Random Variables and Probability Distributions

Animal Health Risk Analysis
Canadian Food Inspection Agency
3851 Fallowfield Road
Ottawa, Ontario, Canada K2H 8P9

Analyse des risques de la santé des animaux
Agence canadienne d'inspection des aliments
3851, chemin Fallowfield
Ottawa (Ontario) Canada K2H 8P9



Canadian Food Inspection Agency (CFIA)

Agence canadienne d'inspection des aliments (ACIA)

Random Variables

A variable X is a random variable if the value that it assumes, corresponding to the outcome of an experiment, is a chance or random event. A random variable is a function that associates a real number with each element in the sample space. If a sample space contains a finite number of possibilities or an unending sequence with as many elements as there are whole numbers, it is called a discrete sample space. For example, if a coin is tossed twice, the sample space S = {HH, HT, TH, TT}. Let X represent the number of heads which can come up. With each simple event one can associate a number for X as shown below:

Simple Event	X	P(E)
HH	2	1/4
HT	1	1/4
TH	1	1/4
TT	0	1/4

If a sample space contains an infinite number of possibilities equal to the number of points on a line segment, it is called a contiuous sample space. When a random variable can take on values on a continuous scale, it is called a continuous random variable. Often the possible values of a continuous random variable are precisely the same values that are contained in the continuous sample space, e.g., weights, temperatures, distances, time.

Random Variables

Definition: A random variable is a variable whose value is determined by the outcomes of a random experiment. It is usually denoted by a capital letter such as X or Y.

Example 1 - The following are random variables:

(i) the weekly number of pieces of litmus paper used in a specific chemistry lab
(ii) the number of plants who survive spraying with a strong herbicide
(iii) the life length of an electron tube
(iv) the height of males in a biostatistics course
(v) the number of children in a family
(vi) the percentage level of glucose in one's blood
(vii) the no. of different solutions tested by a consumer until he finds one that is satisfactory for his slide preparation.

Definition: If the number of possible values of a random variable X is finite or countably infinite (i.e., as many as there are positive integers), then X is called a discrete random variable.

Example 2 - (a) Consider the experiment of tossing a coin twice. Then S = {HH, HT, TH, TT}. If X = the number of heads obtained, then X is a discrete random variable with a finite no. of possible values 0, 1 and 2.
(b) Consider the experiment of tossing a coin until a head appears. Then S = {H, TH, TTH, TTTH, ...}. If X = the No. of tosses required to get a first head, then Y is a discrete random variable with a countably infinite number of possible values 1, 2, 3, ...

Random Variables

Example 3 - In example 1, the random variables (i), (ii), and (v) are discrete with a finite number of possible values while random variable (vii) is discrete with a countably infinite no. of possible outcomes.

Definition: If the possible values of a random variable X form an interval or a collection of intervals, then X is known as a continuous random variable.

Example 4 - In example 1, the random variables (iii), (iv) and (vi) are continuous.

Mathematical Expectation of a Random Variable

We can speak of the mean, variance and standard deviation of a random variable or of a probability distribution in the same way as we do of a data set. Unlike the latter which are sample statistics, these will refer to populations since a random variable may take on any of all its possible values and as such will exist on a long run of the random experiment.
NOTE: We define the following, using the same notation already introduced, for a discrete random variable. The same definitions hold for a continuous random variable where the summation over all possible values of X is merely replaced by the integral over the entire domain of X.

Definition: Let X be a discrete random variable with possible values $x_1, x_2, ..., x_n$ and corresponding probability function f(x) = P(X=x). Then the mathematical expectation or the expected value of X is defined as where N is replaced by ∞ if X is countably infinite. Hence, E(X) is the long run average or process average and is also called the mean of the random variable X or the population mean so that E(X) = μ. It represents the centre of gravity for our possible values of X.

Properties of Expectation:
If a and b are real constants while X and Y are random variables, we have the following properties for the expectation operator.
(i) E(a) = a
(ii) E(aX) = aE(X)
(iii) E(aX+b) = AE(X) + b
(iv) E(X-μ) = 0
(v) E(X±Y) = E(X)±E(Y)
Note: E(1/X) ≠ 1/E(X)

Variance of a Random Variable

The variance of a random variable measures the spread or dispersion of the possible values of the random variable from its expected value μ .
Definition: Let X be a discrete random variable with possible values $x_1, x_2, ..., x_n$ and corresponding probability function f(x) = P(X=x). If μ = E(X) is the mean of X, then the variance of X is defined as

$$V(X) = E\left[(X-\mu)^2\right] = \sum_{i=1}^{N}(x_i-\mu)^2 f(x_i)$$

where N becomes ∞ if X is countably infinite.

Hence, V(X) is the average squared deviation fromE(X) and is also called the population variance so that V(X) = σ^2. The respective standard deviation of X is then simply σ = sd(X) = √V(X) which represents the average deviation of a value of the random variable from its mean μ or E(X).

Now V(X)

$$= E\left[(X-\mu)^2\right]$$
$$= E\left[(X^2-2\mu X+\mu^2)\right]$$
$$= E(X^2)-E(2\mu X)+E(\mu^2)$$
$$= E(X^2)-2\mu E(X)+\mu^2$$
$$= E(X^2)-2\mu^2+\mu^2 \qquad \because E(X)=\mu$$
$$= E(X^2)-\mu^2$$

Variance of a Random Variable

Computational Formula:

$$\sigma^2 = V(X) = E(X^2) - \mu^2 = E(X^2) - \left[E(X)\right]^2$$

Properties of Variance: Let a and b represent real constants and X and Y be random variables.

(i) V(a) = 0
(ii) V(aX) = a^2V(X)
(iii) V(aX+b) =a^2V(X)
(iv) V(X±Y) = V(X) + V(Y) if X and Y are independent
= V(X) + V(Y) ± 2Cov(X, Y) otherwise
where Cov(X, Y) = E{ [X-E(X)] [Y-E(Y)] } = E(XY) - E(X)E(Y)

Probability Distribution for Discrete Random Variables

Definition: The probability distribution for a discrete random variable is a formula, table, or graph that provides P(X=x), the probability associated with each of the values of x. The set of ordered pairs (x, f(x)) is a probability function, probability mass fucntion, or probability distribution of the discrete random variable X if, for each possible outcome x,

f(x) ≥ 0

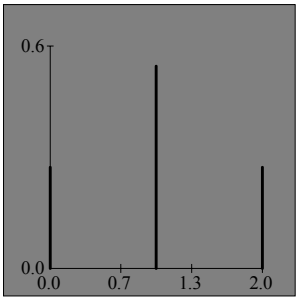
∑ f(x) = 1

P(X=x) = f(x)

Employing the example of the two coin toss the probability function corresponding to the random variable X is tabulated below:

x	f(x)
0	1/4
1	1/2
2	1/4

P(X = 0) = P(TT) = 1/4
P(X = 1) = P(HT∪TH) = P(HT) + P(TH) = 1/4 + 1/4 = ½
P(X = 2) = P(HH) = 1/4



Probability graph

Probability Distribution for Discrete Random Variables

Note: We denote by x_i the distinct possible values that a random variable X can assume and by $f(x_i) = P(X=x_i)$ the probability that X takes on the value x_i

Definition: Let X be a discrete random variable with possible values $x_1, x_2, \ldots$ and corresponding probabilities $f(x_i) = P(X=x_i)$ for i =1, 2, ... Then the function f(x) is called the distribution of X or probability function of X if the following conditions are satisfied:

(i) $f(x_i)$ for all x_i

(ii) $\sum_{i=1}^{\infty} f(x_i) = 1.$

(Here ∞ is replaced by n if X is finite.)

Definition: The probability that X is less than or equal to x in value, $P(x \leq x)$ is called the cumulative function or the distribution function of a random variable X and is denoted by F(x), that is, $F(x) = P(X \leq x)$.

Example: For the double coin tossing experiment where X represents the number of heads obtained, we have

x	0	1	2
f(x) = P(X=x)	¼	½	¼
F(x) = P(X≤ x)	¼	¾	1

Probability Distribution for Discrete Random Variables

Example: Consider the probability function of Y where y_i refers to *i* tosses being required for a head to first appear in the tossing of a fair coin. It may be written as

$Y = y_i$	1	2	3	...	n ...
$f(y_i) = P(Y=y_i)$	½	$(½)^2$	$(½)^3$	...	$(½)^n$ ...

or expressed algebraically as $f(y) = (½)^y$ where y = 1, 2, 3, ...

Since the sum of the infinite geometric series with common ratio r, $a + ar + ar^2 + \ldots$, is given as a/(1-r), then by definition $f(y) = (½)^y$ is a probability function with a = r = ½. As well,

$$F(3) = P(\leq 3) = f(1) + f(2) + f(3) = \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} = \tfrac{7}{8}$$

$$P(Y \geq 4) = 1 - P(Y \leq 3) = 1 - F(3) = 1 - \tfrac{7}{8} = \tfrac{1}{8}$$

$$P(2 \leq Y \leq 4) = f(2) + f(3) = \tfrac{1}{4} + \tfrac{1}{8} = \tfrac{3}{8} = F(3) - F(1)$$

Probability Distribution for Discrete Random Variables

PROBABILITY MASS FUNCTION

The probability distribution of a discrete random variable is represented by its probability mass function. It is a function whose domain contains the set of discrete values that the random variable can assume, with the probabilities of the random variable assuming the values in the domain as its range. The probability mass function of the discrete random variable Y is denoted by pY(y) and is defined as

pY(y) = P{Y=y} is an element of the domain of pY(y)

$0 \le pY(y) \le 1$ for all values of y that Y can assume

pY(y) = 0 for all other values of y

$$\sum_{all-y} p_Y(y) = 1$$

A probability mass function is often depicted graphically by a probability histogram. The potential values of the random variable are plotted on the x-axis, while their associated probabilities are plotted on the y-axis. Rectangles of equal width are centered on each discrete value, and their heights are equal to the probabilities that the random variable can assume those values.

Probability Distribution for Discrete Random Variables

PROBABILITY MASS FUNCTION

Example: A random variable X assumes a value equal to the sum of the rolls of two dice. Since there are 6 possible values for each die, the fundamental principle of counting asserts that there are 6 x 6 = 36 possible outcomes of the dice roll. With unbiased dice, each outcome will have a 1/36 change of occurring.

The following table shows the relationship between the outcome of the dice roll and the value of the random variable X:

Value of X	Dice Roll (x, y) x - roll of first die y - roll of second die	Number of Outcomes
2	(1, 1)	1
3	(1, 2), (2, 1)	2
4	(1, 3), (2, 2), (3, 1)	3
5	(1, 4), (2, 3), (3, 2), (4, 1)	4
6	(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)	5
7	(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)	6
8	(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)	5
9	(3, 6), (4, 5), (5, 4), (6, 3)	4
10	(4, 6), (5, 5), (6, 4)	3
11	(5, 6), (6, 5)	2
12	(6, 6)	1
Total		36

Probability Distribution for Discrete Random Variables

PROBABILITY MASS FUNCTION

Therefore, the probability mass function of X is

x	$p_x(x)$
2	1 / 36
3	2 / 36
4	3 / 36
5	4 / 36
6	5 / 36
7	6 / 36
8	5 / 36
9	4 / 36
10	3 / 36
11	2 / 36
12	1 / 36

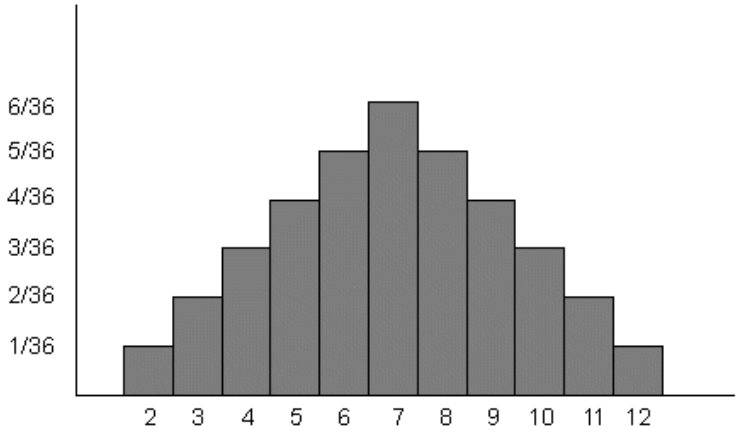
Note that all the probabilities sum up to 1, i.e.

$$\sum_{i=2}^{12} p_x(x) = 1$$

Probability Distribution for Discrete Random Variables

PROBABILITY MASS FUNCTION

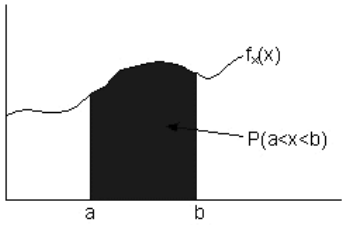
The probability histogram for the random variable X is graphed as follows:



Probability Density Function

The probability density function of a continuous random variable is represented by its probability density function (or density function). It is a function whose graph satisfies the following conditions:

1. **The area bounded by the function curve (or line) and the x-axis is equal to 1.**
2. **The probability that the random variable assumes a value within the interval (a, b) is equal to the area bounded by the curve (or line) of $f_x(x)$, the x-axis, and the two ordinates x = a and x = b.**



Using the definition of a probability density function, the probability that a continuous random variable assumes a specific value will be zero. This is true because the probability that a random variable takes on an exact value *a* is the probability that the random variable assumes a value in the interval (*a*, *a*); in the density function's graph, the area bounded by the curve, the x-axis, and the two parallel lines that both pass through (b, 0), is zero.

Probability Density Function

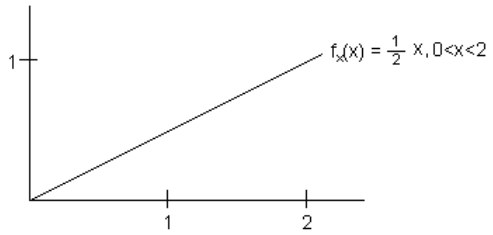
Intuitively, this fact about continuous random variables states that a continuous random variable does not assume a specific value, but rather it takes on a value in a specified interval with a certain probability. For example, the random variable Z whose value is the height of a person in California is a continuous random variable. The probability that Z has the value of 66 (inches) is zero, but the probability that the value of Z is between 64 and 68 is some positive value.

<u>NOTE</u>: For a continuous random variable X, its probability density function satisfies the following equality:

$$P(a < X < b) = \int_a^b f_x(x)dx$$

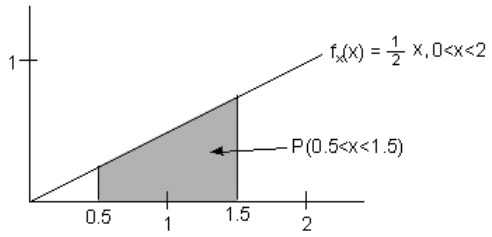
<u>Example</u>: The probability density function of the continuous random variable X is as follows:

$$f_x(x) = \begin{cases} \frac{1}{2}x, 0 < x < 2 \\ 0, otherwise \end{cases}$$



Probability Density Function

To show that is indeed a density, we must show that the area bounded by the line graph of and the x-axis is equal to 1. The area in question is that of a right triangle, with legs of length 1 and 2. Thus, its area is 0.5(1)(2) = 1.



To find the value of P(0.5<X<1.5) we must find the area bounded by the line graph of $f_x(x)$, the x-axis, and the ordinates x = 0.5 and x = 1.5. This area is that of a trapezoid, with parallel sides of length 0.25 and 0.75, and a base of length 1. Thus, its area is 0.5(1)(0.25 + 0.75) = 0.5.

Cumulative Probability Function

The cumulative probability function of a random variable (discrete or continuous) is a function whose domain is similar to that of the probability mass or density function, but whose range is the set of probabilities associated with the possibility that the random variable will assume a value that is less than or equal to the values in the domain. The cumulative probability function of a random variable X is denoted by $F_x(x)$ and is defined as $F_x(x) = P(X \leq x)$.

For a discrete random variable X with a probability mass function of $p_x(x)$ its cumulative probability function $F_x(x)$ is given by

$$F_X(X) = \sum_{a \leq x} p_x(a)$$

<u>NOTE</u>: A continuous random variable Y with a probability density function of has a cumulative distribution function $F_y(y)$ given by

$$F_Y(y) = P(Y \leq y) = P(-\infty \leq Y \leq y) = \int_{-\infty}^{y} f_y(y)dy$$

Cumulative Probability Function

x	pX(x)	FX(x)
2	1 / 36	1/36
3	2 / 36	3/36
4	3 / 36	6/36
5	4 / 36	10/36
6	5 / 36	15/36
7	6 / 36	21/36
8	5 / 35	26/36
9	4 / 36	30/36
10	3 / 36	33/36
11	2 / 36	35/36
12	1 / 36	1

Note that for x = 2, 3, 4, ..., 12.

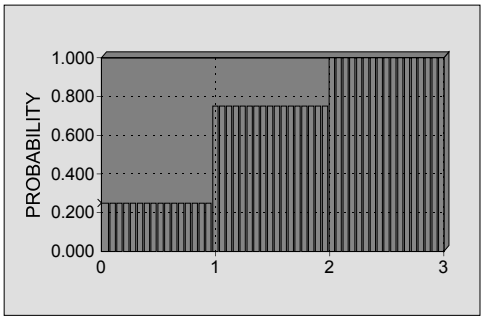
$$F_x(x) = \sum_{a=2}^{12} p_x(a)$$

Probability Distribution for Discrete Random Variables

The cumulative distribution F(x) of a discrete random variable X with probability distribution f(x) is given by:

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), -\infty < x < \infty$$

The graph of F(x) is portrayed at right for the two coin toss experiment. The magnitude of the jumps at 0, 1 and 2 are 1/4, ½ and 1/4.



Probability Distribution for Continuous Random Variables

A continous random variable has a probability of zero of assuming exactly any of its values. Consequently, its probability distribution cannot be given in tabular form. In dealing with continous variables, f(x) is usually called the probability density function or simply the density function of X. A probability density function is constructed so that the area under its curve bounded by the x axis is equal to 1 when computed over the range of X for which f(x) is defined.

Therefore, the function f(x) is a probability density function for the continuous random variable X, defined over the set of real numbers R, if

$$f(x) \geq 0, \text{for all } x \in R$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a < X < b) = \int_{a}^{b} f(x)dx$$

The cumulative distribution F(x) of a continuous random variable X with density function f(x) is given by:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, -\infty < x < \infty$$

Probability Distribution of a Continuous Random Variable

<u>Definition</u>: Let X be a continuous random variable defined over an interval (a, b). Then the function f(x) is called the probability density function (p.d.f.) of X if the following conditions are satisfied:

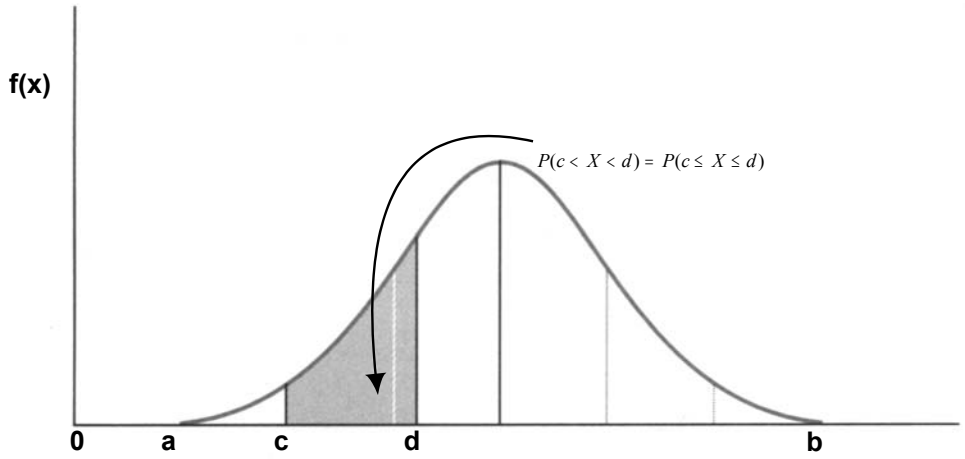
- **(I) $f(x) \geq 0$ for all $a<x<b$**
- **(ii) The total area under the curve f(x) above the X-axis and bounded by the vertical lines x = a and x = b is equal to $P(a \leq X \leq b) = P(a<X<b) = 1$.**

  (Here a and/or b may be replaced by $-\infty$ or ∞).

<u>Note</u>: For a continuous random variable X, the only meaningful probability refers to that of X lying in an interval of its domain. The probability that X = x where x is a specific real value is always zero since probability is equivalent to the area under a curve and at any specific value, x the width of the interval is zero units, that is,

$$\int_{x}^{x} f(x)dx = 0.$$

Probability Distribution of a Continuous Random Variable



$$P(X = a) = P(X = b) = P(X = c) = P(X = d) = 0 \qquad P(a \le X \le b) = 1.$$

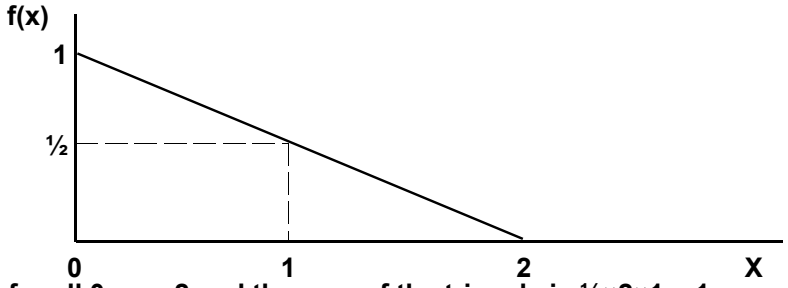
Definition: The cumulative function or distribution function of X is given as

$$F(x) = P(X \le x) = P(a \le X \le x) = \int_a^x f(x)dx.$$

Hence, P(c<X<d) = P(c≤X≤d) = F(d) - F(c) above.

Probability Distribution of a Continuous Random Variable

Example: Let the random variable X have the probability density function whose graph appears as follows:



Since f(x)≥ 0 for all 0≤ x ≤ 2 and the area of the triangle is ½×2×1 = 1 square unit, then f(x) is a p.d.f. by definition. As well,

$$P(1 < X < 2) = \tfrac{1}{2} \times 1 \times \tfrac{1}{2} = \tfrac{1}{4} = 0.25$$

$$P\left(\tfrac{1}{2} \le X \le 1\tfrac{1}{2}\right) = \left(1\tfrac{1}{2} - \tfrac{1}{2}\right)\left\{\left(\tfrac{3}{4} + \tfrac{1}{4}\right)/2\right\} = \tfrac{1}{2} = 0.5$$

$$F(1) = P(X \le 1) = P(0 < X < 1) = 1 - P(1 < X < 2) = 1 - 0.25 = 0.75$$