Probability Fundamentals

Probability and Statistics

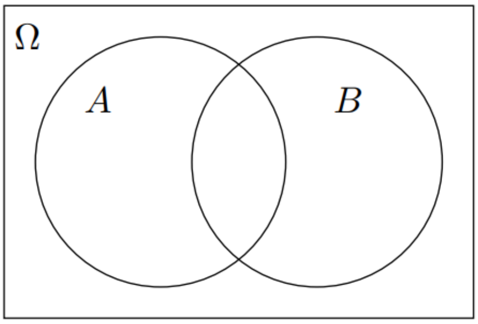
Probability theory is a branch of mathematics that deals with uncertain events.

Statistics is the analysis and interpretation of data.

Foundations of Probability

A **sample space** Ω is the set of possible outcomes of a random experiment. A sample space can be a discrete finite set, a discrete countably infinite set such as the set of integers, or a continuous set such as the set of real numbers.

An **event** is a subset of Ω.



A **probability measure** p is a function that assigns numbers in $\mathbb{R}$ to events, such that the following axioms holds.

> For any event $A \subset \Omega$, the probability of any event is non-negative.

$p(A) \geq 0$

> The probability of the certain event is 1.

$p(\Omega) = 1$

> For any events A and B with empty intersection $A \cap B = \emptyset$, the probability of the union of disjoint events is the sum of their probabilities.

$p(A \cup B) = p(A) + p(B)$

Based on these three axioms, a number of further properties of probability measures can be deduced:

Complement rule:

$p(\Omega - A) = p(\overline{A}) = 1 - p(A)$

General addition rule:

$p(A \cup B) = p(A) + p(B) - p(A \cap B)$

If $A \subseteq B$ then $p(A) \leq p(B)$.

The empty event $\emptyset$ is also called the **impossible** event.

$p(\emptyset) = 0$

The probability of an intersection of events $p(A \cap B)$ is sometimes denoted $p(A, B)$ and called the **joint probability** of the events.

$$p(A, B) + p(A, \overline{B}) = p(A)$$

The **conditional probabilty** of an event conditioned on an event of non-zero probability is defined as the joint probability divided by the probability of the event:

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A \cap B)}{p(B)}$$

If the conditioning event has probability zero, the conditional probability is undefined.

Product rule:

$$p(A, B) = p(A|B)p(B)$$

The conditional probability can be seen as a way to transfer the probability measure on Ω to a re-defined random experiment where the conditional event is the new sample space.

Bayes' theorem:

$$p(B|A) = \frac{p(B,A)}{p(A)} = \frac{p(A|B)p(B)}{p(A)}$$

Random Variables

A **random variable** is a scalar-valued function of the outcomes of a random experiment, i.e., a function that assigns elements in Ω to numbers.

The **probablity distribution** or the **probability mass function** of a random variable X:

$$P_X(x) = p(X = x)$$

The **cumulative probablity function** of X:

$$F_X(x) = p(X \leq x)$$

The joint probability distribution of X and Y:

$$P_{XY}(x, y) = p(X = x \cap Y = y)$$

The conditional probability distribution of Y given X:

$$P_{Y|X}(y|x) = p(Y = y|X = x) = \frac{P_{XY}(x,y)}{P_X(x)}$$

If $\mathcal{Y}$ is the set of all values taken on by the random variable Y:

$$\sum_{y \in \mathcal{Y}} P_{XY}(x, y) = P_X(x)$$

This sum rule for probability distributions is also known as the **marginalisation** of joint probability distributions and allows us to recover probability distributions of individual variables (also called their **marginal** distributions) from their joint distributions.

$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)} = \frac{P_{Y|X}(y|x)P_X(x)}{\sum_{x' \in \mathcal{X}} P_{Y|X}(y|x')P_X(x')}$$

Independence

Two events A and B are said to be **independent** if their joint probability factors into the product of their individual probabilities.

$$p(A, B) = p(A) \cdot p(B)$$

If two events A and B are independent and B has **non-zero** probability, the probability of A knowing B is the same as the probability of A without knowing B.

$p(A|B) = \frac{p(A,B)}{p(B)} = p(A)$

Two random variables X and Y are independent if all the events corresponding to values of X are independent of all the events corresponding to values of Y.

$P_{XY}(x, y) = P_X(x) P_Y(y) \ for \ all \ (x, y) \ in \ \mathcal{X} \times \mathcal{Y}$

Expectation and Entropy

The **expectation** of a random variable, also called the **mean** or **average**, is defined as:

$E[X] = \sum_{x \in \mathcal{X}} x P_X(x)$

$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) P_X(x)$

Expectations are linear operators and fufil the following two **linearity** properties.

$E[X + Y] = E[X] + E[Y]$

$E[cX] = cE[X]$

The expectation of a product of **independent** random variables is the product of their expectations.

$E[XY] = E[X] E[Y]$

The expectation is also called the **first moment** of a distribution while the **second moment** is defined as:

$E[X^2] = \sum_{x^2 \in \mathcal{X}} x P_X(x)$

The **central second moment** or **variance** is defined as:

$\mathrm{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$

The information content of the value of a random variable is defined as:

$h(x) = \log_2 \frac{1}{P_X(x)}$

It is a measure of our surprise when we observe this particular value. If the probability distribution assigns a small probability to the value, then its information content will be large and we will be more surprised if it occurs.

The average of the information content is a measure of our uncertainty about a random variable.

$H(X) = E[h(X)] = \sum_{x \in \mathcal{X}} P_X(x) h(x) = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{1}{P_X(x)}$

It is known as Shannon's **entropy** and its unit is the **bit** when the base of the logarithm is 2.

Discrete Probability Distributions

The Bernoulli Distribution

A binary random variable X with a probability distribution $P_X(1) = p$ and $P_X(0) = 1 - p$ is said to have a **Bernoulli distribution** with parameter p, denoted $X \sim \mathrm{Ber}(p)$.

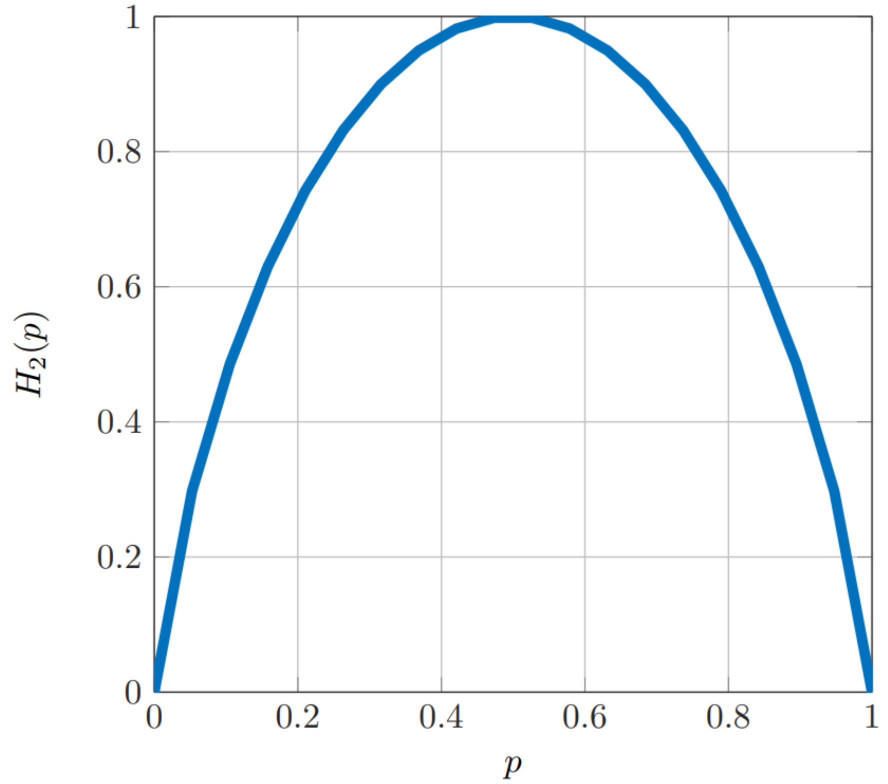
$E[X] = P_X(1) = p$

$E[X^2] = P_X(0) \cdot 0^2 + P_X(1) \cdot 1^2 = p$

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$$

The entropy of a Bernoulli random variable is known as the **binary entropy function** of p.

$$H_2(p) = H(X) = p \log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{1-p}$$

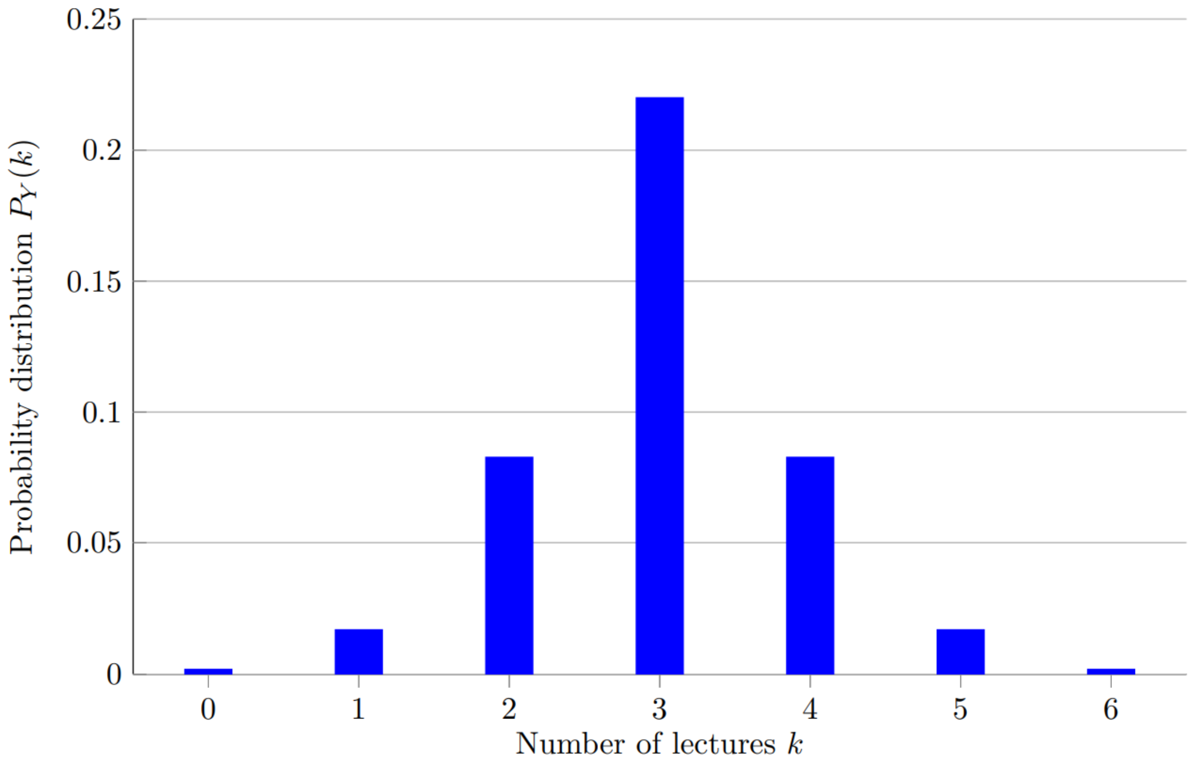


The Binomial Distribution

Consider n independent $\mathrm{Ber}(p)$ distributed random variables $X_1, X_2, \cdots, X_n$. The random variable $Y = \sum_{k=1}^{n} X_k$ is said to follow a **binomial distribution** with parameters n and p, denoted $Y \sim B(n, p)$.

$$P_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{k=1}^{n} P_{X_k}(x_k) = p^{\sum_k x_k}(1-p)^{n-\sum_k x_k}$$

$$P_Y(k) = \binom{n}{k} p^k (1-p)^{n-k}$$



$$E[Y] = E[X_1] + \cdots + E[X_n] = nE[X_1] = np$$

$$E[Y^2] = E[(X_1 + \cdots + X_n)^2] = nE[X_1^2] + 2\binom{n}{2} E[X_1]^2 = np + n(n-1)p^2$$

$$\operatorname{Var}[Y] = E[Y^2] - E[Y]^2 = np + n(n-1)p^2 - (np)^2 = np(1-p)$$

The Geometric Distribution

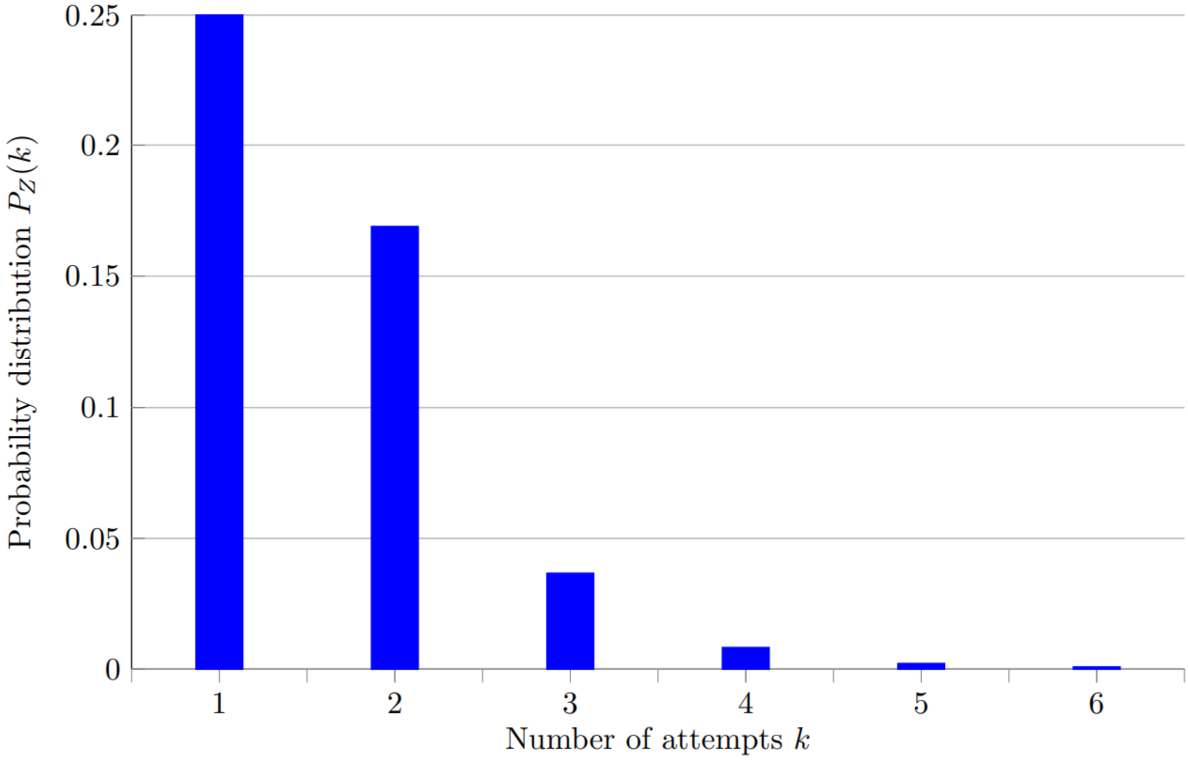
The **geometric distribution** can be derived from a collection of independent Bernoulli random variable as the distribution of the index of the first 1 in the sequence. Hence, Y is a geometric distributed random variable derived from an infinite collection $X_1, X_2, \cdots$ of independent $\mathrm{Ber}(p)$ random variables

$$P_Y(k) = p(1-p)^{k-1}$$

$$E[Y] = \frac{1}{p}$$

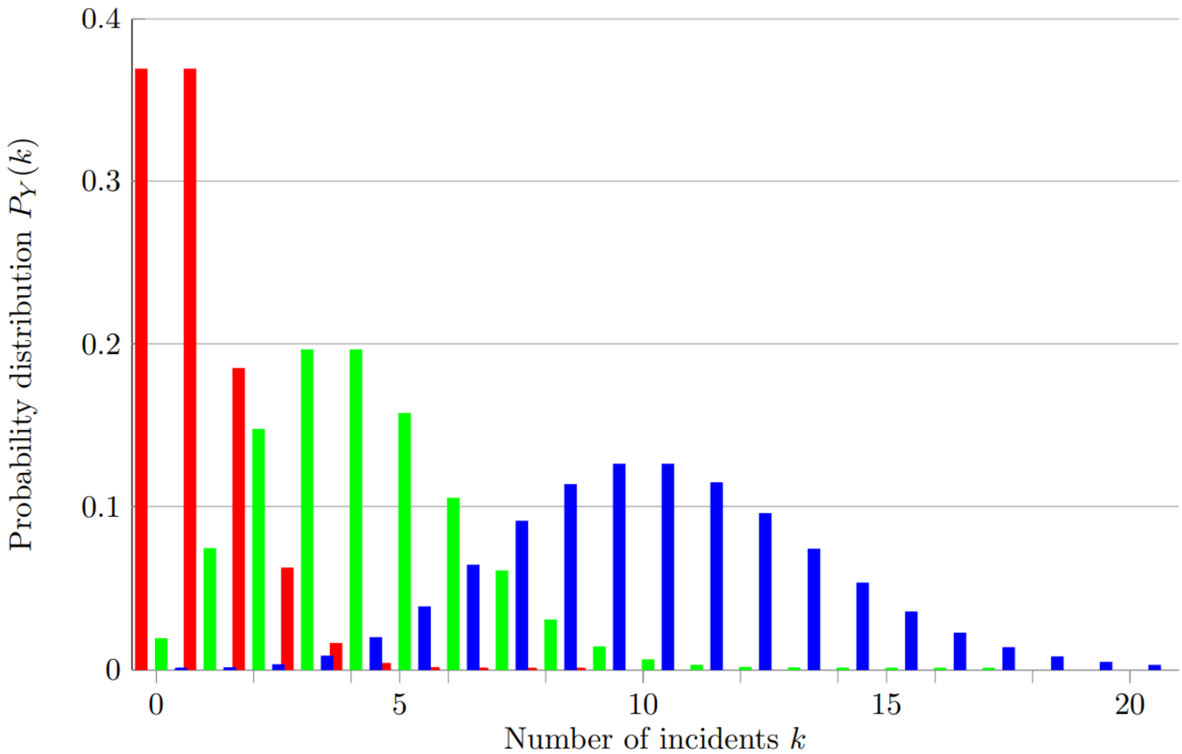
$$\operatorname{Var}(Y) = \frac{1-p}{p^2}$$

$$H(Y) = \frac{H_2(p)}{p}$$



The Poisson Distribution

The **Poisson distribution** is used to model the probability of the number of incidents in a time interval when incidents happen independently at a given rate of λ incidents per time interval. The incidents in this context are assumed to be of zero duration, or if they have a duration we are interested only in the start or the end of the incident, which have zero duration.



Poisson distributions for $\lambda = 1$ (red), $\lambda = 4$ (green) and $\lambda = 10$ (blue)

The Poisson distribution is the limit of a binomial distribution $B(n, \lambda/n)$ as n goes to infinity. Y is the random variable counting the number of incidents in the time interval of interest.

$P_Y(k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$

$E[Y] = \mathrm{Var}(Y) = \lambda$

Continuous Distributions

Fundamentals of Continuous Random Variables

For continuous random variables, the sample space itself must contain a continuum of possible outcomes in order for a random variable to be truly continuous. The probability distributions $P_X(x)$ would in general be zero everywhere. However, we can consider events corresponding to intervals and use the cumulative probability function.

$F_X(x) \geq 0$

$\lim_{x\to-\infty} F_X(x) = 0$

$\lim_{x\to\infty} F_X(x) = 1$

$F_X(x)$ increases with x.

The probability of falling within an interval $[a, b]$ can be expressed in function of the cumulative probability function.

$p(a \leq X \leq b) = p(X \leq b) - p(X \leq a) = F_X(b) - F_X(a)$

Joint cumulative probability functions are defined similarly to joint distributions for discrete random variables.

$F_{XY}(x, y) = p(X \leq x \cap Y \leq y)$

The independence of continuous random variables is defined as the independence of all events associated with the random variables.

$F_{XY}(x, y) = F_X(x)F_Y(y) \; for \; all \; (x, y) \; in \; \mathcal{X} \times \mathcal{Y}$

The definition of conditional cumulative probability functions follows from the definition of conditional probability.

$F_{Y|X}(y|x) = p(Y \leq y|X \leq x) = \frac{F_{XY}(x,y)}{F_X(x)}$

Bayes' theorem:

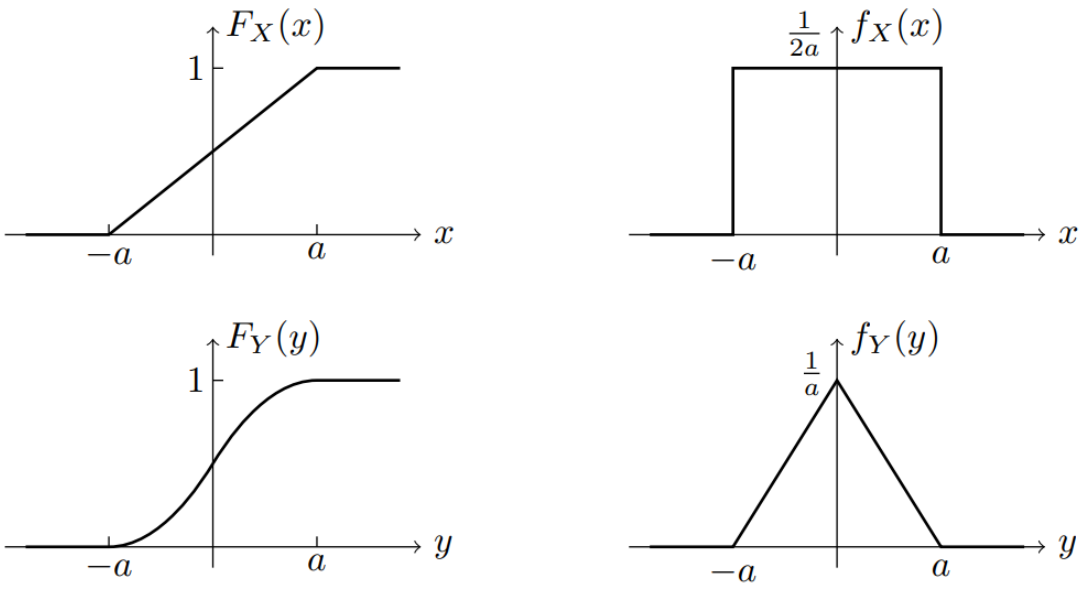
$F_{X|Y}(X|Y) = \frac{F_{Y|X}(y|x)F_X(x)}{F_Y(y)}$

However, there is no marginalisation for cumulative probability functions, because the events $X \leq x$ do intersect.

The Probability Density Function

The cumulative probability function for continuous random variables is often called the **cumulative density function** (CDF). It derivative is known as the **probability density function** (PDF) which is also called the **continuous probability distribution**.

$f_X(x) = \frac{dF_X(x)}{dx} = F'_X(x)$



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

$\int_{-\infty}^{\infty} f_X(x)dx = 1$

Since $F_X(x)$ is a non-decreasing function of x, its derivative, the probability density function, must always be positive or zero (non-negative).

Joint probability density functions are obtained from the cumulative density function through a multiple differentiation.

$f_{XY}(x, y) = \frac{d}{dx}\frac{d}{dy}F_{XY}(x, y)$

Independent random variables satisfy:

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

Marginalisation applies to probability density functions:

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

Conditional probability density functions can be defined as:

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

Bayes' theorem:

$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x')f_X(x')dx'}$

The probability density function can also be used to compute expectations:

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

The Exponential Distribution

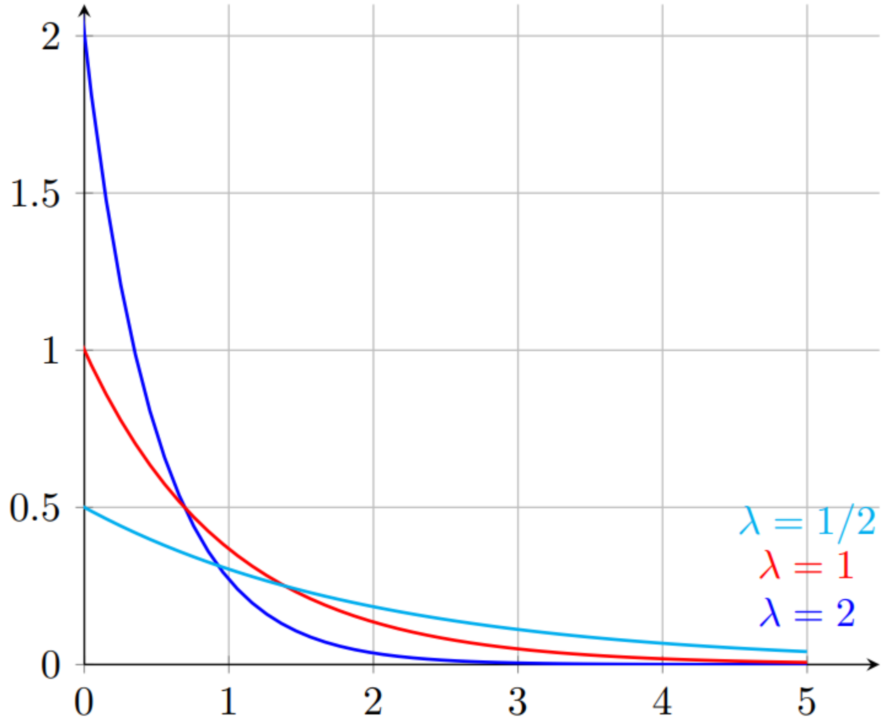
The **exponential distribution** can be derived from the Poisson distribution. Y_t is the Poisson distributed random variable giving the number of arrivals in a given time interval of length t. λ is the rate of packet arrivals per time unit, so that λt is the rate for the interval of length t.

$P_{Y_t}(k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$

The exponential distribution for the continuous random variable X is used to model the time intervals in a Poisson process with independent arrival times. X can only be larger than t if no arrivals occur in the interval.

$p(X > t) = P_{Y_t}(0) = e^{-\lambda t}$

$f_X(t) = \frac{d}{dt} F_X(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t}$



$E[X] = \int_0^\infty t f_X(t) dt = \frac{1}{\lambda}$

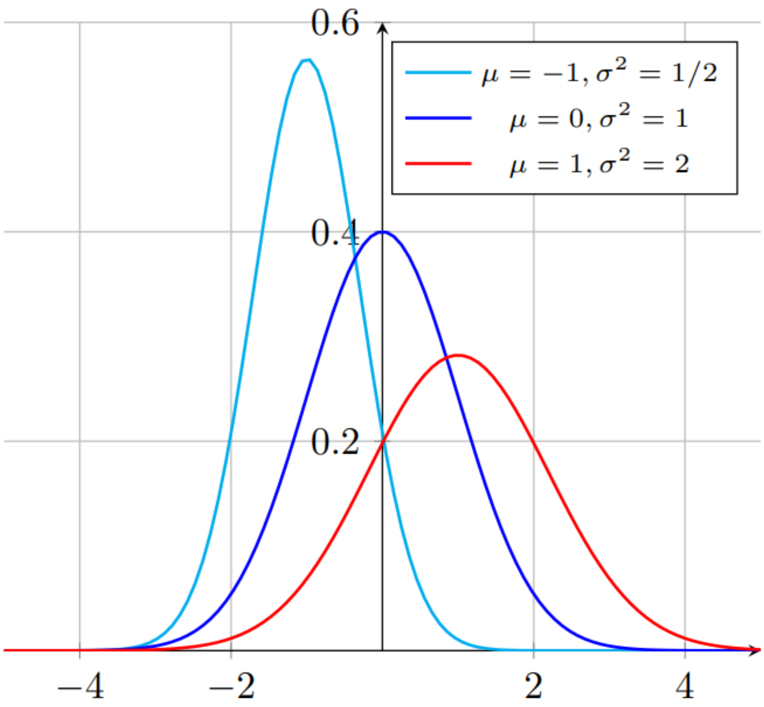
$E[X^2] = \int_0^\infty t^2 f_X(t) dt = \frac{2}{\lambda^2}$

$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{1}{\lambda^2}$

The Gaussian Distribution

If Y follows a Gaussian distribution with parameters μ and σ^2, $Y \sim N(\mu, \sigma^2)$.

$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$



$E[Y] = \mu$

$\mathrm{Var}[Y] = \sigma^2$

The cumulative probability function or CDF of a **standard Gaussian** random variable $X \sim N(0,1)$.

$F_X(x) = p(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x'^2}{2}} dx' = \Phi(x)$

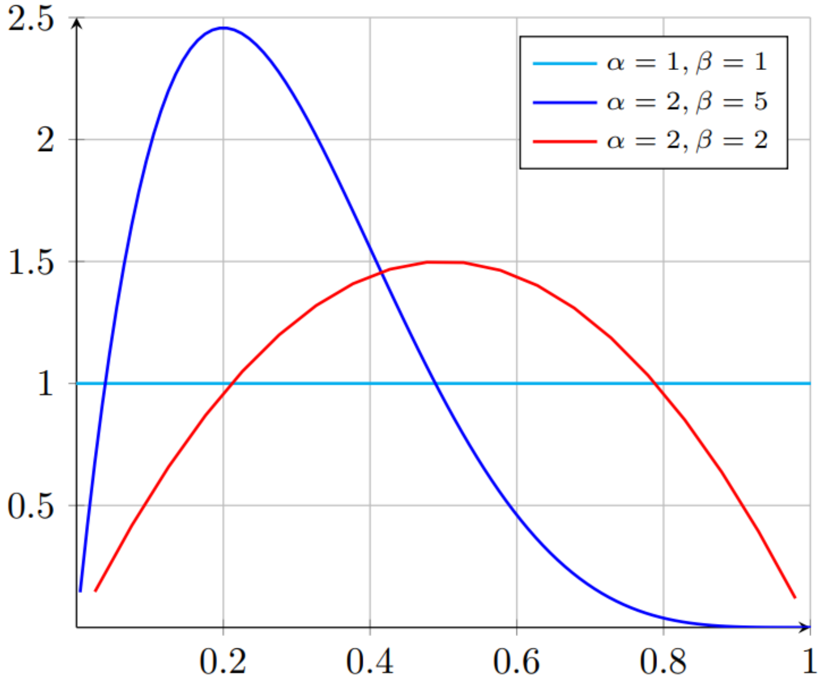
The Beta Distribution

The **Beta distribution** is a continuous probability density function whose range is limited to a finite interval. It is used to model parameters that have a finite range in various disciplines. It can be used to model the parameter p of a Bernoulli distributed random variable.

If π is a random variable that follows a Beta distribution with parameters α, β, denoted $\pi \sim \mathrm{Beta}(\alpha, \beta)$.

$f_\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

$\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy$



For integer arguments n:

$\Gamma(n) = (n-1)!$

For Beta distributions with integer parameters:

$f_\pi(p) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} p^{\alpha-1}(1-p)^{\beta-1} = (\alpha+\beta-1)\binom{\alpha+\beta-2}{\alpha-1} p^{\alpha-1}(1-p)^{\beta-1}$

$\mathrm{Beta}(1,1)$ is the **uniform** probability density function over the interval $[0,1]$.

$E[\pi] = \frac{\alpha}{\alpha+\beta}$

Characterising Distributions

Standard deviation is the square root of the variance.

Mode is the most probable value.

Median is the middle value.

Quartiles are the x values such that $F_X(x) = 1/4$, $F_X(x) = 1/2$, and $F_X(x) = 3/4$.

Interquartile range is the third quartile minus the first quartile.

Skewness is defined as $E\left[(X-\mu)^3\right]/\sigma^3$. If the skewness is positive, the distribution is skewed to the right. Informally the 'tail' of the distribution is longer to the right.

Manipulating and Combining Distributions

Sums of Random Variables

Let X and Y be two **independent** random variables and consider the sum $S = X + Y$.

$E[S] = E[X+Y] = E[X] + E[Y]$

$\mathrm{Var}(S) = \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

For discrete random variables, we assume a sum $S = X + Y$ of two random variables taking values in sets of numbers $\mathcal{X}$ and $\mathcal{Y}$.

$P_S(s) = \sum_{(x,y):x+y=s} P_{XY}(x,y) = \sum_{(x,y):x+y=s} P_X(x)P_Y(y)$

For plain integer addition, $P_S(s) = \sum_{x\in\mathcal{X}} P_X(x)P_Y(s-x)$.

For continuous random variables, we resort to infinitesimal calculus to compute the density of $S = X + Y$ given the densities of the independent continuous random variables X and Y.

$f_S(s) = \int_{-\infty}^{\infty} f_X(x)f_Y(s-x)dx$

Transforms of Distributions

The Probability Generating Function (PGF)

For a discrete random variable X taking values on the set X, the **probability generating function** (PGF) is defined as:

$g_X(z) = \sum_{x\in\mathcal{X}} P_X(x)z^x = E\left[z^X\right]$

The convolution property:

$P_Y(y) = (P_X \star P_X \star \cdots \star P_X)(y) \leftrightarrow g_Y(z) = (g_X(z))^n$

Convolutions can be evaluated efficiently as multiplications in the transform domain.

For a binomial distribution, the probability distribution of Y results from the convolution of the $\mathrm{Ber}(p)$ distribution n times with itself whose PGF is $g_X(z) = 1 - p + pz$.

$g_Y(z) = (g_X(z))^n = (1 - p + pz)^n$

The binomial and Poisson distributions are both **closed** under addition: consider a sum $S = X + Y$ of two random variables X and Y. If X and Y are binomial $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$ with the same parameter p, then S is binomial $S \sim B(n_1 + n_2, p)$.

$g_S(z) = g_X(z)g_Y(z) = (1 - p + pz)^{n_1}(1 - p + pz)^{n_2} = (1 - p + pz)^{n_1+n_2}$

If X and Y are Poisson $X \sim \mathrm{Po}(\lambda_1)$ and $Y \sim \mathrm{Po}(\lambda_2)$, then S is a Poisson distributed random variable $S \sim \mathrm{Po}(\lambda_1 + \lambda_2)$.

$g_S(z) = g_X(z)g_Y(z) = e^{\lambda_1(z-1)}e^{\lambda_2(z-1)} = e^{(\lambda_1+\lambda_2)(z-1)}$

The PGF can be used to compute moments of random variables.

$g_X^{'}(1) = \sum_{x\in\mathcal{X}} xP_X(x)z^{x-1}\big|_{z=1} = E[X]$

$g_X^{''}(1) = \sum_{x\in\mathcal{X}} x(x-1)P_X(x)z^{x-2}\big|_{z=1} = E\left[X^2\right] - E[X]$

$g_X^{(k)}(1) = E[X(X-1)(X-2)\cdots(X-k+1)]$

The Moment Generating Function (MGF)

For a continuous random variable X, the **moment generating function** (MGF) is defined as:

$g_X(s) = \int_{-\infty}^{\infty} f_X(x)e^{sx}dx = E\left[e^{sX}\right]$

The MGF can be seen as a two-sided generalisation of the Laplace transform.

The MGF of the standard Gaussian distribution can be derived:

$g_X(s) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{sx}dx = e^{s^2/2}$

The MGF of general Gaussian variables $Y \sim N(\mu, \sigma^2)$ can be determined from the standard Gaussian X where $Y = \sigma X + \mu$.

$g_Y(s) = e^{\mu s} g_X(\sigma s) = e^{\mu s + \sigma^2 s^2/2}$

The Gaussian density is **closed** under addition of random variables: consider two independent Gaussian random variables $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and their sum $Z = X + Y$, then Z is a Gaussian random variable with the sum of the means and the sum of the variances, $Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$g_Z(s) = e^{\mu_1 s + \frac{\sigma_1^2 s^2}{2}} e^{\mu_2 s + \frac{\sigma_2^2 s^2}{2}} = e^{(\mu_1+\mu_2)s + \frac{(\sigma_1^2+\sigma_2^2)s^2}{2}}$

The MGF can also be used to compute moments of random variables.

$g_X^{'}(0) = \int_{-\infty}^{\infty} x f_X(x) e^{sx} dx\big|_{s=0} = E[X]$

$g_X^{''}(0) = \int_{-\infty}^{\infty} x^2 f_X(x) e^{sx} dx\big|_{s=0} = E[X^2]$

$g_X^{(n)}(0) = E[X^n]$

The Central Limit Theorem

Let $X_1, X_2, \cdots$ be independent random variables with means $\mu_1, \mu_2, \cdots$ and variances $\sigma_1, \sigma_2, \cdots$. Assume that the random variables are all continuous with any probability density functions whose MGF exist. In particular, the densities can all be different. Then the random variable $Y_n = X_1 + X_2 + \cdots + X_n$ tends to a Gaussian random variable Y as n grows to infinity:

$Y \sim N(\mu_1 + \mu_2 + \cdots + \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$

Multivariate Gaussians

The random vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ is multivariate Gaussian $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$, if:

$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} |\mathbf{\Sigma}|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$

$\mathbf{\Sigma}$ is the $n \times n$ **covariance matrix** whose elements are:

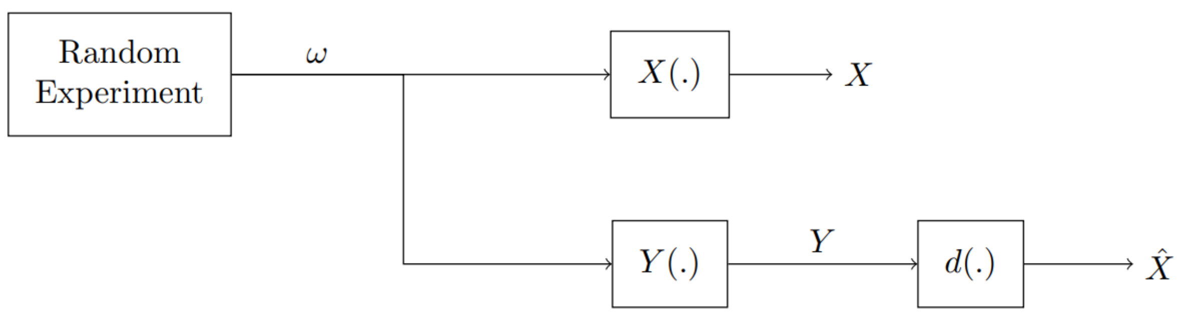
$\Sigma_{km} = E[(X_k - \mu_k)(X_m - \mu_m)] = E[X_k X_m] - \mu_k \mu_m$

$\mu_k = E[X_k]$ for all k and $\mu = (\mu_1, \mu_2, \cdots, \mu_n)$ is the **mean vector**.

Decision, Estimation and Hypothesis Testing

Decision and Estimation theory

In decision theory, X is a discrete random variable and the role of the decision block $d(.)$ is to decide the value of X based on the observation Y, which may be discrete or continuous. In estimation theory, X is a continuous random variable and $d(.)$ aims to provide an estimate of X based on the observation Y. In all cases, the conditional probability distribution $P_{Y|X}(.|.)$ or density $f_{Y|X}(.|.)$ is known to the **decider** or **estimator**.



The Maximum A-Posteriori (MAP) rule: for an observation $Y = y$, pick $\hat{X} = x$ to maximise $P_{X|Y}(x|y)$.

The Maximum Likelihood (ML) rule: for an observation $Y = y$, pick $\hat{X} = x$ to maximise $P_{Y|X}(y|x)$ (or $f_{Y|X}(y|x)$ for continuous observations.)

The ML rule is equivalent to the MAP rule when X is uniform, but is often also used in cases where the prior distribution P_X is unknown to the decider.

In estimation problems, X is continuous and we cannot reconstruct X exactly. We aim to find a $\hat{X}$ that approximates X as closely as possible given the observation Y. The closeness is commonly defined to minimise the **Mean Squared Error** (MSE):

$$E\left[(\hat{X} - X)^2 | Y = y\right] = \text{Var}\left[X | Y = y\right] + \left(\hat{X} - E\left[X | Y = y\right]\right)^2 \geq \text{Var}\left[X | Y = y\right]$$

The Minimum Mean Squared Error (MMSE) estimator: $\hat{X} = E\left[X | Y = y\right]$

Hypothesis Testing

Hypothesis testing is a branch of classical statistics that establishes rules for making certain statements about uncertain events, sometimes qualifying them with a soft "p-value".

Given an observed random variable Y, a **simple** hypothesis H is one for which the probabilities $p(Y = y|H)$ and $p(Y = y|\overline{H})$ are well defined, where $\overline{H}$ is the complement of H. H is often called the **null** hypothesis $H_0 = H$, and $\overline{H}$ the **alternative hypothesis** $H_1 = \overline{H}$.

The outcome of a hypothesis test is a statement concluding either H_1 is true (H_0 is false) or H_1 is false (H_0 is true), possibly with a numerical p-value indicating the strength of the statement. If X is an indicator random variable for our statement, it is useful to distinguish between the types of error that we can make in our statement:

X \ H_1	false	true
0	$\checkmark$	type II
1	type I	$\checkmark$

For **composite hypotheses**, it is not easy or impossible to express a probability distribution of the data.