Queen Mary
University of London

MAS 108 Probability I

Notes 9 Autumn 2005

Cumulative distribution function

We have already defined the *cumulative distribution function* (abbreviated to c.d.f.) of a random variable. The c.d.f. of the random variable X is defined by

$$F_X(x) = P(X \leq x) \qquad \text{for } x \text{ in } \mathbb{R}$$

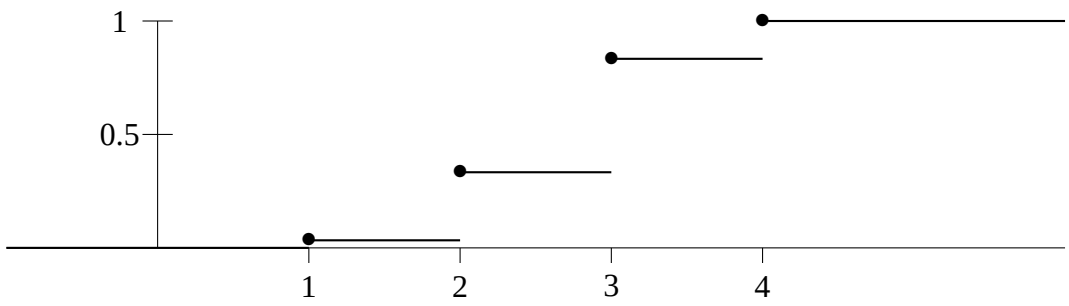
As usual, we write $F(x)$ rather than $F_X(x)$ if it is clear which random variable is meant.

Example I have 3 copper coins and 7 silver coins in my pocket. I randomly take out 4 coins. Let X be the number of silver coins in my sample. The p.m.f. of X is

x	1	2	3	4
$P(X=x)$	$\frac{1}{30}$	$\frac{9}{30}$	$\frac{15}{30}$	$\frac{5}{30}$

So the cumulative distribution function is

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{30} & \text{if } 1 \leq x < 2 \\ \frac{10}{30} & \text{if } 2 \leq x < 3 \\ \frac{25}{30} & \text{if } 3 \leq x < 4 \\ 1 & \text{if } 4 \leq x \end{cases}$$



This graph of F is typical of the graph of a c.d.f. of a discrete random variable. At each value x that X takes, the graph of F 'jumps'; in other words, F is discontinuous. Between neighbouring values of X, the graph is flat.

Here are some properties of the c.d.f.

(a) $0 \leq F(x) \leq 1$ for all x in $\mathbb{R}$.

(b) As $x \to \infty$, $F(x) \to 1$.

(c) As $x \to -\infty$, $F(x) \to 0$.

(d) If $x < y$ then $F(x) \leq F(y)$ (this means that F is a *non-deccreasing* function).

(e) If $x < y$ then $P(x < X \leq y) = F(y) - F(x)$.

(f) $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$.

Continuous random variables

Example Archer: part 1 An archer shoots an arrow at a circular target with radius 60 cm. Suppose that the arrow always hits the target. Let X be the distance from the centre of the target to the point where the arrow hits, measured in cm.

Given any region A of the target, it is reasonable to assume that the

$$P(\text{arrow lands in } A) = \frac{\text{area of } A}{\text{area of target}}.$$

In particular, if A is the circle of radius x cm centered at the origin of the target then

$$F(x) = P(X \leq x) = P(\text{arrow lands in } A) = \frac{\pi x^2}{\pi 60^2} = \frac{x^2}{60^2}.$$

So the c.d.f. is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x^2}{60^2} & \text{if } 0 \leq x < 60 \\ 1 & \text{if } 60 \leq x \end{cases}$$

The graph of F is continuous; it has no holes. Also, it has non-zero slope apart from at the two ends.

Roughly speaking, a random variable is *continuous* if there are no gaps between its possible values. For example, the height of a randomly chosen student in the class could in principle be any real number between certain extreme limits. A random variable whose values range over an interval of real numbers, or even over all real numbers, is continuous.

There are two crucial properties. One is that there are no gaps. The other is that, for any real number x, we have $P(X = x) = 0$; that is, the probability that the height of a random student, or the time I have to wait for a bus, is *precisely* x, is zero. So we can't use the probability mass function for continuous random variables; it would always be zero and give no information.

We use the *cumulative distribution function* or c.d.f. instead. Here is the formal definition. A random variable X is *continuous* if its cumulative distribution function F is a continuous function and it has non-zero slope apart (possibly) from at the two ends of the real line.

Now let X be a continuous random variable. Then, since the probability that X takes the precise value x is zero, there is no difference between $P(X \leq x)$ and $P(X < x)$. Thus, in addition to the previous properties of c.d.f., we also have

(a) $P(X = x) = 0$ for all x in $\mathbb{R}$;

(b) if $a < b$ then $P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = F(b) - F(a)$;

(c) $P(X \leq x) = F(x) = P(X < x)$ for all x in $\mathbb{R}$;

(d) $P(X > x) = 1 - F(x) = P(X \geq x)$ for all x in $\mathbb{R}$.

Since F is continuous and non-decreasing, a result from Calculus tells us that it is differentiable 'almost everywhere', that is, everywhere apart possibly from a few corners (there is just one corner in the archer's c.d.f.). So we define the *probability density function* f_X to be this derivative. We often abbreviate 'probability density function' to 'p.d.f.'. As usual, we write just $f(x)$ if the random variable X is clear from the context.

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x).$$

Now $f_X(x)$ is non-negative, since it is the derivative of an increasing function. If we know $f_X(x)$, then F_X is obtained by integrating. Because $F_X(-\infty) = 0$, we have

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

Note the use of the "dummy variable" t in this integral. Note also that

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t)\mathrm{d}t.$$

Letting a tend to $-\infty$ and b tend to $+\infty$ gives

$$\int_{-\infty}^{\infty} f_X(t)\mathrm{d}t = P(-\infty < X < \infty) = 1.$$

You can think of the p.d.f. like this: the probability that the value of X lies in a very small interval from x to $x+h$ is approximately $f_X(x) \cdot h$. This is because, if h is small,

$$\frac{F_X(x+h) - F(x)}{h} \approx f_X(x).$$

So, although the probability of getting exactly the value x is zero, the probability of being close to x is proportional to $f_X(x)$.

There is a mechanical analogy which you may find helpful. Remember that we modelled a discrete random variable X by placing at each value a of X a mass equal to $P(X = a)$. Then the total mass is one, and the expected value of X is the centre of mass. For a continuous random variable, imagine instead a wire of variable thickness, so that the density of the wire (mass per unit length) at the point x is equal to $f_X(x)$. Then again the total mass is one; the mass to the left of x is $F_X(x)$; and again it will hold that the centre of mass is at $E(X)$.

Example Archer: part 2 Differentation gives

$$f(x) = F'(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{2x}{60^2} & \text{if } 0 < x < 60 \\ 0 & \text{if } 60 \leq x \end{cases}$$

Most definitions and facts about continuous random variables are obtained by replacing the p.m.f. by the p.d.f. and replacing sums by integrals. Thus, if X is a continuous random variable with p.d.f. f, and g is a real function then

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x.$$

In particular, the *expected value* of X is given by

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x,$$

and the *variance* is (as before)

$$\mathrm{Var}(X) = E((X-\mu))^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) \mathrm{d}x,$$

where $\mu = E(X)$.

Note that Theorem 3 (equality of two different formulae for variance), Theorem 4 (properties of expectation), Theorem 5 (properties of variance) and Proposition 7 (symmetric random variables) are still true for continuous random variables. In particular

$$\mathrm{Var}(X) = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x - \mu^2.$$

Example Archer: part 3

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x = \int_0^{60} \frac{2x^2}{60^2} \mathrm{d}x = \left[\frac{2x^3}{3 \times 60^2} \right]_{x=0}^{x=60} = \frac{2 \times 60}{3} = 40.$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x = \int_0^{60} \frac{2x^3}{60^2} \mathrm{d}x = \left[\frac{2x^4}{4 \times 60^2} \right]_{x=0}^{x=60} = \frac{1}{2} \times 60^2 = 1800,$$

and so

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 = 1800 - 40^2 = 1800 - 1600 = 200.$$

Notice how I began by writing the formal definition of $E(X)$, with limits of integration $-\infty$ and ∞, but immediately changed the limits to 0 and 60 because $f(x)$ is zero outside the interval $[0,60]$. We use this idea very often. The *support* of a continuous random variable X is defined to be the smallest interval containing all values of x where $f_X(x) > 0$. All integrals can be taken just over the support of the interval, but you must still take care to define the c.d.f. and the p.d.f. on the whole real line.

Example Suppose that the random variable X has p.d.f. given by

$$f_X(x) = \begin{cases} \frac{1}{2}x^{-1/2} & \text{if } 0 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The support of X is the interval $[0,1]$. We check the integral:

$$\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = \int_0^1 \frac{1}{2}x^{-1/2}\,\mathrm{d}x = \left[x^{1/2}\right]_{x=0}^{x=1} = 1.$$

The cumulative distribution function of X is

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t = \begin{cases} 0 & \text{if } x < 0, \\ x^{1/2} & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

(Study this carefully to see how it works.) We have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x = \int_0^1 \frac{1}{2}x^{1/2}2\mathrm{d}x = \frac{1}{3}, \\ E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)\mathrm{d}x = \int_0^1 \frac{1}{2}x^{3/2}\mathrm{d}x = \frac{1}{5}, \\ \mathrm{Var}(X) &= \frac{1}{5} - \left(\frac{1}{3}\right)^2 = \frac{4}{45}. \end{aligned}$$

Median, quartiles, percentiles

Another measure commonly used for continuous random variables is the *median*; this is the value m such that "half of the distribution lies to the left of m and half to the right". More formally, m should satisfy $F_X(m) = 1/2$. It is not the same as the mean or expected value.

In the example at the end of the last section, we saw that $E(X) = 1/3$. The median of X is the value of m for which $F_X(m) = 1/2$. Since $F_X(x) = x^{1/2}$ for $0 \leq x \leq 1$, we see that $m^{1/2} = 1/2$, or $m = 1/4$, which is not the same.

If there is a value m such that the graph of $y = f_X(x)$ is symmetric about $x = m$, then both the expected value and the median of X are equal to m.

The *lower quartile* l and the *upper quartile* u are similarly defined by

$$F_X(l) = 1/4, \qquad F_X(u) = 3/4.$$

Thus, the probability that X lies between l and u is $3/4 - 1/4 = 1/2$, so the quartiles give an estimate of how spread-out the distribution is. More generally, we define the nth *percentile* of X to be the value of x_n such that

$$F_X(x_n) = n/100,$$

that is, the probability that X is smaller than x_n is n%. In the same vein, the *top decile* is the value t such that

$$F(t) = \frac{9}{10}.$$

So in the earlier example, the quartiles are $1/16$ and $9/16$. The probability of the event $1/16 < X < 9/16$ is equal to $1/2$.

Example Archer: part 4 The median m satisfies $1/2 = F(m) = m^2/60^2$ so $m^2 = 1800$ and $m = 30\sqrt{2} \approx 42.43$. Similarly, the lower quartile is 30 and the upper quartile is $30\sqrt{3} \approx 51.96$.

The *support* of a continuous random variable X is an interval, possibly a semi-infinite interval or even the whole real line. We don't care whether or not the endpoints of the interval are included, since as we have seen, the probability of getting one precise value is zero. In a sense, the support of X stretches from the 0th to the 100th percentile!

The median, quartiles and so on are also defined for discrete random variables, but there is a problem: sometimes there is *no* solution to the equation $F(x) = 1/2$ (see the first example in these notes) and sometimes there are *too many* solutions. So in general we have to define the median to be *any* value m which satisfies

$$P(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq m) \geq \frac{1}{2}.$$

Similarly, the lower quartiles is any value l which satisfies

$$P(X \leq l) \geq \frac{1}{4} \quad \text{and} \quad P(X \geq l) \geq \frac{3}{4}$$

and the upper quartile is any value u which satisfies

$$P(X \leq u) \geq \frac{3}{4} \quad \text{and} \quad P(X \geq u) \geq \frac{1}{4}.$$

If X is continuous then $P(X \geq x) = 1 - P(X \leq x)$ for all x and so the median is indeed the unique solution m of $F(m) = 1/2$, and similarly for the quartiles and other percentiles.

Worked examples

Let X be a continuous random variable whose probability density function f is given by

$$f(x) = \begin{cases} 0 & \text{if } x < 4 \\ \theta & \text{if } 4 \leq x \leq 10 \\ 0 & \text{if } x > 10 \end{cases}$$

for some constant θ.

To find θ, we use the fact that the integral of f over the whole real line is 1. Thus

$$1 = \int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_4^{10} \theta \mathrm{d}x = [\theta x]_{x=4}^{x=10} = 6\theta,$$

so $\theta = 1/6$.

To find the c.d.f. F, we use the fact that

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

If $x \leq 4$ then $f(t) = 0$ for all t in $(-\infty, x)$ so $F(x) = 0$. If $4 \leq x \leq 10$ then

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \int_4^x \frac{1}{6}\mathrm{d}t = \left[\frac{t}{6}\right]_{t=4}^{t=x} = \frac{x-4}{6}.$$

If $x \geq 10$ then $f(t) = 0$ for all t in $(-\infty, 4)$ and all t in $(10, x)$ so

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \int_4^{10} \frac{1}{6}\mathrm{d}t = 1.$$

In summary:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 4 \\ \frac{x-4}{6} & \text{if } 4 \leq x \leq 10 \\ 1 & \text{if } x \geq 10 \end{cases}$$

We use F to find some particular probabilities:

$$\begin{aligned} P(X \leq 8) &= F(8) = \frac{2}{3} \\ P(X \geq 5) &= 1 - F(5) = \frac{5}{6} \\ P(X \geq 12) &= 1 - F(12) = 0 \\ P(7 \leq X \leq 9) &= F(9) - F(7) = \frac{1}{3}. \end{aligned}$$

We use f to find the expectation and variance.

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x = \int_4^{10} \frac{x}{6} \mathrm{d}x = \left[\frac{x^2}{12}\right]_{x=4}^{x=10} = \frac{100-16}{12} = 7.$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x = \int_4^{10} \frac{x^2}{6} \mathrm{d}x = \left[\frac{x^3}{18}\right]_{x=4}^{x=10} = \frac{1000-64}{18} = 52,$$

so

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 = 52 - 7^2 = 3.$$

In the second example we start with the c.d.f. Let X be a continuous random variable whose cumulative distribution function F is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \dfrac{1-\cos x}{2} & \text{if } 0 \le x \le \pi \\ 1 & \text{if } \pi < x \end{cases}$$

We differentiate this to find the probability density function.

$$f(x) = F'(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2}\sin x & \text{if } 0 < x < \pi \\ 0 & \text{if } \pi < x \end{cases}$$

The easy way to find the expectation is to notice that X is symmetric about $\pi/2$ so $E(X) = \pi/2$. The traditional way is to use f and integrate. We need to integrate by parts.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) \mathrm{d}x = \int_0^{\pi} \frac{1}{2} x \sin x \, \mathrm{d}x = \left[-\frac{1}{2} x \cos x\right]_{x=0}^{x=\pi} + \int_0^{\pi} \frac{1}{2} \cos x \, \mathrm{d}x \\ &= \frac{\pi}{2} + \left[\frac{1}{2}\sin x\right]_{x=0}^{x=\pi} = \frac{\pi}{2}. \end{aligned}$$

To find the median, again the easy way is to use symmetry to deduce that the median is $\pi/2$. Otherwise we use F: the median m satisfies

$$\frac{1}{2} = F(m) = \frac{1-\cos m}{2},$$

so $\cos m = 0$, so m is equal to $\pi/2$ plus some multiple of π. But the formula we have used for $F(m)$ holds only if $0 \le m \le \pi$, so $m = \pi/2$.

Similarly, the lower quartile l satisfies $0 \le l \le \pi$ and $1/4 = F(l) = (1-\cos l)/2$ so $\cos l = 1/2$ and $l = \pi/3$. Likewise, the upper quartile u satisfies $0 \le u \le \pi$ and $3/4 = F(u) = (1-\cos u)/2$ so $\cos u = -1/2$ and $u = 2\pi/3$.