

Probability Functions

I. The Basic Idea

We use probability models for Random Variables to determine which events are likely to occur by chance and which are not. This determination is fundamental to the statistical inference we shall shortly encounter. There are two kinds of probability models used in statistics we will cover in this course, the finite discrete and the continuous. Probabilities are presented for each one in two different ways, just different enough to be confusing. Throughout this discussion, X and Y stand for Random Variables (abbreviated RV) and x , and y are real numbers.

II. Finite Models

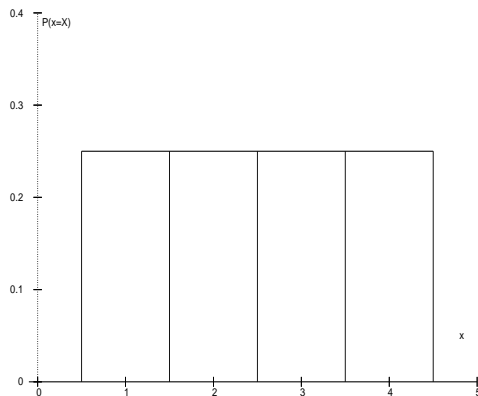
A finite model has a sample space which contains a finite number of numerical outcomes. The probabilities can be given in a list. Such a list is a function, often called the probability mass function, reflecting the fact that the probability is assigned in lumps or masses, which must sum to one. The usual abbreviation is pmf, and the usual symbol is

$$f_X(x) = P(X = x_i),$$

and though we will not use these, you might run into them in other work. Such a probability model is the one for the toss of a tetrahedral die, for which we have

i	x_i	P_i
1	1	$\frac{1}{4}$
2	2	$\frac{1}{4}$
3	3	$\frac{1}{4}$
4	4	$\frac{1}{4}$

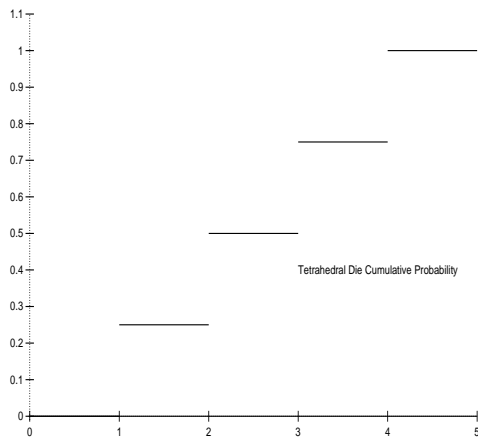
We also can provide a graphical representation of such a probability model



The other method of presenting probabilities is the Cumulative Density or Distribution function, abbreviated cdf, which represents the probability that the RV X is less than or equal to the real number x . In symbols, this is

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

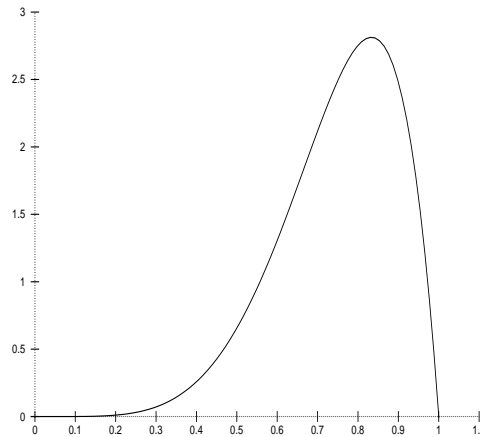
This also may be graphed and is a staircase function which jumps every at every x which is in the sample space.



III. Continuous Models

A continuous RV is characterized by the fact that it takes values in intervals, and so the probability assigned to a single value must be zero; only intervals can have positive probability assigned to them. Probability is given by a probability density function (symbol $f_Y(y)$) which must start at zero and

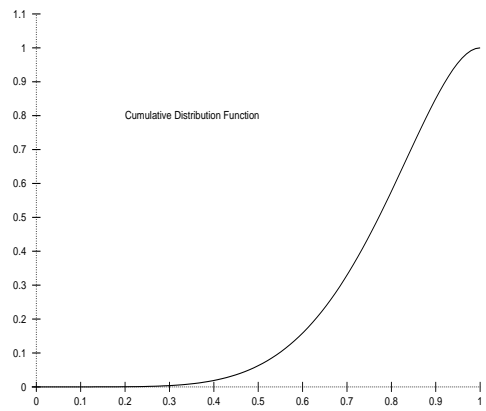
return to zero as one progresses along the x axis. The area under the curve is one. The height of the curve has no probability meaning, the probability that the RV Y lies between the real numbers, a and b , in symbols $a \leq Y \leq b$ is given by the area above the interval $[a, b]$ and below the curve. Examples are the uniform (watchman) and the triangular model given in class, but there is no requirement other than that the area under the curve be one. The usual symbol for this is $f_Y(y)$, the same as is used for the probability mass function in the discrete case, but the meaning is very different. A graph of such a function is shown below, you will have to take my word for it that the area underneath is one, although for those of you with a bit of calculus, it is easy to demonstrate that it is, the equation plotted is $y = 42x^5(1 - x)$.



The other way of presenting probability information for a continuous RV is identical to the second method for discrete RV's, the cumulative. The same definition and symbology is used

$$F_Y(y) = P(Y \leq y).$$

The only difference is that in the continuous case, probability accrues continuously instead of in lumps. The cdf associated with the above RV, Y , is graphed below, its equation is $y = 7x^6 - 6x^7$.



The height of the cumulative represents the area under the curve of the pdf from the left end of the distribution, where probabilities begin to accumulate.