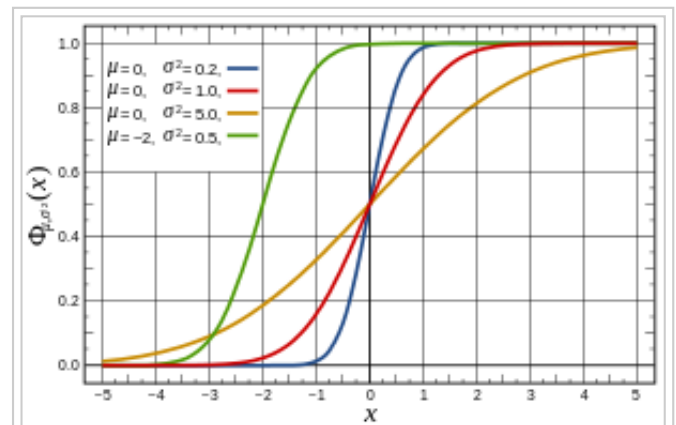


Cumulative distribution function

From Wikipedia, the free encyclopedia

In probability theory and statistics, the **cumulative distribution function (CDF)**, or just **distribution function**, evaluated at 'x', is the probability that a real-valued random variable X will take a value less than or equal to x . In other words, $\text{CDF}(x) = \text{Pr}(X \leq x)$, where Pr denotes probability.

In the case of a 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 normal distribution

Contents

- 1 Definition
- 2 Properties
- 3 Examples
- 4 Derived functions
 - 4.1 Complementary cumulative distribution function (tail distribution)
 - 4.2 Folded cumulative distribution
 - 4.3 Inverse distribution function (quantile function)
- 5 Multivariate case
- 6 Use in statistical analysis
 - 6.1 Kolmogorov–Smirnov and Kuiper's tests
- 7 See also
- 8 References
- 9 External links

Definition

The cumulative distribution function of a real-valued random variable X is the function given by

$$F_X(x) = P(X \leq x),$$

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

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

In the definition above, the "less than or equal to" sign, " \leq ", is a convention, not a universally used one (e.g. Hungarian literature uses " $<$ "), but 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, \dots 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.

The CDF of a continuous random variable X can be expressed as the integral of its probability density function f_X as follows:

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

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

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

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

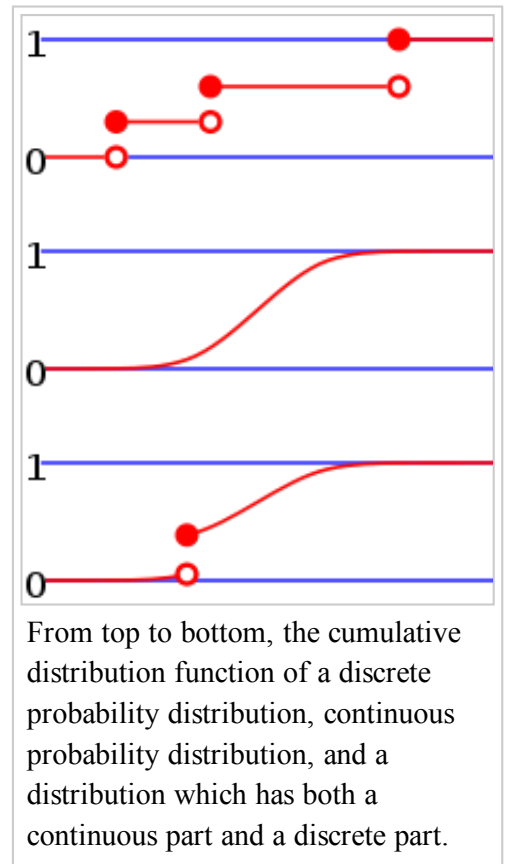
Properties

Every cumulative distribution function F is non-decreasing and right-continuous, which makes it a càdlàg function. Furthermore,

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

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, \dots with probability $p_i = P(x_i)$, and the CDF of X will be discontinuous at the points x_i and constant in between:



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

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

$$F(b) - F(a) = P(a < X \leq b) = \int_a^b f(x) dx$$

for all real numbers a and b . The function f is equal to the derivative of F 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

$$F(x) = \begin{cases} 0 & : x < 0 \\ x & : 0 \leq x < 1 \\ 1 & : x \geq 1. \end{cases}$$

Suppose instead that X takes only the discrete values 0 and 1, with equal probability. Then the CDF of X is given by

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

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

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

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

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

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

Properties

- For a non-negative continuous random variable having an expectation, Markov's inequality states that^[1]

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

- As $x \rightarrow \infty$, $\bar{F}(x) \rightarrow 0$, and in fact $\bar{F}(x) = o(1/x)$ provided that $\mathbb{E}(X)$ is finite.

Proof: Assuming X has a density function f , for any $c > 0$

$$\mathbb{E}(X) = \int_0^\infty x f(x) dx \geq \int_0^c x f(x) dx + c \int_c^\infty f(x) dx$$

Then, on recognizing $\bar{F}(c) = \int_c^\infty f(x) dx$ and rearranging terms,

$$0 \leq c\bar{F}(c) \leq \mathbb{E}(X) - \int_0^c x f(x) dx \rightarrow 0 \text{ as } c \rightarrow \infty$$

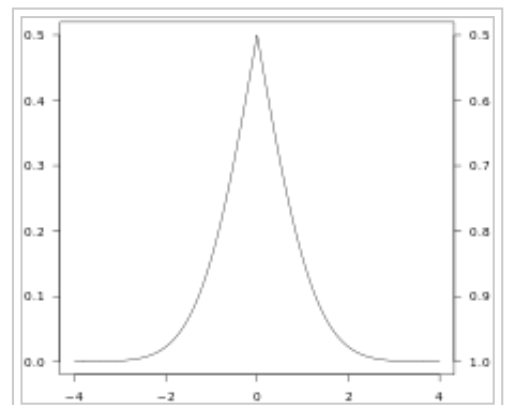
as claimed.

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,^{[2][3]} thus using two scales, one for the upslope and another for the downslope. This form of illustration emphasises the median and dispersion (the mean absolute deviation from the median^[4]) of the distribution or of the empirical results.

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.



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

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 $y \leq F(x)$
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.

Multivariate case

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 is given by

$$F(x, y) = P(X \leq x, Y \leq y),$$

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 .

Every multivariate CDF is:

1. Monotonically non-decreasing for each of its variables
2. Right-continuous for each of its variables.
3. $0 \leq F(x_1, \dots, x_n) \leq 1$
4. $\lim_{x_1, \dots, x_n \rightarrow +\infty} F(x_1, \dots, x_n) = 1$ and $\lim_{x_i \rightarrow -\infty} F(x_1, \dots, x_n) = 0$, for all i

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

References

1. Zwillinger, Daniel; Kokoska, Stephen (2010). *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press. p. 49. ISBN 978-1-58488-059-2.
2. Gentle, J.E. (2009). *Computational Statistics*. Springer. ISBN 978-0-387-98145-1. Retrieved 2010-08-06.
3. Monti, K.L. (1995). "Folded Empirical Distribution Function Curves (Mountain Plots)". *The American Statistician* **49**: 342–345. doi:10.2307/2684570. JSTOR 2684570.
4. Xue, J. H.; Titterton, D. M. (2011). "The p-folded cumulative distribution function and the mean absolute deviation from the p-quantile". *Statistics & Probability Letters* **81** (8): 1179–1182. doi:10.1016/j.spl.2011.03.014.<

External links

-  Media related to Cumulative distribution functions at Wikimedia Commons

Retrieved from "https://en.wikipedia.org/w/index.php?"

title=Cumulative_distribution_function&oldid=705175314"

Categories: Theory of probability distributions

- This page was last modified on 15 February 2016, at 23:20.
- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.