Analytical models of discrete random phenomena

5.1 DISCRETE RANDOM VARIBLE: In probability and statistics, a **random variable**, **aleatory variable** or **stochastic variable** is a variable whose value is subject to variations due to chance (i.e. randomness, in a mathematical sense). A random variable can take on a set of possible different values (similarly to other mathematical variables), each with an associated probability (if discrete) or a probability density function (if continuous), in contrast to other mathematical variables.

A random variable's possible values might represent the possible outcomes of a yet-to-be-performed experiment, or the possible outcomes of a past experiment whose already-existing value is uncertain (for example, as a result of incomplete information or imprecise measurements). They may also conceptually represent either the results of an "objectively" random process (such as rolling a die) or the "subjective" randomness that results from incomplete knowledge of a quantity. The meaning of the probabilities assigned to the potential values of a random variable is not part of probability theory itself but is instead related to philosophical arguments over the interpretation of probability. The mathematics works the same regardless of the particular interpretation in use.

The mathematical function describing the possible values of a random variable and their associated probabilities is known as a probability distribution. *Random variables* can be ***discrete***, that is, taking any of a specified finite or countable list of values, endowed with a probability mass function, characteristic of a probability distribution; or ***continuous***, taking any numerical value in an interval or collection of intervals, via a probability density function that is characteristic of a probability distribution; or a mixture of both types. The realizations of a random variable, that is, the results of randomly choosing values according to the variable's probability distribution function, are called random variates.

The formal mathematical treatment of random variables is a topic in probability theory. In that context, a random variable is understood as a function defined on a sample space whose outputs are numerical values.

Definition

Random variable is usually understood to mean a real-valued random variable; this discussion assumes real values. A random variable is a real-valued function defined on a set of possible outcomes, the sample space Ω. That is, the random

variable is a function that maps from its domain, the sample space Ω, to its range, the real numbers or a subset of the real numbers. It is typically some kind of a property or measurement on the random outcome (for example, if the random outcome is a randomly chosen person, the random variable might be the person's height, or number of children).

The fine print: the admissible functions for defining random variables are limited to those for which a probability distribution exists, derivable from a probability measure that turns the sample space into a probability space. That is, for the mapping to be an admissible random variable, it must be theoretically possible to compute the probability that the value of the random variable is less than any particular real number. Equivalently, the preimage of any range of values of the random variable must be a subset of Ω that has a defined probability; that is, there exists a subset of Ω, an event, the probability of which is the same probability as the random variable being in the range of real numbers that that event maps to. Furthermore, the notion of a "range of values" here must be generalizable to the non-pathological subset of reals known as Borel sets.

Random variables are typically distinguished as discrete versus continuous ones. Mixtures of both types also exist.

Discrete random variables can take on either a finite or at most accountably infinite set of discrete values (for example, the integers). Their probability distribution is given by a probability mass function which directly maps each value of the random variable to a probability; for each possible value of the random variable, the probability is equal to the probability of the event containing all possible outcomes in Ω that map to that value.

Continuous random variables, on the other hand, take on values that vary continuously within one or more real intervals, and have a cumulative distribution function (CDF) that is absolutely continuous. As a result, the random variable has an uncountably infinite number of possible values, all of which have probability 0, though ranges of such values can have nonzero probability. The resulting probability distribution of the random variable can be described by a probability density. (Some sources refer to this class as "absolutely continuous random variables", and allow a wider class of "continuous random variables", including those with singular distributions, but note that these are typically not encountered in practical situations. Random variables with discontinuities in their CDFs can be treated as mixtures of discrete and continuous random variables.

Examples

For example, in an experiment a person may be chosen at random, and one random variable may be the person's height. Mathematically, the random variable is interpreted as a function which maps the person to the person's height. Associated with the random variable is a probability distribution that allows the computation of the probability that the height is in any non-pathological subset of possible values, such as probability that the height is between 180 and 190 cm, or the probability that the height is either less than 150 or more than 200 cm.

Another random variable may be the person's number of children; this is a discrete random variable with non-negative integer values. It allows the computation of probabilities for individual integer values – the probability mass function (PMF) – or for sets of values, including infinite sets. For example, the event of interest may be "an even number of children". For both finite and infinite event sets, their probabilities can be found by adding up the PMFs of the elements; that is, the probability of an even number of children is the infinite sum PMF(0) + PMF(2) + PMF(4) + ...

In examples such as these, the sample space (the set of all possible persons) is often suppressed, since it is mathematically hard to describe, and the possible values of the random variables are then treated as a sample space. But when two random variables are measured on the same sample space of outcomes, such as the height and number of children being computed on the same random persons, it is easier to track their relationship if it is acknowledged that both height and number of children come from the same random person, for example so that questions of whether such random variables are correlated or not can be posed.

Probability density

The probability distribution for continuous random variables can be defined using a probability density function (PDF or p.d.f), which indicates the "density" of probability in a small neighborhood around a given value. The probability that a random variable is in a particular range can then be computed from the integral of the probability density function over that range. The PDF is the derivative of the CDF.

Mixtures

Some random variables are neither discrete nor continuous, but a mixture of both types. Their CDF is not absolutely continuous, and a PDF does not exist. For example, a typical "sparse" continuous random variable may be exactly 0 with probability 0.9, and continuously distributed otherwise, so its CDF has a big jump discontinuity at 0. The PDF therefore does not exist as an ordinary function in this case, though such situations are easily handled by using a distribution instead of a function to represent a PDF, or by using other representations of measure.

5.2 function of probability and distribution: In probability and statistics, a probability distribution assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference. Examples are found in experiments whose sample space is non-numerical, where the distribution would be a categorical distribution; experiments whose sample space is encoded by discrete random variables, where the distribution can be specified by a probability mass function; and experiments with sample spaces encoded by continuous random variables, where the distribution can be specified by a probability density function. More complex experiments, such as those involving stochastic processes defined in continuous time, may demand the use of more general probability measures.

In applied probability, a probability distribution can be specified in a number of different ways, often chosen for mathematical convenience:

- by supplying a valid probability mass function or probability density function
- by supplying a valid cumulative distribution function or survival function
- by supplying a valid hazard function
- by supplying a valid characteristic function
- by supplying a rule for constructing a new random variable from other random variables whose joint probability distribution is known.

A probability distribution can either be univariate or multivariate. A univariate distribution gives the probabilities of a single random variable taking on various alternative values; a multivariate distribution (a joint probability distribution) gives the probabilities of a random vector—a set of two or more random variables—taking on various combinations of values. Important and commonly encountered univariate probability distributions include the binomial distribution, the

hypergeometric distribution, and the normal distribution. The multivariate normal distribution is a commonly encountered multivariate distribution.

Introduction

The probability mass function (pmf) p(S) specifies the probability distribution for the sum S of counts from two dice. For example, the figure shows that p(11) = 1/18. The pmf allows the computation of probabilities of events such as P(S > 9) = 1/12 + 1/18 + 1/36 = 1/6, and all other probabilities in the distribution.

To define probability distributions for the simplest cases, one needs to distinguish between discrete and continuous random variables. In the discrete case, one can easily assign a probability to each possible value: for example, when throwing a fair die, each of the six values 1 to 6 has the probability 1/6. In contrast, when a random variable takes values from a continuum, probabilities can be nonzero only if they refer to intervals: in quality control one might demand that the probability of a "500 g" package containing between 490 g and 510 g should be no less than 98%.

The probability density function (pdf) of the normal distribution, also called Gaussian or "bell curve", the most important continuous random distribution. As notated on the figure, the probabilities of intervals of values correspond to the area under the curve.

If the random variable is real-valued (or more generally, if a total order is defined for its possible values), the cumulative distribution function (CDF) gives the probability that the random variable is no larger than a given value; in the real-valued case, the CDF is the integral of the probability density function (pdf) provided that this function exists.

Terminology

As probability theory is used in quite diverse applications, terminology is not uniform and sometimes confusing. The following terms are used for non-cumulative probability distribution functions:

Probability mass, Probability mass function, p.m.f.: for discrete random variables.
Categorical distribution: for discrete random variables with a finite set of values.
Probability density, Probability density function, p.d.f: most often reserved for continuous random variables.

The following terms are somewhat ambiguous as they can refer to non-cumulative or cumulative distributions, depending on authors' preferences:

- Probability distribution function: continuous or discrete, non-cumulative or cumulative.
- Probability function: even more ambiguous, can mean any of the above or other things.

Finally,

Probability distribution: sometimes the same as probability distribution function, but usually refers to the more complete assignment of probabilities to all measurable subsets of outcomes, not just to specific outcomes or ranges of outcomes.

Basic terms

- Mode: for a discrete random variable, the value with highest probability (the location at which the probability mass function has its peak); for a continuous random variable, the location at which the probability density function has its peak.

- Support: the smallest closed set whose complement has probability zero.

- Head: the range of values where the pmf or pdf is relatively high.

- Tail: the complement of the head within the support; the large set of values where the pmf or pdf is relatively low.

- Expected value or mean: the weighted average of the possible values, using their probabilities as their weights; or the continuous analog thereof.

- Median: the value such that the set of values less than the median has a probability of one-half.

- Variance: the second moment of the pmf or pdf about the mean; an important measure of the dispersion of the distribution.

- Standard deviation: the square root of the variance, and hence another measure of dispersion.

- Symmetry: a property of some distributions in which the portion of the distribution to the left of a specific value is a mirror image of the portion to its right.
- Skewness: a measure of the extent to which a pmf or pdf "leans" to one side of its mean

A continuous probability distribution is a probability distribution that has a probability density function. Mathematicians also call such a distribution absolutely continuous, since its cumulative distribution function is absolutely continuous with respect to the Lebesgue measure λ. If the distribution of X is continuous, then X is called a continuous random variable. There are many examples of continuous probability distributions: normal, uniform, chi-squared, and others.

Intuitively, a continuous random variable is the one which can take a continuous range of values—as opposed to a discrete distribution, where the set of possible values for the random variable is at most countable. While for a discrete distribution an event with probability zero is impossible (e.g., rolling 3 1⁄2 on a standard die is impossible, and has probability zero), this is not so in the case of a continuous random variable. For example, if one measures the width of an oak leaf, the result of 3½ cm is possible, however it has probability zero because there are uncountably many other potential values even between 3 cm and 4 cm. Each of these individual outcomes has probability zero, yet the probability that the outcome will fall into the interval (3 cm, 4 cm) is nonzero. This apparent paradox is resolved by the fact that the probability that X attains some value within an infinite set, such as an interval, cannot be found by naively adding the probabilities for individual values. Formally, each value has an infinitesimally small probability, which statistically is equivalent to zero.

In probability theory, the expected value of a random variable is intuitively the long-run average value of repetitions of the experiment it represents. For example, the expected value of a die roll is 3.5 because, roughly speaking, the average of an extremely large number of die rolls is practically always nearly equal to 3.5. Less roughly, the law of large numbers guarantees that the arithmetic mean of the values almost surely converges to the expected value as the number of repetitions goes to infinity. The expected value is also known as the expectation, mathematical expectation, EV, mean, or first moment.

More practically, the expected value of a discrete random variable is the probability-weighted average of all possible values. In other words, each possible

value the random variable can assume is multiplied by its probability of occurring, and the resulting products are summed to produce the expected value. The same works for continuous random variables, except the sum is replaced by an integral and the probabilities by probability densities. The formal definition subsumes both of these and also works for distributions which are neither discrete nor continuous: the expected value of a random variable is the integral of the random variable with respect to its probability measure.

The expected value does not exist for random variables having some distributions with large "tails", such as the Cauchy distribution. For random variables such as these, the long-tails of the distribution prevent the sum/integral from converging.

The expected value is a key aspect of how one characterizes a probability distribution; it is one type of location parameter. By contrast, the variance is a measure of dispersion of the possible values of the random variable around the expected value. The variance itself is defined in terms of two expectations: it is the expected value of the squared deviation of the variable's value from the variable's expected value.

The expected value plays important roles in a variety of contexts. In regression analysis, one desires a formula in terms of observed data that will give a "good" estimate of the parameter giving the effect of some explanatory variable upon a dependent variable. The formula will give different estimates using different samples of data, so the estimate it gives is itself a random variable. A formula is typically considered good in this context if it is an unbiased estimator—that is, if the expected value of the estimate (the average value it would give over an arbitrarily large number of separate samples) can be shown to equal the true value of the desired parameter.

In decision theory, and in particular in choice under uncertainty, an agent is described as making an optimal choice in the context of incomplete information. For risk neutral agents, the choice involves using the expected values of uncertain quantities, while for risk averse agents it involves maximizing the expected value of some objective function such as a von Neumann-Morgenstern utility function.