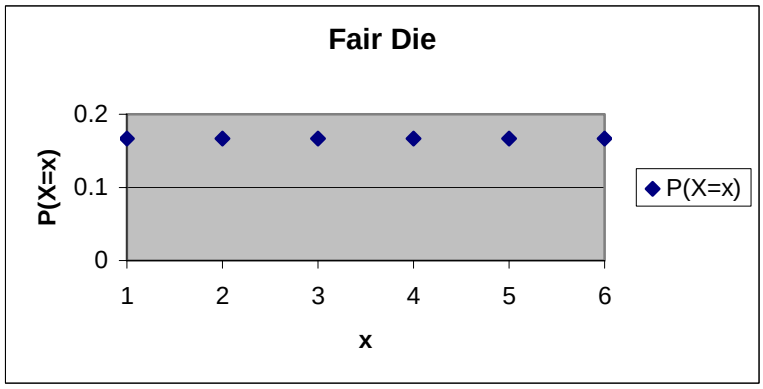
Some Basic Concepts in Statistics

One of the most important basic concepts in statistics is that of a *random variable*. A random variable is one which can assume certain values, dependent on the situation, with certain probabilities. We usually refer to the random variable with a capital letter, and individual values with lower case.

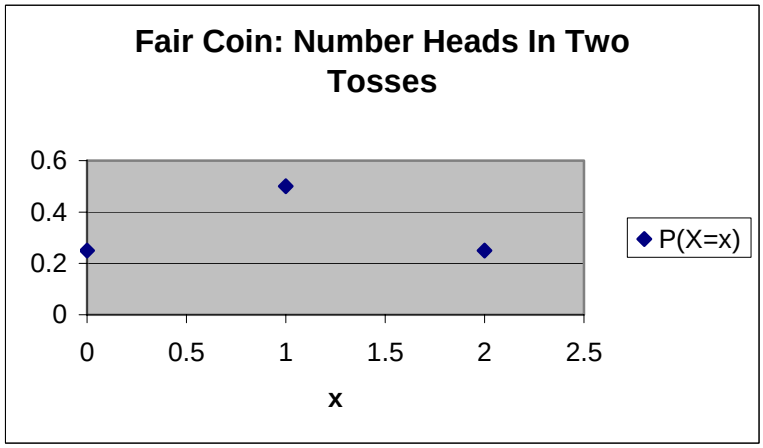
Example 1
Let X be the face value of a fair die when rolled. Then X can take on the values 1, 2, 3, 4 , 5 and 6, each with probability $\frac{1}{6}$. So for any x in the set {1, 2, 3, 4, 5, 6} we can say $P(X = x) = \frac{1}{6}$. The graph of the probability distribution looks like this.



♦♦♦

Example 2
Suppose a fair coin is tossed twice. Let X be the number of times a head is obtained. Then the table below gives the possible outcomes and their probabilities.

x	0	1	2
$P(X=x)$	1/4	1/2	1/4



♦♦♦

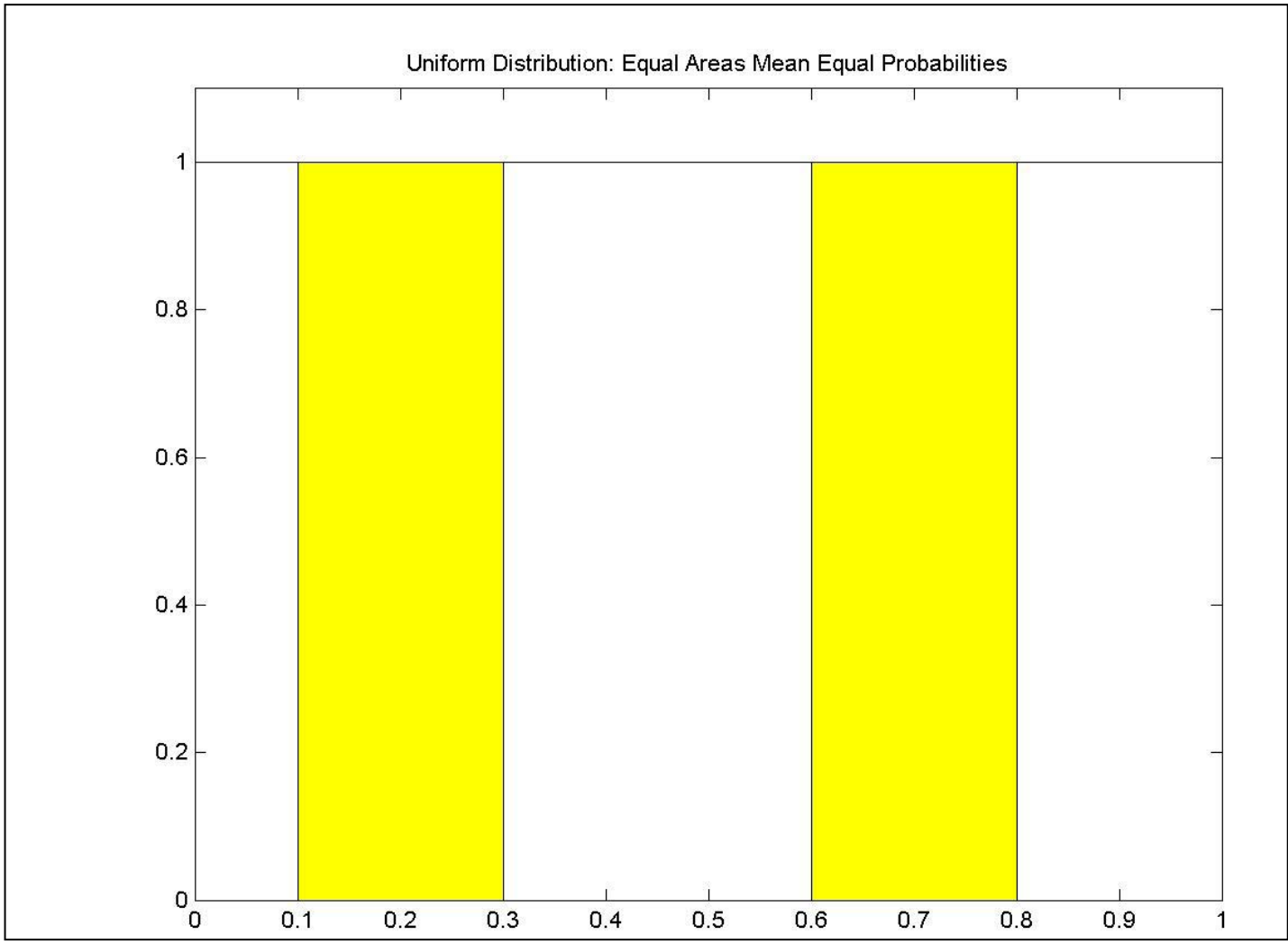
Note the basic principle that the sum of all probabilities is always 1. It is impossible for the sum to be greater than 1, as that would mean that there was more than a 100% chance that something would happen, which is absurd. If the sum was less than 1 then we would have neglected some of the possible outcomes. Of course all probabilities must be positive or zero.

The random variables above are called *discrete* random variables. This is because they have a finite number of outcomes. It is possible for an infinite number of outcomes to also be discrete, for example when that the outcomes can be matched up with the integers.

Example 3
Suppose the random variable Y is given by the number of cosmic rays being detected in a one hour period. What is the range of Y? Certainly Y cannot be less than 0. What upper limit do we accept? If we accept any upper limit at all, then we are limiting the number of cosmic rays entering the Earths atmosphere. No matter what upper limit you choose, it is possible to envisage a scenario whereby more cosmic rays could be detected in a one hour period. This might require so many supernovas exploding next to the Earth that life would be impossible, but nevertheless we cannot accept an actual limit. Hence we accept a range for Y of $0 \le y < \infty$, even though in practice we know that there will be some practical limit beyond which Y is unlikely to go.
♦♦♦

A *continuous* random variable, by contrast, is one which can take on an infinite number of values, and those values cannot be matched up with the integers. For example X might be allowed to assume any real number between 0 and 1, inclusive. It is a basic result in real analysis that no such interval of the real numbers can be matched up with integers ("mapped" to the integers in technical parlance). We shall assume this result without proof.

Example 4
Allow the random variable Z to be a continuous one whose values can range from 0 to 1, and every value is equally likely. Actually, for any $z \in [0,1]$ $P(Z = z) = 0$, as there are infinitely many values, all equally likely. So the height of the graph at any z value does not represent the actual probability that z will occur. Instead we accept that the *area* between, say, $z=a$ and $z=b$ is the probability that $a < z < b$. Of course if the interval $[a,b]$ is the same length as the interval $[c,d]$, and both intervals are completely within the interval [0,1], then $P(a \le z \le b) = P(c \le z \le d)$. The diagram below illustrates that when the areas beneath the graph are equal, then the probability that Z will take a value in either of those two areas is equal. Of course this can also be true for distributions that are not uniform, but in this case provided the two intervals are of equal width then the probabilities will always be equal.


Uniform Distribution: Equal Areas Mean Equal Probabilities
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

♦♦♦

Probability Density and Distribution Functions

So far we have used the term *distribution* without really defining it. Informally, it describes the way in which the probability is distributed between the different values of X. Formally, we have two different concepts, that of a *probability density* function and that of a *probability distribution* function. In fact the density function describes how much probability is *at or near* a value of X. The distribution function shows how much probability there is *up to* the value being considered. The formal definitions are as follows.

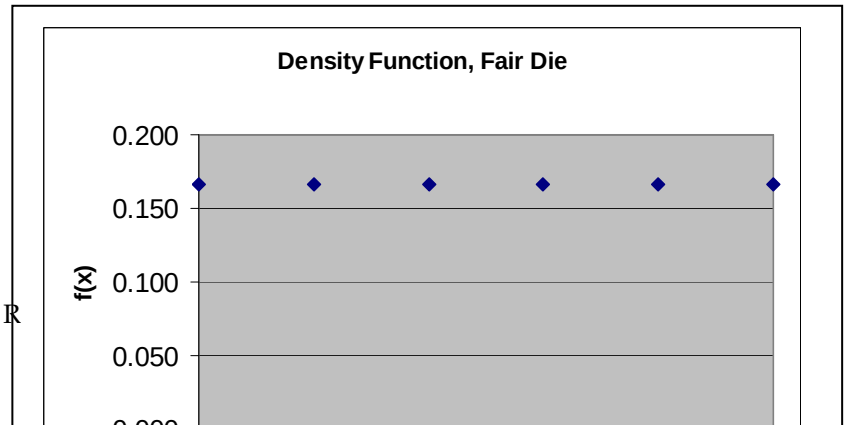
Density And Cumulative Distribution Functions For The Discrete Case

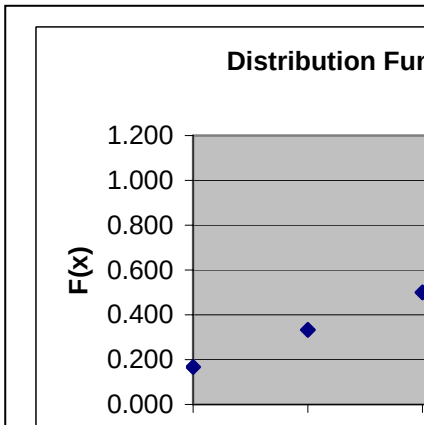
In the discrete case, the probability density function $f(x)$ gives the probability that the random variable X will assume the value x. The probability distribution function $F(x)$gives the probability that X assumes a value less than or equal to x.

Discrete Case	
Density Function	$f(x)=P(X=x)$
Cumulative Distribution Function	$F(x)=P(X\leq x)$

Example 5

For the case of a fair die being rolled and the value on the face being assigned to X, clearly the density function is given by $f(x)=\frac{1}{6}$. In this case the density function is constant. However the cumulative distribution function is the probability that X is less than or equal to x. In this case, $F(x)=\frac{x}{6}$. The graphs are below.





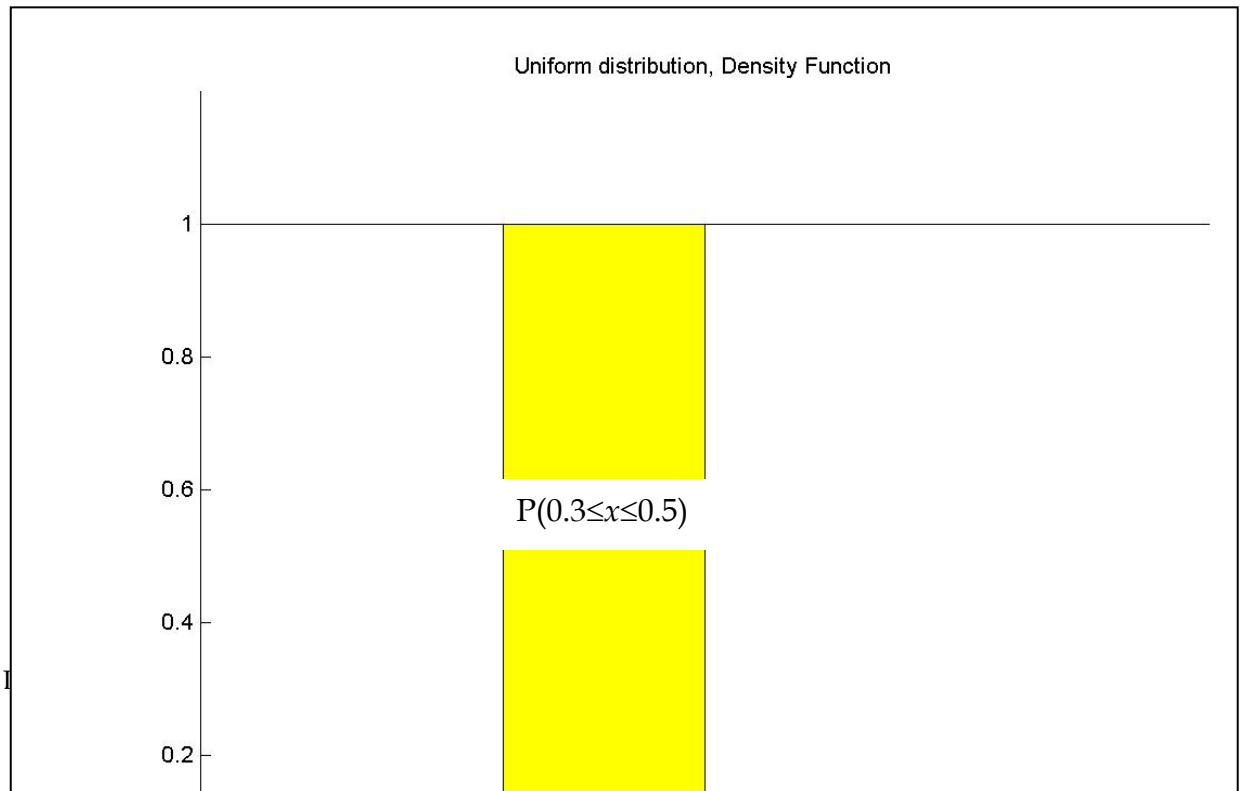
G R

♦♦♦

Density And Cumulative Distribution Functions For The Continuous Case

Because any particular value of X has a zero probability of occurring, because there are infinitely many possible values, the density function $f(x)$ satisfies the property that $P(a \le X \le b) = \int_a^b f(x)dx$. In other words the probability that X will be between a and b is equal to the area beneath $f(x)$ from $x=a$ to $x=b$. The distribution function $F(x)$, by contrast, gives the probability that X will take a value of at most x. In other words, $P(X \le x) = \int_{-\infty}^{x} f(u)du$. Of course sometimes $f(x)=0$ in most of its domain, so the lower limit of integration won't always be $-\infty$.

Example 6
In the case of the uniform distribution, the probability $P(a \le x \le b) = \text{area underneath line}$. Now earlier we showed a graph with the height at 1, without justifying that. In fact, if the distribution is from 0 to 1, then the total area beneath that must be 1 as always. So the height of the line must be 1, as the area of a rectangle is base*height. So what is the area between $x=a$ and $x=b$? It must be $\frac{1}{b-a}$. Hence $P(a \le x \le b) = \frac{1}{b-a}$.



GRI

What about the distribution function? We know that the density function $f(x)=1$, when $0\le x\le 1$. Hence we obtain the distribution function by integrating, ie

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u)du \\ &= \int_{0}^{x} 1du \\ &= \left[u\right]_0^x \\ &= x \end{aligned}$$

← Note that $x\le 1$ here.

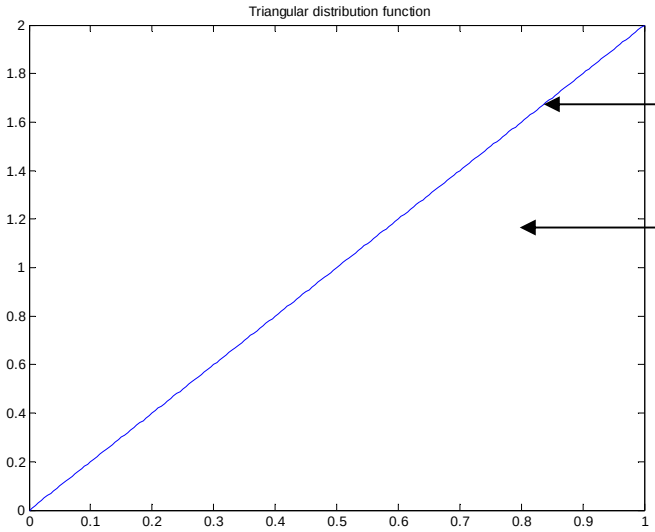
Of course below 0 or above 1, $f(x)=0$, so $F(x)=0$ whenever $x\le 0$. And x must be no more than 1, so $F(x)=1$ whenever $x\ge 1$. Hence our distribution function turns out to be

$$F(x)=\begin{cases} 0, & x<0 \\ x, & 0\le x\le 1 \\ 1, & x>1 \end{cases}$$

♦ ♦ ♦

Example 7

Suppose the random variable X is given by a probability distribution function that consists of a straight line through the origin, reaching its maximum value when $X=1$, which is the largest value X can achieve. Then the graph of $f(x)$ must be as below, bearing in mind that the total area must be 1.



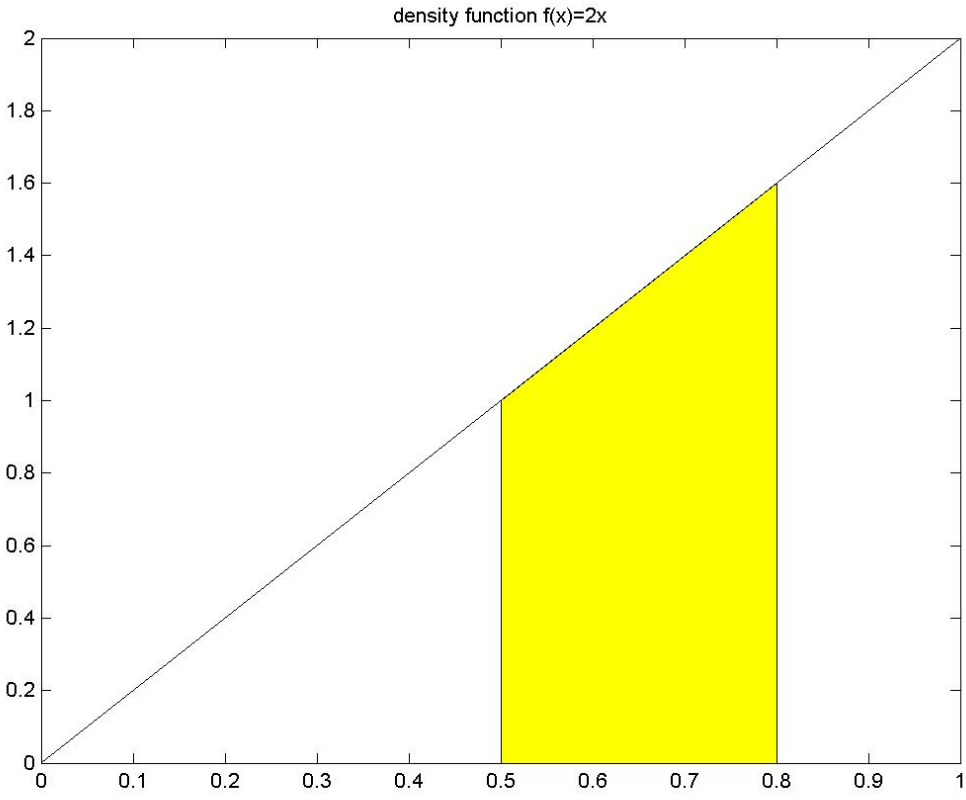
$y=2x$

Area=1/2**base***height*
(Triangle)

Thus the density function is

$$f(x)=\begin{cases} 0, & x<0 \\ 2x, & 0\le x\le 1 \\ 0, & 1<x \end{cases}$$

Hence the probability that the random variable x is between, say, 0.5 and 0.8 will be given by the area shown.



This area can either be found using some basic geometry (area of a triangle is 0.5*base*height) or by performing a formal integration. In this case, we find that $P(0.5 \le x \le 0.8) = 0.39$. The distribution function can be found by integration as follows.

$F(x) = \int_0^x 2u\,du = \left[u^2\right]_0^x = x^2$.

This enables us to answer questions such as which is the x value which ensures $P(X \le x)=0.5$. We solve the equation $x^2 = \frac{1}{2}$, ie $x \approx 0.7071$. This makes perfect sense in light of the above graph, which shows clearly that most of the probability occurs in the top half of the domain, ie more than 50% of the probability is above x=0.5.

♦♦♦

Expected Value And Variance

These concepts convey the same sort of information as the mean and variance for a sample of measured data. In other words the mean, or expected value, of a distribution is a measure of the "middle" of the distribution. It gives us an idea of where the centre of the distribution is located. By contrast, the variance of the distribution gives us an understanding of how spread out the distribution is, or how far from the mean most of the values are likely to be. Their definitions follow.

Mean (Expected Value)

1. Discrete case: $E(X) = \mu = \sum_i x_i * P(x_i) = \sum_i x_i * f(x_i)$
2. Continuous case: $E(X) = \mu = \int_{-\infty}^{\infty} f(x)dx$

Note that in the continuous case $f(x)$ is considered to be 0 outside the relevant domain. Hence this definition suffices for all cases, including those where the domain is restricted.

Variance[1]

1. Discrete case: $\mathrm{Var}(X) = \sigma^2 = \sum_i (x_i - \mu)^2 f(x_i)$
2. Continuous case: $\mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$

Example 8

Find the mean and variance of X if X is a random variable whose value is given by the roll of a fair die.

[1] Of course the standard deviation, σ, is just the square root of the variance.

x_i	$f(x_i)$	$x_i * f(x_i)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6
Total		**21/6**

Hence the expected value, or mean, is 21/6 or 7/2=3.5. **(**Observe that we would expect this value, from the symmetry of the *pdf*, ie it should be in the middle of the set 1,2,…,6.). Note that the mean is not necessarily a value that *X* can actually take. What about the variance? The next table shows how that might be found.

x_i	$f(x_i)$	$(x_i-\mu)^2$	$(x_i-\mu)^2 * f(x_i)$*
1	1/6	6.25	1.041
2	1/6	2.25	0.375
3	1/6	0.25	0.042
4	1/6	0.25	0.042
5	1/6	2.25	0.375
6	1/6	6.25	1.042
		Total	**2.92**

Hence the variance is 2.92.

♦♦♦

For the sake of computational efficiency, and for accuracy because round-off error is reduced, it is common to use the form $\sigma^2 = \sum_i x_i^2 f(x_i) - \mu^2$, as below.

Example 9

x_i	$f(x_i)$	x_i^2	$x_i^2 * f(x_i)$
1	1/6	1	1/6
2	1/6	4	4/6
3	1/6	9	9/6
4	1/6	16	16/6
5	1/6	25	25/6
6	1/6	36	36/6
		Total	**91/6**

Hence the variance is $\frac{91}{6} - (3.5)^2 = 2.92$.

♦♦♦

Example 10

Find the mean and variance of the uniform distribution from *Examples 4* and *6*. In this case, $f(x)=1$.

$$E(X) = \mu = \int_0^1 x * 1 dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

$$Var(X) = \sigma^2 = \int_0^1 (x - \mu)^2 * 1 dx$$

$$= \int_0^1 (x - 0.5)^2 \, dx$$

Now we make a change of variable, setting $u = x - 0.5$.

$du = dx$, of course, and when $x = 0$, $u = -0.5$. When $x = 1$, $u = 0.5$.

Hence,

$$\sigma^2 = \int_{u=-0.5}^{0.5} u^2 du = \left[\frac{u^3}{3} \right]_{-0.5}^{0.5} = \frac{(0.5)^3 - (-0.5)^3}{3} = 0.083$$

♦ ♦ ♦

Example 11

Two fair dice are tossed and the sum is recorded. Find the mean and variance of the sum.

Die 1 \ Die 2	*1*	*2*	*3*	*4*	*5*	*6*
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

x	$P(x)$	$x*P(x)$
2	1/36	2/36
3	2/36=1/18	6/36
4	3/36=1/12	12/36
5	4/36=1/9	20/36
6	5/36	30/36
7	6/36=1/6	42/36
8	5/36	40/36
9	4/36=1/9	36/36
10	3/36=1/12	30/36
11	2/36=1/18	22/36
12	1/36	12/36
	Total	252/36=7

Hence the mean is 7, which in this case is an achievable value. Now we can find the variance.

x	$P(x)$	$x^2*P(x)$
2	1/36	4/36
3	2/36	18/36
4	3/36	48/38
5	4/36	100/36
6	5/36	180/36
7	6/36	294/36
8	5/36	320/36
9	4/36	324/36
10	3/36	300/36
11	2/36	242/36
12	1/36	144/36
	Total	1974/36

Hence the variance is $\frac{1974}{36} - 7^2 = \frac{210}{36} = 35/6$.

♦ ♦ ♦

Example 12

Find the mean and variance of the triangular distribution of *Example 7*. This being a continuous distribution, we need to use the integration formulae.

$$\mu = \int_0^1 xf(x)dx$$
$$= \int_0^1 2x^2 dx$$
$$= 2\left[\frac{x^3}{3}\right]_0^1$$
$$= \frac{2}{3}$$

$$\sigma^2 = \int_0^1 (x-\mu)^2 f(x)dx$$

$$= \int_0^1 \left(x - \frac{2}{3}\right)^2 * 2x dx$$

$$= 2\int_0^1 \left(x^3 - \frac{4}{3}x^2 + \frac{4}{9}x\right)dx$$

$$= 2\left[\frac{x^4}{4} - \frac{4}{9}x^3 + \frac{2}{9}x^2\right]_0^1$$

$$= 2\left(\frac{1}{4} - \frac{4}{9} + \frac{2}{9}\right)$$

$$= 2\left(\frac{9}{36} - \frac{8}{36}\right) = \frac{1}{18}$$

♦ ♦ ♦