Probability theory

Central subjects in probability theory include discrete and continuous random variables, probability distributions, and stochastic processes.

Mathematical abstractions of non-deterministic or uncertain processes and quantities.

Although it is not possible to perfectly predict random events, much can be said about their behaviour. Two major results in probability theory describing such behaviour are *the law of large numbers* and *the central limit theorem*.

Treatment

Most introductions to probability theory treat discrete probability distributions and continuous probability distributions separately.

Consider an experiment that can produce a number of outcomes. The set of all outcomes is called the sample space of the experiment.

Discrete probability distributions

Discrete probability theory deals with events that occur in countable sample spaces.

Let's start with a finite or countable set called the sample space, which is the set of all possible outcomes, denoted by Ω. It is then assumed that to each element $x \in \Omega$, an intrinsic "probability" value $f(x)$ is attached, with the following properties:

1. $f(x) \in [0,1]$ for all $x \in \Omega$,

2. $\sum_{x \in \Omega} f(x) = 1$.

That is, the probability function $f(x)$ lies between 0 and 1 for every value of x in the sample space Ω, and the sum of $f(x)$ over all possible values of x in the sample space Ω is equal to 1. An event is defined as any subset E of the sample space Ω. The probability of the event E is defined as

$$P(E) = \sum_{x \in E} f(x).$$

The function $f(x)$ mapping a point in the sample space to the "probability" value is called a probability mass function abbreviated as pmf. The modern definition does not try to answer how probability mass functions are obtained; instead, it builds a theory that assumes their existence.

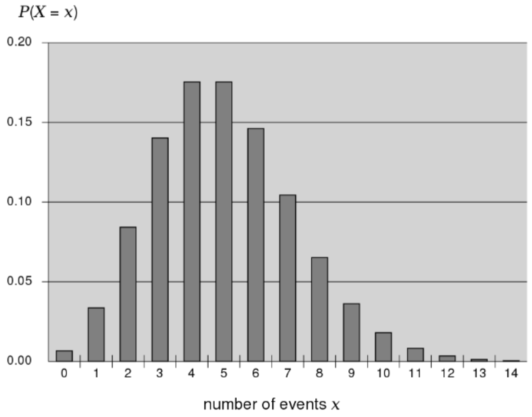

Figure. A discrete probability distribution.

Continuous probability distributions

Continuous probability theory deals with events that occur in a continuous sample space.

If the outcome space of a random variable X is the set of real numbers ($\mathbb{R}$) or a subset thereof, then a function called the cumulative distribution function (or CDF) F exists, defined by $F(x) = P(X \le x)$. That is, $F(x)$ returns the probability that X will be less than or equal to x. The CDF necessarily satisfies the following properties:

1. F is a monotonically non-decreasing, right-continuous function,

2. $\lim\limits_{x \to -\infty} F(x) = 0$,

3. $\lim\limits_{x \to \infty} F(x) = 1$.

If the derivative of F exists and integrating the derivative gives us the CDF back again, then the random variable X is said to have a probability density function or PDF, given by $f(x) = \dfrac{dF(x)}{dx}$.

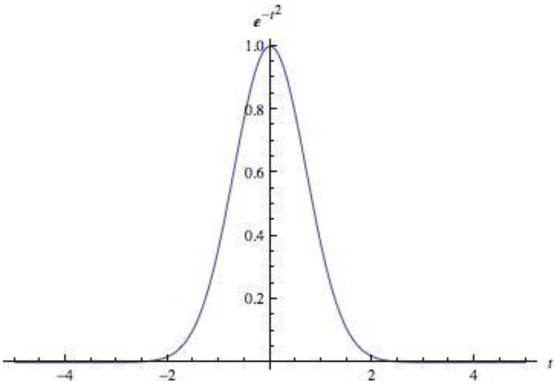

Figure. A continuous probability distribution (not normalized).

Note: Whereas the PDF exists only for continuous random variables, the CDF exists for all random variables (including discrete random variables) that take values in $\mathbb{R}$. These concepts can be generalized for multidimensional cases on $\mathbb{R}^n$ and other continuous sample spaces.

Law of large numbers

Common intuition suggests that if a fair coin is tossed many times, then roughly half of the times it will turn up heads, and the other half it will turn up tails. Furthermore, the more often the coin is tossed, the more likely it should be that the ratio of the number of heads to the number of tails will approach *unity*. Modern probability theory provides a formal version of this intuitive idea, known as *the law of large numbers*. This law is remarkable because it is not assumed in the foundations of probability theory, but instead emerges from those foundations as a theorem. Since it links theoretically derived probabilities to their actual frequency of occurrence in the real world, the law of large numbers is considered as a pillar in the history of statistical theory and has had widespread influence.

Consider a sequence of independent and identically distributed random variables X_k with $k=1,...,n$. Let's define the sample average $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ and assume that the common expectation is $E[X_k] = \mu$ and the variance $Var(X_k) = E[(X_k - \mu)^2]$ is finite. It can be shown that the expectation of $\bar{X}_n$ is still $E[\bar{X}_n] = \mu$ and the variance of $\bar{X}_n$ is given by

$$Var(\bar{X}_n) = E[(\bar{X}_n - \mu)^2] = \frac{1}{n} Var(X_k),$$

which approaches *zero* as $n \to \infty$. It follows that the standard deviation of $\bar{X}_n$, $\sqrt{Var(\bar{X}_n)}$, scales with n as $1/\sqrt{n} \to 0$.

Note 1: The derivation of $Var(\bar{X}_n) = \frac{1}{n} Var(X_k)$ is based on the statistical independence, which states that the random variables X_k are *independent*.

Note 2: The law of large numbers is essential to the validity of macroscopic thermodynamics for systems in the thermodynamic limit.

Normal distribution

The normal (or Gaussian) distribution is a very common continuous probability distribution. Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. The normal distribution is useful because of the central limit theorem, which is to be discussed below.

The PDF of the normal distribution is given by

$$f(x\,|\,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

in which μ is the mean or expectation, σ is the standard deviation, and σ^2 is the variance. The standard normal distribution is $f(x\,|\,\mu=0,\sigma^2=1)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ with zero expectation and unit variance. As to the notation, the normal distribution is often denoted by $N(\mu,\sigma^2)$. If a random variable X is normally distributed, then one may write $X\sim N(\mu,\sigma^2)$.

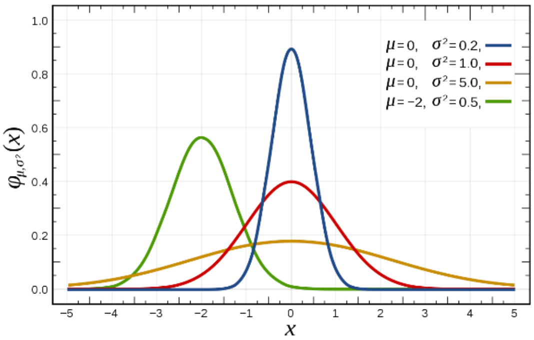

Figure. A few normal distributions parametrized by expectation μ and variance σ^2.

In probability theory, a distribution is said to be *stable* if a linear combination of two independent random variables with this distribution has the same distribution, up to location and scale parameters. The Gaussian distribution belongs to the family of stable distributions. Here is a simple and straightforward exercise.

Consider three random variables X, Y, and $Z=X+Y$, with X and Y being independent. Given $X\sim N(0,\sigma^2)$ and $Y\sim N(0,\sigma^2)$, show $Z\sim N(0,2\sigma^2)$.

Outline of the proof: We start from an integral expression for the CDF of $Z=X+Y$. The CDF, denoted by $F_Z(z)$, reads $F_Z(z)=\int_{-\infty}^{\infty}dx\int_{-\infty}^{z-x}dy\,f_X(x)\,f_Y(y)$. It follows that the PDF, denoted by $f_Z(z)$, is given by $f_Z(z)=\int_{-\infty}^{\infty}dx\,f_X(x)\,f_Y(z-x)$, which can be directly derived from $f_Z(z)=\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dy\,\delta(z-x-y)\,f_X(x)\,f_Y(y)$.

Central limit theorem

In probability theory, the central limit theorem establishes that, in many circumstances, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are *not* normally distributed.

Let $\{X_1,X_2,\cdots,X_n\}$ be a sequence of independent and identically distributed random variables drawn from a distribution of expectation μ and finite variance σ^2. Suppose we are interested in the sample average $Y=\frac{X_1+X_2+\cdots+X_n}{n}$ of these random variables. By the law of large numbers, the sample average converges almost surely to the expectation μ as $n \to \infty$. The classical central limit theorem describes the size and the distributional form of the stochastic fluctuations around the deterministic number μ during this convergence. More precisely, it states that for large enough *n*, the distribution of $Y = (X_1 + X_2 + \cdots + X_n)/n$ is close to the normal distribution with expectation μ and variance σ^2/n. The usefulness of the theorem is that the normal distribution is approached regardless of the shape of the distribution of the individual X_i.

Outline of the proof:

(1) Let $f(x_i)$ denote the PDF of X_i and Y denote the sample average $Y = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, with the expectation value $\int_{-\infty}^{\infty} x_i f(x_i)dx_i = \mu$ and the variance value $\int_{-\infty}^{\infty}(x_i-\mu)^2 f(x_i)dx_i = \int_{-\infty}^{\infty}(x_i^2-\mu^2)f(x_i)dx_i = \sigma^2$.

(2) The PDF of Y, denoted by $g(y)$, is given by $g(y) = \int \prod_{i=1}^{n} dx_i \delta\left(y - \dfrac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\prod_{i=1}^{n} f(x_i)\right)$ based on the statistical independence.

(3) The Fourier transform of $g(y)$, defined by $g_Y(k) = \int_{-\infty}^{\infty} dy e^{-iky} g(y)$, is given by

$$g_Y(k) = \int_{-\infty}^{\infty} dy e^{-iky} \int \prod_{i=1}^{n} dx_i \delta\left(y - \frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\prod_{i=1}^{n} f(x_i)\right) = \int \prod_{i=1}^{n} dx_i \exp\left(-i\frac{k}{n}\sum_{i=1}^{n} x_i\right)\left(\prod_{i=1}^{n} f(x_i)\right),$$

which can be written as

$$g_Y(k) = \left[\int dx_i \exp\left(-i\frac{k}{n}x_i\right) f(x_i)\right]^n.$$

(4) With the understanding that n is large enough, we have the Taylor expansion $\exp\left(-i\dfrac{k}{n}x_i\right) = 1 - i\dfrac{k}{n}x_i + \dfrac{1}{2}\left(-i\dfrac{k}{n}x_i\right)^2$ and

$$g_Y(k) = \left[\exp\left(-i\frac{k}{n}\mu\right)\exp\left(-\frac{k^2}{2n^2}\sigma^2\right)\right]^n = \exp(-ik\mu)\exp\left(-\frac{k^2}{2n}\sigma^2\right).$$

(5) Through the inverse Fourier transform $g(y) = \dfrac{1}{2\pi}\int_{-\infty}^{\infty} dk e^{iky} g_Y(k)$, we obtain $g(y)$ as

$$g(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk e^{iky} \exp(-ik\mu)\exp\left(-\frac{k^2}{2n}\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2/n}\right],$$

which is a normalized normal distribution with expectation μ and variance σ^2/n.

Maximum likelihood estimation

Maximum likelihood estimation is a method that estimates the parameters in a distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data become most probable.

Consider a random sample $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$ taken from an unknown distribution. In each probability distribution $f(y;\theta)$, there is a unique vector $\theta = [\theta_1, \theta_2, \cdots, \theta_k]^T$ of parameters that parametrize the probability distribution. As θ changes in value, different probability distributions are generated. The joint probability of the sample data, $f(y_1, y_2, \cdots, y_n;\theta)$, is given by $f(y_1, y_2, \cdots, y_n;\theta) = \prod_{i=1}^{n} f(y_i;\theta)$ for independent and identically distributed random variables. The likelihood function is defined as

$$L(\theta; y_1, y_2, \cdots, y_n) = f(y_1, y_2, \cdots, y_n;\theta) = \prod_{i=1}^{n} f(y_i;\theta),$$

which treats θ as a variable. The goal is then to find the value of θ that maximizes the likelihood function over the parameter space of θ.

For illustration purpose, we discuss a case of *continuous* distribution and *continuous* parameter space. The PDF of the random variable *X* is a normal distribution, i.e., $X \sim N(\mu, \sigma^2)$ in which μ and σ^2 are to be treated as variables in maximum likelihood estimation. For a sample given by $\{x_1, x_2, \cdots, x_n\}$, the likelihood function is

$$L(\mu, \sigma^2) = f(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right],$$

which is to be maximized with respect to μ and σ^2. From $\left.\frac{\partial L}{\partial \mu}\right|_{(\hat{\mu}, \hat{\sigma}^2)} = 0$, we obtain

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Furthermore, from $\left.\frac{\partial L}{\partial \sigma}\right|_{(\hat{\mu}, \hat{\sigma}^2)} = 0$, we obtain

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Formally, we say that the maximum likelihood estimator for $\theta = (\mu, \sigma^2)$ is given by $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$.

Bayesian inference

Bayesian inference is a method of *statistical inference* in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

Bayes' theorem is stated mathematically as

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

in which *A* and *B* are events and $P(B) \neq 0$. In this equation,

- $P(A|B)$ is a conditional probability: the probability of event *A* occurring given that *B* is true.
- $P(B|A)$ is also a conditional probability: the probability of event *B* occurring given that *A* is true.
- $P(A)$ and $P(B)$ are the probabilities of observing *A* and *B* independently of each other; known as the marginal probability.

Bayes' theorem may be derived from

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

with $P(A \cap B) = P(B \cap A)$. Here $P(A \cap B)$ and $P(B \cap A)$ are the joint probability of both *A* and *B* being true, and $P(A|B)$ and $P(B|A)$ are the conditional probabilities introduced above.

Here is an example for the theorem's application to drug testing.

Drug testing Example for Conditional Probability and Bayes Theorem

Suppose that a drug test for an illegal drug is such that it is 98% accurate in the case of a user of that drug (e.g. it produces a positive result with probability .98 in the case that the tested individual uses the drug) and 90% accurate in the case of a non-user of the drug (e.g. it is negative with probability .9 in the case the person does not use the drug). Suppose it is known that 10% of the entire population uses this drug.

You test someone and the test is positive. What is the probability that the tested individual uses this illegal drug?

What is the probability of a false positive with this test (e.g. the probability of obtaining a positive drug test given that the person tested is a non-user)?

What is the probability of obtaining a false negative for this test (e.g. the probability that the test is negative, but the individual tested is a user)?

Let:
+ = the event that the drug test is positive for an individual
- = the event that the drug test is negative for an individual
A = the event that the person tested does use the drug that is being tested for

We want to find: $P[A \mid +]$, $P[+ \mid \bar{A}]$ (false positive), and $P[- \mid A]$ (false negative)

We know that $P[A] = .1$, $P[+ \mid A] = .98$, and $P[- \mid \bar{A}] = .9$ and from this we know that $P[+ \mid \bar{A}] = .9$ which allows us to find using Bayes Theorem:

$$P[A \mid +] = \frac{P[+ \mid A]\, P[A]}{P[+]}$$

so

$$P[A \mid +] = \frac{P[+ \mid A]\, P[A]}{P[+ \mid A]\, P[A] + P[+ \mid \bar{A}]\, P[\bar{A}]}$$

or

$$P[A \mid +] = .98\ \frac{(.1)}{.98\,(.1) + .1(.9)} = .52$$

Also, the probability of a false positive is

$$P[+ \mid \bar{A}] = 1 - P[- \mid \bar{A}] = .1$$

and the probability of a false negative is

$$P[- \mid A] = 1 - P[+ \mid A] = .02$$

A short review of delta function is needed here.

Review
Purpose

Fourier transform

$$f(x) = \frac{1}{2\pi}\int \hat{f}(k)\, e^{ikx}\, dk$$

$$\hat{f}(k) = \int f(x)\, e^{-ikx}\, dx$$

$$f(x) = \frac{1}{2\pi}\int \underset{\underset{\int f(x')\, e^{-ikx'}\, dx'}{\uparrow}}{\hat{f}(k)}\, e^{ikx}\, dk$$

$$= \frac{1}{2\pi}\int f(x')\, e^{ik(x-x')}\, dk\, dx'$$

Using $\frac{1}{2\pi}\int e^{ik(x-x')}\, dk = \delta(x-x')$

we have $f(x) = \int f(x')\, \delta(x-x')\, dx'$

$$= f(x)$$

$f(x) \longleftrightarrow \hat{f}(k)$ via e^{ikx}

$f(t) \longleftrightarrow \hat{f}(\omega)$ via $e^{i\omega t}$

Fourier transform

Fourier transform

$$\int f^2(x)\,dx$$

$$= \int \frac{1}{2\pi}\hat{f}(k)e^{ikx}\,\frac{1}{2\pi}\hat{f}(q)e^{iqx}\,dk\,dq\,dx$$

$$= \frac{1}{(2\pi)^2}\int \hat{f}(k)\,\hat{f}(q)\,e^{i(k+q)x}\,dx\,dk\,dq$$

$$= \frac{1}{2\pi}\int \hat{f}(k)\hat{f}(q)\,\delta(k+q)\,dk\,dq$$

$$= \frac{1}{2\pi}\int \hat{f}(k)\,\hat{f}(-k)\,dk = \frac{1}{2\pi}\int \hat{f}(k)\hat{f}^{*}(k)\,dk$$

Energy of signal.