not matter - if there is a power of P which does not contain zeros. So *if there is a number of steps n such that you can go from any state to any state in that amount of steps, then this power of P will converge.*

The long-term probabilities π are such that multiplying P again by π doesn't change anything. Therefore, the row vector π can be computed as the eigenvector of the transposed matrix P^T for which the sum of the entries is equal to one.

Example:

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.4 & 0.4 \\ 0.4 & 0.3 & 0.3 \end{bmatrix}, P^2 = \begin{bmatrix} 0.39 & 0.33 & 0.28 \\ 0.34 & 0.34 & 0.32 \\ 0.38 & 0.33 & 0.29 \end{bmatrix},$$

$$P^5 = \begin{bmatrix} 0.37 & 0.33 & 0.3 \\ 0.37 & 0.33 & 0.3 \\ 0.37 & 0.33 & 0.3 \end{bmatrix}, P^{100} = \begin{bmatrix} 0.37 & 0.33 & 0.3 \\ 0.37 & 0.33 & 0.3 \\ 0.37 & 0.33 & 0.3 \end{bmatrix},$$

$$\pi P^{100} = \begin{bmatrix} 0.37 & 0.33 & 0.3 \end{bmatrix} P^{100} = \begin{bmatrix} 0.37 & 0.33 & 0.3 \end{bmatrix} = \pi$$

Definition

Let $X_0, X_1, X_2, \dots$ be a Markov chain with transition matrix P.

A state j is **accessible** from the state i if there exists an integer n such that $P^n_{i,j} > 0$.

Two states i and j are **communicating** if they are each accessible to another: $i \leftrightarrow j$.

A set of communicating states is an **equivalence class**.

A Markov chain in which all states are communicating is called **irreducible**.

Theorem

Communicating states have the following properties:

- Reflexivity: $i \leftrightarrow i$ (the 0-step)
- Symmetry: If $i \leftrightarrow j$ then $j \leftrightarrow i$.
- Transitivity: If $i \leftrightarrow k$ and $k \leftrightarrow j$ then $i \leftrightarrow j$.

Definition

Let i be a state of a Markov chain with transition matrix P. The period $d(i)$ of the state i is the greatest common divisor of all integers $n \geq 1$ with positive return probabilities $P^n_{i,j}$:

$$d(i) := \gcd(\{n \geq 1 : P^n_{i,j} > 0\})$$

Periodicity: State i is called periodic if $d(i) > 1$ or aperiodic if $d(1) = 1$.

Definition

Let $i \leftrightarrow i$ be two communicating states. Then $d(i) = d(j)$.

Definition

Let i be a state of a Markov chain. Let ρ_i be the probability that the Markov chain returns to the state i in a finite non-zero number of steps:

$$\rho_i = P(X_n = i \text{ for some } n > m | X_m = i)$$

If $\rho_i = 1$, the state i is **recurrent**.
If $\rho_i < 1$, it is **transient**.

Theorem

Let $i \leftrightarrow j$ be two communicating states. Then i is recurrent if and only if j is recurrent.

It can be shown that the sum of n-step probabilities diverges if a state is recurrent - a useful check for recurrence in more complicated chains:

Theorem

A state i is recurrent if and only if $\sum_{n=1}^{\infty} P^n_{i,j} = \infty$

Example: Consider a random walk with $r = 1 - p - q = 0$ and

- $P(X_{n+1} = x_n + 1 | X_n = x_n) = p > 0$
- $P(X_{n+1} = x_n - 1 | X_n = x_n) = q = 1 - p > 0$

If $p \neq q$, the random walk is transient as there is a non-zero probability that the chain converges to ∞ or $-\infty$.

Each state is recurrent if and only if $p = q$.

9.2 Finite-State Markov Chains

Definition — Theorem

Description: Text

$$\mathbf{v} \cdot \mathbf{w} = 0 \iff \alpha = \frac{\pi}{2} \iff \mathbf{v} \perp \mathbf{w}$$

9.3 Steady-State Behaviour of Markov Chains

Definition — Theorem

Description: Text

$$\mathbf{v} \cdot \mathbf{w} = 0 \iff \alpha = \frac{\pi}{2} \iff \mathbf{v} \perp \mathbf{w}$$

9.4 Absorption Probabilities and Expected Time to Absorption

Definition — Theorem

Description: Text

$$\mathbf{v} \cdot \mathbf{w} = 0 \iff \alpha = \frac{\pi}{2} \iff \mathbf{v} \perp \mathbf{w}$$

10 Bayesian Inference

10.1 Information Theory

The variance is one way of measuring uncertainty.

Information theory provides perhaps the cleanest derivations for measuring uncertainty and randomness in some of the learning algorithms we will derive!

In this section, we answer the following questions in terms of bits:

1. How do we measure how random an event is?
2. How do we measure how random a random variable or a distribution is?
3. How do we measure how different two distributions are?
4. How much information do two random variables share?

Information theory is often used to show what the best possible performance we should even hope an inference algorithm can achieve such as fundamental limits to how accurate we can make a prediction. And if you can show that your inference algorithm's performance meets the fundamental limit, then that certifies that your inference algorithm is optimal! Inference and information theory are heavily intertwined!

10.1.1 Shannon Information Content

Definition

The Shannon Information Content is a measure of how random an event is in terms of bits: How many yes / no questions do I have to ask to get the answer?

$$\log_2\left(\frac{1}{P(A)}\right)$$

Example: The number of bits needed to store an event A, e.g. an integer from $0, 1, \ldots, 63$ is $\log_2(64) = 6$ bits.

We don't a priori know which of the 64 possible outcomes is going to be stored, and so each outcome is equally likely with probability $\frac{1}{64}$. Then the number of bits needed to store such an event A is given by its Shannon Information Content:

$$\log_2\left(\frac{1}{P(\text{integer is } x)}\right) = \log_2\left(\frac{1}{1/64}\right) = \log_2(64) = 6 \text{ bits}$$

10.1.2 Shannon Entropy

Whereas variance measures how far a random variable is expected to deviate from its expected value, $\text{Var}(X) = E[(X - E[X])^2]$, entropy measures how many bits are needed on average to store each i.i.d. sample of a random variable X.

Definition

The entropy of a random variable is the expectation of its Shannon information content:

$$H(X) = E\left[\log_2\left(\frac{1}{p_X(x)}\right)\right] = \sum_x p_X(x)\log_2\left(\frac{1}{p_X(x)}\right)$$

That is, on avergae, the number of bits needed to encode each i.i.d. sample of a random variable X is $H(X)$.

Example: If X is a fair coin toss, then

$$H(X) = p_X(H)\log_2\frac{1}{p_X(H)} + p_X(T)\log_2\frac{1}{p_X(T)} = \frac{1}{2}\log_2\frac{1}{1/2} + \frac{1}{2}\log_2\frac{1}{1/2} = 1 \text{ bits}$$

If X is a biased coin toss where heads occurs with probability one, then

$$H(X) = p_X(H)\log_2\frac{1}{p_X(H)} + p_X(T)\log_2\frac{1}{p_X(H)} = 1\log_2\frac{1}{1} + 0\log_2\frac{1}{0} = 0 \text{ bits}$$

Theorem

Given a n i.i.d. random samples from p_X,

1. there is an algorithm that is able to store these n samples in $nH(X)$ bits, and
2. this is a lower bound - it is not possible to store the sequence in fewer than $nH(X)$ bits.

10.1.3 Kullback-Leibler / Information Divergence (Relative Entropy)

Definition

KL / Information Divergence (or relative entropy) is a measure of how different two distributions p and q are (over the same alphabet) by the number of bits needed to encode a sample from p using information content according to q instead of according to p:

$$\begin{aligned}
D(p||q) &= E_{X\sim p}\Big[\log_2 \frac{1}{q(X)}\Big] - E_{X\sim p}\Big[\log_2 \frac{1}{p(X)}\Big] \\
&= \sum_x p(x)\log_2 \frac{1}{q(x)} - \sum_x p(x)\log_2 \frac{1}{p(x)} \\
&= \sum_x p(x)\Big(\log_2 \frac{1}{q(x)} - \log_2 \frac{1}{p(x)}\Big) \\
&= \sum_x p(x)\log_2 \frac{p(x)}{q(x)}
\end{aligned}$$

Gibbs' Inequality

For any two distributions p, q defined over the same alphabet,

$$D(p||q) \geq 0$$

where equality holds if and only if p, q are the same distribution, i.e. $p(x) = q(x)$ for all x.

Theorem

Gibbs' Inequality makes information divergence seem a bit like a distance. However, it is not like a distance, as it is not symmetric: in general,

$$D(p||q) \neq D(q||p)$$

Example: Suppose p is the distribution for a fair coin flip, while q is the distribution for a biased coin that always comes up heads. Then

$$\begin{aligned}
D(p||q) &= p(H)\log_2 \frac{p(H)}{q(H)} + p(T)\log_2 \frac{p(T)}{q(T)} \\
&= \frac{1}{2}\log_2 \frac{\frac{1}{2}}{1} + \frac{1}{2}\log_2 \frac{\frac{1}{2}}{0} \\
&= \infty \text{ bits}
\end{aligned}$$

$$\begin{aligned}
D(q||p) &= q(H)\log_2 \frac{q(H)}{p(H)} + q(T)\log_2 \frac{q(T)}{p(T)} \\
&= 1\log_2 \frac{1}{\frac{1}{2}} + 0\log_2 \frac{0}{\frac{1}{2}} \\
&= 1 \text{ bit}
\end{aligned}$$

10.1.4 Mutual Information

Definition

For two discrete random variables X, Y, the mutual information between X and Y, denoted as $I(X;Y)$, measures how much information they share. Specifically,

$$I(X;Y) \triangleq D(p_{X,Y}||p_X p_Y)$$

where $p_X p_Y$ is the distribution we get if X and Y were actually independent (i.e., if X and Y were actually independent, then we know that the joint probability table would satisfy $P(X = x, Y = y) = p_X(x)p_Y(y)$).

Mutual information can be thought of as how far X and Y are from being independent, since if indeed they were independent, then $I(X;Y) = 0$.

On the other hand, if X and Y are the same, then the number of bits they share is exactly the average number of bits needed to store X (or Y), namely $H(X)$ bits:

$$\begin{aligned}
I(X;Y) &= D(p_{X,Y}||p_X p_Y) \\
&= \sum_x \sum_y p_{X,Y}(x,y) \log_2 \frac{1}{p_X(x)p_Y(y)} - \sum_x \sum_y p_{X,Y}(x,y) \log_2 \frac{1}{p_{X,Y}(x,y)} \\
&= \sum_x \sum_y p_X(x) \mathbf{1}\{x=y\} \log_2 \frac{1}{p_X(x)p_Y(y)} - \sum_x \sum_y p_X(x) \mathbf{1}\{x=y\} \log_2 \frac{1}{p_X(x)\mathbf{1}\{x=y\}} \\
&= \sum_x p_X(x) \log_2 \left(\frac{1}{p_X(x)}\right)^2 - \sum_x p_X(x) \log_2 \frac{1}{p_X(x)} \\
&= 2\sum_x p_X(x) \log_2 \frac{1}{p_X(x)} - \sum_x p_X(x) \log_2 \frac{1}{p_X(x)} \\
&= \sum_x p_X(x) \log_2 \frac{1}{p_X(x)} \\
&= H(X)
\end{aligned}$$

10.2 Graphical Models

Inference with graphical models is a specialized technique within Bayesian inference that takes advantage of the graphical structure for efficient computations in problems with complex variable dependencies.

Definition

An **undirected pairwise graphical model** for random variables $X_1, \ldots, X_n$ consists of an undirected graph with vertices $V = \{1, \ldots, n\}$ and edges E, and tables ϕ_i's and $\varphi_{i,j}$'s that have non-negative entries. The joint probability table of $X_1, \ldots, X_n$ is given by

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i \in V} \phi_i(x_i) \prod_{(i,j)\in E} \varphi_{i,j}(x_i, x_j), \text{ where}$$

- Z is the normalisation constant that ensures that the probability distribution actually sums to 1
- each table ϕ_i depends only on random variable X_i and is called the **node potential function** or node potential of node i
- each table $\varphi_{i,}$ depends only on random variables X_i and X_j and is called the **pairwise function** or pairwise potential or edge potential of nodes i and j
- the potential tables need to consist of non-negative entries but each potential table does not have to sum to 1; the constant Z will ensure that the joint probability table actually sums to 1

Definition

A **tree** is a graph for which there are no loops, and we can reach from any node to any other node (moving along edges in the graph).

Theorem

For any graph that has n nodes, if the graph is a tree, then it will always have exactly $n-1$ edges.

Proof by induction:

Base case $n = 1$: there is only 1 node and so there are no edges, so the claim clearly holds.

Inductive step: suppose the claim holds for every tree of size (i.e. number of nodes) up to k. Thus, every tree of size k nodes has $k-1$ edges. Now consider a tree T with $k+1$ nodes. Take a leaf node v from T and note that the tree T with v removed is a tree T' of size k, which by the inductive hypothesis has $k-1$ edges. Since v is a leaf node though, it has exactly 1 neighbour, which means that the tree T has 1 more edge than the tree T', i.e. T has k edges.

Definition

Given a graphical model with graph $G = (V, E)$ and its associated node and edge potentials, the two fundamental inference tasks we focus on are as follows:

- **Marginalisation**: Compute marginal probability table p_{X_i} for every $i \in V$.

- **Most probable configuration**: Compute the most probable configuration $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ such that

$$(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) = \arg \max_{x_1,x_2,\ldots,x_n} p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

The most probable configuration given observation(s) is precisely the MAP estimate!

10.2.1 Maximum A Posteriori (MAP) Estimation

Often times, we want to report which particular value of X achieves the highest posterior probability $p_{X|Y}(\cdot|y)$, i.e. the most probable value x that X can take on given that we have observed $Y = y$.

Definition

The value that X can take on that maximises the posterior distribution is called the **maximum a posteriori (MAP) estimate** of X given $Y = y$:

$$\hat{x}_{MAP}(y) = \arg\max_{x} p_{X|Y}(x|y)$$

10.3 Learning Probabilistic Models

Thus far in the course, we've been given a probabilistic model of the uncertain world, from which we produced predictions given observations. But where do these probabilistic models come from? We now turn to the problem of learning such models (also referred to as **model selection** since we are selecting which model to use).

There are two levels of learning we consider:

- **Parameter learning**: Suppose we know what the edges are in an undirected graphical model but we don't know what the table entries should be for the potentials – how do we estimate these entries?
- **Structure learning**: What if we know neither the parameters nor which edges are present in an undirected graphical model? In this case, we could first figure out what edges are present. After we decide on which edges are present, then the problem reduces to the first problem of parameter learning.

In both cases, the high-level setup is the same: there is some underlying probability distribution p that we don't know details for but want to learn. The distribution p has some parameter (or a set of parameters) θ. We will assume that we can collect n independent samples $X^{(1)}, \ldots, X^{(n)}$ from the distribution p (these n samples are often referred to as "training data"). Given these samples, we aim to estimate θ using what's called **maximum likelihood**, which tries to learn a model that in some sense best fits the training data we have available.

10.3.1 Parameter Learning

Maximum Likelihood Estimation (MLE):

Consider a random variable X, that for simplicity can be thought of as a (possibly biased) coin flip, taking on values in the set $\chi = \{\text{heads}, \text{tails}\}$. The probability of heads, θ, is unknown, and we'd like to estimate (or "learn") this probability. The probability that $X = x$ is denoted as $p_X(x; \theta)$.

We assume we have flipped the coin n times to get outcomes $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ which are i.i.d. samples from the same distribution as X. The **likelihood** of the data is defined as the probability of seeing the observed data as a function of the unknown parameter θ (where the observed data are treated as fixed constants):

$$p_{X^{(1)},X^{(2)},\ldots,X^{(n)}}(x^{(1)}, x^{(2)}, \ldots, x^{(n)}; \theta) = \prod_{i=1}^{n} p_X(x^{(i)}; \theta), \text{ for example} = \theta \cdot (1-\theta) \cdot \cdots \cdot \theta$$

Maximum likelihood estimation maximises the likelihood function L over possible values of the parameter θ:

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \prod_{i=1}^{n} p_X(x^{(i)}; \theta)$$

Note that the value of θ that maximises the likelihood is the same as the value of θ that maximises the log of the likelihood, because the log function is strictly increasing; and mathematically it's often (not always!) easier to work with the log of the likelihood.

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} L(\theta) = \arg\max_{\theta \in [0,1]} \log(L(\theta)) = \arg\max_{\theta \in [0,1]} l(\theta)$$

Naive Bayes Classification:

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Laplace smoothing or additive smoothing can be applied for labels that didn't appear in the train set, to avoid zero probability for these labels.

10.3.2 Structure Learning

At this point we've covered parameter learning for undirected trees, where we assume we know the tree structure. But what if we don't know what tree to even use? We now look at an algorithm called the Chow-Liu algorithm that learns which tree to use from training data, again using maximum likelihood. Once more, information measures appear, where mutual information plays a pivotal role. Recall that mutual information tells us how far two random variables are from being independent. A key idea here is that to determine whether an edge between two random variables should be present in a graphical model, we can look at mutual information.

Trees: The Chow-Liu Algorithm

References

[1] MIT. *Computational Probability and Inference*. 2024. URL: https://www.edx.org/learn/probability/massachusetts-institute-of-technology-computational-probability-and-inference/.

[2] Ross, Sheldon. *A First Course in Probability*. 2012.

[3] TU Delft. *Mastering Probability Theory and Statistics*. 2024. URL: https://online-learning.tudelft.nl/programs/mastering-probability-and-statistics/.