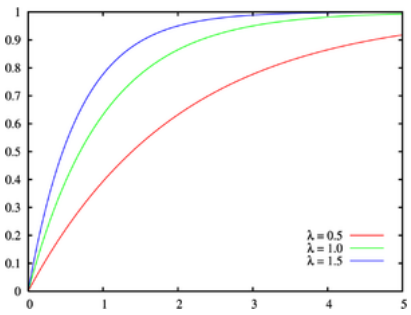
Cumulative distribution function

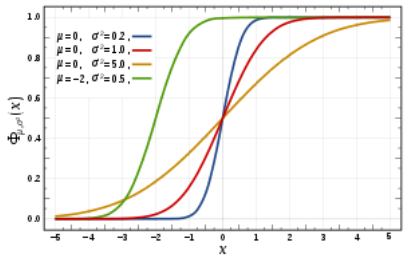
In probability theory and statistics, the **cumulative distribution function** (**CDF**) of a real-valued random variable X, or just **distribution function** of X, evaluated at x, is the probability that X will take a value less than or equal to x.[1]

Every probability distribution supported on the real numbers, discrete or "mixed" as well as continuous, is uniquely identified by an *upwards continuous*[2] *monotonic increasing* cumulative distribution function $F : \mathbb{R} \to [0, 1]$ satisfying $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

In the case of a scalar continuous distribution, it gives the area under the probability density function from minus infinity to x. Cumulative distribution functions are also used to specify the distribution of multivariate random variables.



Cumulative distribution function for the exponential distribution



Cumulative distribution function for the normal distribution

Contents

Definition

The cumulative distribution function of a real-valued random variable X is the function given by[3]:p. 77

$$F_X(x) = \mathrm{P}(X \leq x) \qquad \text{(Eq.1)}$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x.

The probability that X lies in the semi-closed interval $(a, b]$, where $a < b$, is therefore[3]:p. 84

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a) \qquad \text{(Eq.2)}$$

In the definition above, the "less than or equal to" sign, "≤", is a convention, not a universally used one (e.g. Hungarian literature uses "<"), but the distinction is important for discrete distributions. The proper use of tables of the binomial and Poisson distributions depends upon this convention. Moreover, important formulas like Paul Lévy's inversion formula for the characteristic function also rely on the "less than or equal" formulation.

If treating several random variables $X, Y, \ldots$ etc. the corresponding letters are used as subscripts while, if treating only one, the subscript is usually omitted. It is conventional to use a capital F for a cumulative distribution function, in contrast to the lower-case f used for probability density functions and probability mass functions. This applies when discussing general distributions: some specific distributions have their own conventional notation, for example the normal distribution uses Φ and ϕ instead of F and f, respectively.

The probability density function of a continuous random variable can be determined from the cumulative distribution function by differentiating[4] using the Fundamental Theorem of Calculus; i.e. given $F(x)$,

$$f(x) = \frac{dF(x)}{dx}$$

as long as the derivative exists.

The CDF of a continuous random variable X can be expressed as the integral of its probability density function f_X as follows:[3]:p. 86

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

In the case of a random variable X which has distribution having a discrete component at a value b,

$$\mathrm{P}(X = b) = F_X(b) - \lim_{x \to b^-} F_X(x).$$

If F_X is continuous at b, this equals zero and there is no discrete component at b.

Properties

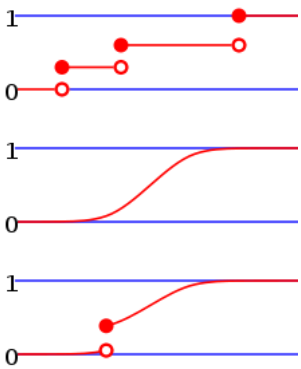
Every cumulative distribution function F_X is non-decreasing[3]:p. 78 and right-continuous,[3]:p. 79 which makes it a càdlàg function. Furthermore,

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to +\infty} F_X(x) = 1.$$

Every function with these four properties is a CDF, i.e., for every such function, a random variable can be defined such that the function is the cumulative distribution function of that random variable.

If X is a purely discrete random variable, then it attains values $x_1, x_2, \ldots$ with probability $p_i = p(x_i)$, and the CDF of X will be discontinuous at the points x_i:

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{x_i \leq x} \mathrm{P}(X = x_i) = \sum_{x_i \leq x} p(x_i).$$



From top to bottom, the cumulative distribution function of a discrete probability distribution, continuous probability distribution, and a distribution which has both a continuous part and a discrete part.

If the CDF F_X of a real valued random variable X is continuous, then X is a continuous random variable; if furthermore F_X is absolutely continuous, then there exists a Lebesgue-integrable function $f_X(x)$ such that

$$F_X(b) - F_X(a) = \mathrm{P}(a < X \leq b) = \int_a^b f_X(x)\,dx$$

for all real numbers a and b. The function f_X is equal to the derivative of F_X almost everywhere, and it is called the probability density function of the distribution of X.

Examples

As an example, suppose X is uniformly distributed on the unit interval $[0, 1]$.

Then the CDF of X is given by

$$F_X(x) = \begin{cases} 0 & : x < 0 \\ x & : 0 \leq x \leq 1 \\ 1 & : x > 1 \end{cases}$$

Suppose instead that X takes only the discrete values 0 and 1, with equal probability.

Then the CDF of X is given by

$$F_X(x) = \begin{cases} 0 & : x < 0 \\ 1/2 & : 0 \leq x < 1 \\ 1 & : x \geq 1 \end{cases}$$

Suppose X is exponential distributed. Then the CDF of X is given by

$$F_X(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Here $\lambda > 0$ is the parameter of the distribution, often called the rate parameter.

Suppose X is normal distributed. Then the CDF of X is given by

$$F(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt.$$

Here the parameter μ is the mean or expectation of the distribution; and σ is its standard deviation.

Suppose X is binomial distributed. Then the CDF of X is given by

$$F(k; n, p) = \Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

Here p is the probability of success and the function denotes the discrete probability distribution of the number of successes in a sequence of n independent experiments, and $\lfloor k \rfloor$ is the "floor" under k, i.e. the greatest integer less than or equal to k.

Derived functions

Complementary cumulative distribution function (tail distribution)

Sometimes, it is useful to study the opposite question and ask how often the random variable is *above* a particular level. This is called the **complementary cumulative distribution function** (**ccdf**) or simply the **tail distribution** or **exceedance**, and is defined as

$$\bar{F}_X(x) = \mathrm{P}(X > x) = 1 - F_X(x).$$

This has applications in statistical hypothesis testing, for example, because the one-sided p-value is the probability of observing a test statistic *at least* as extreme as the one observed. Thus, provided that the test statistic, *T*, has a continuous distribution, the one-sided p-value is simply given by the ccdf: for an observed value t of the test statistic

$$p = \mathrm{P}(T \geq t) = \mathrm{P}(T > t) = 1 - F_T(t).$$

In survival analysis, $\bar{F}_X(x)$ is called the **survival function** and denoted $S(x)$, while the term *reliability function* is common in engineering.

Z-table:

One of the most popular application of cumulative distribution function is standard normal table, also called the **unit normal table** or **Z table**,[5] is the value of cumulative distribution function of the normal distribution. It is very useful to use Z-table not only for probabilities below a value which is the original application of cumulative distribution function, but also above and/or between values on standard normal distribution, and it was further extended to any normal distribution.

Properties

- For a non-negative continuous random variable having an expectation, Markov's inequality states that[6]

  $$\bar{F}_X(x) \leq \frac{\mathrm{E}(X)}{x}.$$

- As $x \to \infty, \bar{F}_X(x) \to 0$, and in fact $\bar{F}_X(x) = o(1/x)$ provided that $\mathrm{E}(X)$ is finite.
  Proof:
  Assuming X has a density function f_X, for any $c > 0$

  $$\mathrm{E}(X) = \int_0^\infty x f_X(x)\, dx \geq \int_0^c x f_X(x)\, dx + c \int_c^\infty f_X(x)\, dx$$

  Then, on recognizing

  $$\bar{F}_X(c) = \int_c^\infty f_X(x)\, dx$$

  and rearranging terms,

  $$0 \leq c\bar{F}_X(c) \leq \mathrm{E}(X) - \int_0^c x f_X(x)\, dx \to 0 \text{ as } c \to \infty$$

  as claimed.
- For a random variable having an expectation,

  $$\mathrm{E}(X) = \int_0^\infty \bar{F}_X(x)\, dx - \int_{-\infty}^0 F_X(x)\, dx$$

  and for a non-negative random variable the second term is 0.
  If the random variable can only take non-negative integer values, this is equivalent to

$$\mathrm{E}(X) = \sum_{n=0}^{\infty} \bar{F}_X(n).$$

Folded cumulative distribution

While the plot of a cumulative distribution often has an S-like shape, an alternative illustration is the **folded cumulative distribution** or **mountain plot**, which folds the top half of the graph over,[7][8] thus using two scales, one for the upslope and another for the downslope. This form of illustration emphasises the median, dispersion (specifically, the mean absolute deviation from the median[9]) and skewness of the distribution or of the empirical results.

Example of the folded cumulative distribution for a normal distribution function with an expected value of 0 and a standard deviation of 1.

Inverse distribution function (quantile function)

If the CDF F is strictly increasing and continuous then $F^{-1}(p), p \in [0,1],$ is the unique real number x such that $F(x) = p$. In such a case, this defines the **inverse distribution function** or quantile function.

Some distributions do not have a unique inverse (for example in the case where $f_X(x) = 0$ for all $a < x < b$, causing F_X to be constant). This problem can be solved by defining, for $p \in [0,1]$, the **generalized inverse distribution function**:

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

- Example 1: The median is $F^{-1}(0.5)$.
- Example 2: Put $\tau = F^{-1}(0.95)$. Then we call τ the 95th percentile.

Some useful properties of the inverse cdf (which are also preserved in the definition of the generalized inverse distribution function) are:

1. F^{-1} is nondecreasing
2. $F^{-1}(F(x)) \leq x$
3. $F(F^{-1}(p)) \geq p$
4. $F^{-1}(p) \leq x$ if and only if $p \leq F(x)$
5. If Y has a $U[0,1]$ distribution then $F^{-1}(Y)$ is distributed as F. This is used in random number generation using the inverse transform sampling-method.
6. If $\{X_\alpha\}$ is a collection of independent F-distributed random variables defined on the same sample space, then there exist random variables Y_α such that Y_α is distributed as $U[0,1]$ and $F^{-1}(Y_\alpha) = X_\alpha$ with probability 1 for all α.

The inverse of the cdf can be used to translate results obtained for the uniform distribution to other distributions.

Empirical distribution function

The empirical distribution function is an estimate of the cumulative distribution function that generated the points in the sample. It converges with probability 1 to that underlying distribution. A number of results exist to quantify the rate of convergence of the empirical distribution function to the underlying cumulative distribution function.

Multivariate case

Definition for two random variables

When dealing simultaneously with more than one random variable the **joint cumulative distribution function** can also be defined. For example, for a pair of random variables X, Y, the joint CDF F_{XY} is given by[3]:p. 89

$$F_{X,Y}(x,y) = \mathrm{P}(X \leq x, Y \leq y) \qquad \text{(Eq.3)}$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x **and** that Y takes on a value less than or equal to y.

Example of joint cumulative distribution function:

For two continuous variables X and Y:

$$\Pr(a < X < b \text{ and } c < Y < d) = \int_a^b \int_c^d f(x,y)\, dy\, dx;$$

For two discrete random variables, it is beneficial to generate a table of probabilities and address the cumulative probability for each potential range of X and Y, and here is the example:[10]

given the joint probability mass function in tabular form, determine the joint cumulative distribution function.

	$Y = 2$	$Y = 4$	$Y = 6$	$Y = 8$
$X = 1$	0	0.1	0	0.1
$X = 3$	0	0	0.2	0
$X = 5$	0.3	0	0	0.15
$X = 7$	0	0	0.15	0

Solution: using the given table of probabilities for each potential range of X and Y, the joint cumulative distribution function may be constructed in tabular form:

	$Y < 2$	$2 \le Y < 4$	$4 \le Y < 6$	$6 \le Y < 8$	$Y \ge 8$
$X < 1$	0	0	0	0	0
$1 \le X < 3$	0	0	0.1	0.1	0.2
$3 \le X < 5$	0	0	0.1	0.3	0.4
$5 \le X < 7$	0	0.3	0.4	0.6	0.85
$X \ge 7$	0	0.3	0.4	0.75	1

Definition for more than two random variables

For N random variables $X_1, \ldots, X_N$, the joint CDF $F_{X_1,\ldots,X_N}$ is given by

$$F_{X_1,\ldots,X_N}(x_1, \ldots, x_N) = \mathrm{P}(X_1 \le x_1, \ldots, X_N \le x_N) \qquad \text{(Eq.4)}$$

Interpreting the N random variables as a random vector $\mathbf{X} = (X_1, \ldots, X_N)^T$ yields a shorter notation:

$$F_{\mathbf{X}}(\mathbf{x}) = \mathrm{P}(X_1 \le x_1, \ldots, X_N \le x_N)$$

Properties

Every multivariate CDF is:

1. Monotonically non-decreasing for each of its variables,
2. Right-continuous in each of its variables,
3. $0 \le F_{X_1 \ldots X_n}(x_1, \ldots, x_n) \le 1,$
4. $\lim_{x_1,\ldots,x_n \to +\infty} F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = 1 \text{ and } \lim_{x_i \to -\infty} F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = 0, \text{for all } i.$

Any function satisfying the above four properties is not a multivariate CDF, unlike in the single dimension case. For example, let $F(x, y) = 0$ for $x < 0$ or $x + y < 1$ or $y < 0$ and let $F(x, y) = 1$ otherwise. It is easy to see that the above conditions are met, and yet F is not a CDF since if it was, then $\mathrm{P}\left(\frac{1}{3} < X \le 1, \frac{1}{3} < Y \le 1\right) = -1$ as explained below.

The probability that a point belongs to a hyperrectangle is analogous to the 1-dimensional case:[11]

$$F_{X_1,X_2}(a, c) + F_{X_1,X_2}(b, d) - F_{X_1,X_2}(a, d) - F_{X_1,X_2}(b, c) = \mathrm{P}(a < X_1 \le b, c < X_2 \le d) = \int \ldots$$

Complex case

Complex random variable

The generalization of the cumulative distribution function from real to complex random variables is not obvious because expressions of the form $P(Z \le 1 + 2i)$ make no sense. However expressions of the form $P(\Re(Z) \le 1, \Im(Z) \le 3)$ make sense. Therefore, we define the cumulative distribution of a complex random variables via the joint distribution of their real and imaginary parts:

$$F_Z(z) = F_{\Re(Z),\Im(Z)}(\Re(z), \Im(z)) = P(\Re(Z) \le \Re(z), \Im(Z) \le \Im(z)).$$

Complex random vector

Generalization of **Eq.4** yields

$$F_{\mathbf{Z}}(\mathbf{z}) = F_{\Re(Z_1),\Im(Z_1),\ldots,\Re(Z_n),\Im(Z_n)}(\Re(z_1),\Im(z_1),\ldots,\Re(z_n),\Im(z_n)) = \mathrm{P}(\Re(Z_1) \leq \Re(z_1), \Im(Z_1) \leq \Im(z_1), \ldots, \Re(Z_n) \leq \Re(z_n), \Im(Z_n) \leq \Im(z_n)$$

as definition for the CDS of a complex random vector $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$.

Use in statistical analysis

The concept of the cumulative distribution function makes an explicit appearance in statistical analysis in two (similar) ways. Cumulative frequency analysis is the analysis of the frequency of occurrence of values of a phenomenon less than a reference value. The empirical distribution function is a formal direct estimate of the cumulative distribution function for which simple statistical properties can be derived and which can form the basis of various statistical hypothesis tests. Such tests can assess whether there is evidence against a sample of data having arisen from a given distribution, or evidence against two samples of data having arisen from the same (unknown) population distribution.

Kolmogorov–Smirnov and Kuiper's tests

The Kolmogorov–Smirnov test is based on cumulative distribution functions and can be used to test to see whether two empirical distributions are different or whether an empirical distribution is different from an ideal distribution. The closely related Kuiper's test is useful if the domain of the distribution is cyclic as in day of the week. For instance Kuiper's test might be used to see if the number of tornadoes varies during the year or if sales of a product vary by day of the week or day of the month.

See also

- Descriptive statistics
- Distribution fitting
- Ogive (statistics)

References

1. Deisenroth, Marc Peter; Faisal, A. Aldo; Ong, Cheng Soon (2020). *Mathematics for Machine Learning* (https://github.com/mml-book/mml-book.github.io). Cambridge University Press. p. 181. ISBN 9781108455145.
2. Hüseyin Çakallı (2015). "Upward and Downward Statistical Continuities" (https://www.jstor.org/stable/24898386). *Filomat*. **29** (10): 2265–2273. doi:10.2298/FIL1510265C (https://doi.org/10.2298%2FFIL1510265C). JSTOR 24898386 (https://www.jstor.org/stable/24898386). S2CID 58907979 (https://api.semanticscholar.org/CorpusID:58907979).
3. Park, Kun Il (2018). *Fundamentals of Probability and Stochastic Processes with Applications to Communications*. Springer. ISBN 978-3-319-68074-3.
4. Montgomery, Douglas C.; Runger, George C. (2003). *Applied Statistics and Probability for Engineers* (http://www.um.edu.ar/math/montgomery.pdf) (PDF). John Wiley & Sons, Inc. p. 104. ISBN 0-471-20454-4.
5. "Z Table" (https://www.ztable.net/). *Z Table*. Retrieved 2019-12-11.
6. Zwillinger, Daniel; Kokoska, Stephen (2010). *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press. p. 49. ISBN 978-1-58488-059-2.
7. Gentle, J.E. (2009). *Computational Statistics* (https://books.google.com/books?id=m4r-KVxpLsAC&pg=PA348). Springer. ISBN 978-0-387-98145-1. Retrieved 2010-08-06.
8. Monti, K. L. (1995). "Folded Empirical Distribution Function Curves (Mountain Plots)". *The American Statistician*. **49** (4): 342–345. doi:10.2307/2684570 (https://doi.org/10.2307%2F2684570). JSTOR 2684570 (https://www.jstor.org/stable/2684570).
9. Xue, J. H.; Titterington, D. M. (2011). "The p-folded cumulative distribution function and the mean absolute deviation from the p-quantile" (https://hal.archives-ouvertes.fr/hal-00753950/file/PEER_stage2_10.1016%252Fj.spl.2011.03.014.pdf) (PDF). *Statistics & Probability Letters*. **81** (8): 1179–1182. doi:10.1016/j.spl.2011.03.014 (https://doi.org/10.1016%2Fj.spl.2011.03.014).
10. "Joint Cumulative Distribution Function (CDF)" (https://math.info/Probability/Joint_CDF/). *math.info*. Retrieved 2019-12-11.
11. "Archived copy" (https://web.archive.org/web/20160222051842/http://www.math.wustl.edu/~hgan/Prob2014/slides.259-327.pdf) (PDF). *www.math.wustl.edu*. Archived from the original (http://www.math.wustl.edu/~hgan/Prob2014/slides.259-327.pdf) (PDF) on 22 February 2016. Retrieved 13 January 2022.

External links

- Media related to Cumulative distribution functions at Wikimedia Commons