Queen Mary
University of London

MTH4106 Introduction to Statistics

Notes 7 Spring 2011

Continuous random variables

If X is a random variable (abbreviated to r.v.) then its *cumulative distribution function* (abbreviated to c.d.f.) F is defined by

$$F(x) = \mathbb{P}(X \leqslant x) \qquad \text{for } x \text{ in } \mathbb{R}.$$

We write $F_X(x)$ if we need to emphasize the random variable X.
(Take care to distinguish between X and x in your writing.)

We say that X is a *continuous* random variable if its cdf F is a continuous function. In this case, F is differentiable almost everywhere, and the *probability density function* (abbreviated to p.d.f.) f of F is defined by

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x} \qquad \text{for } x \text{ in } \mathbb{R}.$$

Again, we write $f_X(x)$ if we need to emphasize X.

If $a < b$ and X is continuous then

$$\begin{aligned} & \mathbb{P}(a < X \leqslant b) = F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x \\ = \; & \mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leqslant X < b). \end{aligned}$$

Moreover,

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

and

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1.$$

The *support* of a continuous random variable X is $\{x \in \mathbb{R} : f_X(x) \neq 0\}$.

The *median* of X is the solution of $F(x) = \frac{1}{2}$; the *lower quartile* of X is the solution of $F(x) = \frac{1}{4}$; and the *upper quartile* of X is the solution of $F(x) = \frac{3}{4}$. More generally, the n-th *percentile* of X is the solution of $F(x) = n/100$.

The *expectation* of X is defined by

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

Similarly, if g is any real function then

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

In particular, the n-th moment of X is

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

and

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

where $\mu = \mathbb{E}(X)$.

Two special continuous random variables

Uniform random variable $U(a,b)$ also known as uniform$[a,b]$

Let a and b be real numbers with $a < b$. A uniform random variable on the interval $[a,b]$ is, roughly speaking, "equally likely to be anywhere in the interval". In other words, its probability density function is some constant c on the interval $[a,b]$ (and zero outside the interval). What should the constant value c be? The integral of the p.d.f. is the area of a rectangle of height c and base $b-a$; this must be 1, so $c = 1/(b-a)$. Thus, the p.d.f. of the random variable $X \sim U(a,b)$ is given by

$$f_X(x) = \begin{cases} 0 & \text{if } x < a, \\ \dfrac{1}{(b-a)} & \text{if } a < x < b, \\ 0 & \text{if } x > b. \end{cases}$$

So the support of X is the interval $[a,b]$, as we would expect. By integration, we find that the c.d.f. is

$$F_X(x) = \begin{cases} 0 & \text{if } x < a, \\ \dfrac{(x-a)}{(b-a)} & \text{if } a \leqslant x \leqslant b, \\ 1 & \text{if } x > b. \end{cases}$$

(Strictly speaking, the support of X is the open interval (a,b), because $f_X(a)$ and $f_X(b)$ are not defined, as F_X is not differentiable at $x=a$ or at $x=b$. However, the results from Calculus that we need to use in Theorem 7 are usually stated in terms of closed intervals. For the purposes of *MTH4106 Introduction to Statistics*, it makes no difference whether we regard the support as a closed interval or an open one.)

To find the expectation and variance, we use a little trick: first find them for the special case $U(0,1)$ and then use standard theorems from *MTH4107 Introduction to Probability*. If $X \sim \text{uniform}[0,1]$ then

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

and

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

so

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Now if $Y \sim \text{uniform}[a,b]$ then $Y = (b-a)X + a$ where $X \sim \text{uniform}[0,1]$. Then

$$\mathbb{E}(Y) = (b-a)\,\mathbb{E}(X) + a = \frac{a+b}{2}$$

and

$$\mathrm{Var}(Y) = (b-a)^2\,\mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

The median m is given by $F_Y(m) = 1/2$, that is,

$$\frac{m-a}{b-a} = \frac{1}{2},$$

whence $m = (a+b)/2$. Note that the expected value and the median of Y are both given by $(a+b)/2$ (the midpoint of the interval). This agrees with the fact that the p.d.f. is symmetrical about the mid-point of the interval.

Exponential random variable $\mathrm{Exp}(\lambda)$

The exponential random variable arises in the same situation as the Poisson: be careful not to confuse them! We have events which occur randomly but at a constant average rate of λ per unit time (e.g. radioactive decays, people joining a queue, people leaving a post-office counter, fish biting). The Poisson random variable, which is discrete, counts how many events will occur in the next unit of time. The exponential random variable, which is continuous, measures exactly how long from now it is until the next event occurs. Note that it takes non-negative real numbers as values and that λ must be positive.

If $X \sim \mathrm{Exp}(\lambda)$, the p.d.f. of X is

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda \mathrm{e}^{-\lambda x} & \text{if } x > 0. \end{cases}$$

So the support of X is the set $(0, \infty)$ of positive real numbers. By integration, we find the c.d.f. to be

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - \mathrm{e}^{-\lambda x} & \text{if } x \geqslant 0. \end{cases}$$

Further calculation gives

$$\mathbb{E}(X) = 1/\lambda, \qquad \mathrm{Var}(X) = 1/\lambda^2.$$

This involves some integration by parts, so brush up your calculus before you try it for yourself.

The median m satisfies $1 - \mathrm{e}^{-\lambda m} = 1/2$, so that $m = \ln 2/\lambda$. (The logarithm is the natural logarithm to base e, so that $\ln 2 = 0.69314718056$ approximately.)

You should compare this value with the value $\mathbb{E}(X) = 1/\lambda$, which is about 40% greater. This kind of situation often arises for random variables which can take only non-negative values. For example, suppose that I select a member of the population at random and let X be his or her annual income. The median of X is the value m such that half the population earn less than m. The expected value is likely to be larger than m, because a few people with very large incomes pull the average up.

In Practical 6, you learnt how to plot the probability density function of uniform random variables and exponential random variables. This would be a good place for you to insert those graphs into your notes.

Functions of continuous random variables

Before doing the final special continuous random variable, we make a diversion about functions of random variables. This needs some ideas from Calculus.

Suppose that the support of X is $[a,b]$, where we include the possibility that $a = -\infty$ or $b = \infty$. If $y < a$ then $f(t) = 0$ for all $t \leqslant y$, and so

$$F(y) = \int_{-\infty}^{y} f(t)\mathrm{d}t = 0.$$

If $y > b$ then $f(t) = 0$ for all $t \geqslant y$, so

$$\mathbb{P}(X \geqslant y) = \int_{y}^{\infty} f(t)\mathrm{d}t = 0$$

and so $F(y) = 1 - \mathbb{P}(X \geqslant y) = 1$.

Suppose that I is an interval in $\mathbb{R}$ and that $g\colon I \longrightarrow \mathbb{R}$ is a real function. Then g is defined to be *monotonic increasing* if $g(x) < g(y)$ whenever $x < y$ and x and y are both in I, while g is *monotonic decreasing* if $g(x) > g(y)$ whenever $x < y$ and x and y are both in I.

For example, if $g(x) = x^3$ then g is monotonic increasing on $\mathbb{R}$; if $g(x) = x^2$ then g is monotonic increasing on $[0,\infty)$ and g is monotonic decreasing on $(-\infty,0]$; and if $g(x) = -3x$ then g is monotonic decreasing on $\mathbb{R}$.

Suppose that $I = [a,b]$. Put $J = g(I) = \{g(x) : x \in I\}$. Calculus gives us the following facts. If g is continuous and monotonic increasing then

(a) J is the interval $[g(a), g(b)]$;

(b) g has an inverse function $h\colon J \longrightarrow I$ such that $g(x) = y$ if and only if $x = h(y)$;

(c) g and h are both differentiable almost everywhere, and $g'(x) \geqslant 0$ and $h'(y) \geqslant 0$ whenever $g'(x)$ and $h'(y)$ exist.

On the other hand, if g is continuous and monotonic decreasing then

(a) J is the interval $[g(b), g(a)]$;

(b) g has an inverse function $h\colon J \longrightarrow I$ such that $g(x) = y$ if and only if $x = h(y)$;

(c) g and h are both differentiable almost everywhere, and $g'(x) \leqslant 0$ and $h'(y) \leqslant 0$ whenever $g'(x)$ and $h'(y)$ exist.

Theorem 7 Let X be a continuous random variable with probability density function f_X and support I, where $I = [a,b]$. Let $g\colon I \to \mathbb{R}$ be a continuous monotonic function with inverse function $h\colon J \to I$, where $J = g(I)$. Let $Y = g(X)$. Then the probability density function f_Y of Y satsfies

$$f_Y(y) = \begin{cases} f_X(h(y))\,|h'(y)| & \text{if } y \in J \\ 0 & \text{otherwise.} \end{cases}$$

Proof Let $J = [c,d]$. If $y < c$ then Y takes no values less than or equal to y, so $F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = 0$. Differentiation gives $f_Y(y) = F_Y'(y) = 0$.

If $y > d$ then Y takes no values greater than or equal to y. Therefore $\mathbb{P}(Y \geqslant y) = \mathbb{P}(g(X) \geqslant y) = 0$; that is, $F_Y(y) = 1 - \mathbb{P}(Y \geqslant y) = 1$. Again, differentiation gives $f_Y(y) = F_Y'(y) = 0$.

If $y \in J$ then $y = g(h(y))$ and so

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant g(h(y))). \tag{1}$$

If g is increasing then $g(X) \leqslant g(h(y))$ if and only if $X \leqslant h(y)$, so

$$F_Y(y) = \mathbb{P}(X \leqslant h(y)) = F_X(h(y)).$$

Differentiating with respect to y gives

$$f_Y(y) = F_Y'(y) = F_X'((h(y))h'(y) = f_X(h(y))\left|h'(y)\right|$$

because $h'(y) \geqslant 0$ when g is increasing and so $|h'(y)| = h'(y)$.

On the other hand, if g is decreasing then $g(X) \leqslant g(h(y))$ if and only if $X \geqslant h(y)$, so Equation (1) gives

$$F_Y(y) = \mathbb{P}(X \geqslant h(y)) = 1 - F_X(h(y)).$$

Differentiation gives

$$f_Y(y) = F_Y'(y) = -F_X'(h(y))h'(y) = F_X'(h(y))\left|h'(y)\right|$$

because $h'(y) \leqslant 0$ when g is decreasing and so $|h'(y)| = -h'(y)$. ■

Corollary If X is a continuous random variable and $Y = aX + b$, where a and b are constants wtih $a \neq 0$, then

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right)\frac{1}{|a|}.$$

Proof If $y = g(x) = ax + b$ then $x = (y-b)/a$, so $h(y) = (y-b)/a$ and $|h'(y)| = |1/a| = 1/|a|$. Now substitute in Theorem 7. ■

Normal random variables

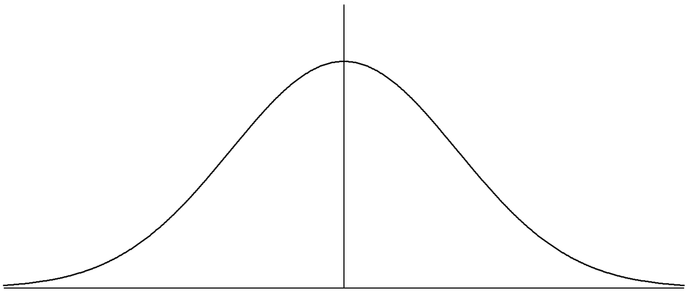
As with uniform random variables, we deal first with the simplest case.

Standard normal random variable $N(0,1)$

A continuous random variable Z is *standard normal* if its p.d.f. is

$$f_Z(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2} \qquad \text{for } x \text{ in } \mathbb{R}.$$

The picture below shows the graph of this function, the familiar 'bell-shaped curve'.



The curve is symmetrical about 0, so the expected value and median are both equal to 0. The support of Z is the whole real line.

Using techniques from Calculus that you have probably not yet met, you can show that

$$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\mathrm{d}x = \sqrt{2\pi},$$

from which it follows that

$$\int_{\infty}^{\infty} f_Z(x)\mathrm{d}x = 1,$$

as it should. See if you can use integration by parts to show that $\mathrm{Var}(Z) = 1$.

We write $Z \sim N(0,1)$: the 'N' is for normal, the '0' is the expectation and the '1' is the variance.

The c.d.f. F_Z of Z is often written Φ. That is,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-t^2/2}\,\mathrm{d}t,$$

which cannot be integrated analytically, so we have to use tables. The c.d.f. of the standard normal is given in Table 4 of the *New Cambridge Statistical Tables* [1].

Using the tables

Here is an example of how to find a probability. To find $\mathbb{P}(Z < 1.25)$, simply look up the value 1.25 in the table: we find $\Phi(1.25) = 0.8944$. So $\mathbb{P}(Z < 1.25) = 0.8944$.

A slightly more complicated example:

$$\mathbb{P}(0 < Z < 1.25) = \Phi(1.25) - \Phi(0) = 0.8944 - 0.5000 = 0.3944.$$

Some tables do not give $\Phi(x)$ for $x < 0$, so we have to work it out like this. If $x > 0$ then

$$\begin{aligned}
\Phi(-x) &= \int_{-\infty}^{-x} f_Z(t)\,\mathrm{d}t \\
&= \int_{x}^{\infty} f_Z(t)\,\mathrm{d}t, \qquad \text{because } f_Z \text{ is symmetric about zero,} \\
&= \mathbb{P}(Z \geqslant x) \\
&= 1 - \mathbb{P}(Z < x) \\
&= 1 - \Phi(x).
\end{aligned}$$

For example, $\mathbb{P}(Z < -1.25) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056$.

Interpolation Supposing that we want to know $\Phi(0.283)$? The table gives

$$\begin{aligned}
\Phi(0.28) &= 0.6103 \\
\Phi(0.29) &= 0.6141.
\end{aligned}$$

We assume that Φ is roughly linear over this short range, so that $\Phi(0.283)$ should be $3/10$ of the way from 0.6103 to 0.6141, which is

$$\begin{aligned}
0.6103 + \frac{3}{10}(0.6141 - 0.6103) &= 0.6103 + \frac{3}{10}(0.0038) \\
&= 0.6103 + 0.0011 = 0.6114.
\end{aligned}$$

Using the tables in reverse Suppose that we want to find the value of x such that $\mathbb{P}(Z \leqslant x) = 0.9750$; that is, $\Phi(x) = 0.9750$. We simply look for the value 0.9750 in the body of the table and find that $x = 1.96$.

What about finding x such that $\mathbb{P}(Z \leqslant x) = 0.9000$? We look in the body of the table to find the entries just below, and just above, 0.9000. They are

$$\begin{aligned}
\Phi(1.28) &= 0.8997 \\
\Phi(1.29) &= 0.9015.
\end{aligned}$$

Then
$$\frac{0.9000-0.8997}{0.9015-0.8997}=\frac{0.0003}{0.0018}=\frac{1}{6},$$
so x is about $1/6$ of the way from 1.28 to 1.29, so $x \approx 1.2817$.

Using the *New Cambridge Statistical Tables*, we can verify this result by using Table 5. $\mathbb{P}(Z \leqslant x) = 0.90 \Rightarrow \mathbb{P}(Z \geqslant x) = 0.10$, from which Table 5 gives $x = 1.2816$.

Take care when using Table 5: the shaded portion is on the right, not the left, and the probabilities have been multiplied by 100. Also, this table has too few values in it for it to be used directly for interpolation.

Some important values from the tables are as follows. For a standard normal random variable, approximately

$$\begin{array}{ll}68\% & \text{of all the values lie within } [-1,1]\\ 95\% & \text{of all the values lie within } [-2,2]\\ 99\tfrac{3}{4}\% & \text{of all the values lie within } [-3,3].\end{array}$$

In other words, if $Z \sim N(0,1)$ then $\mathbb{P}(-1 \leqslant Z \leqslant 1) = 0.68$, $\mathbb{P}(-2 \leqslant Z \leqslant 2) = 0.95$ and $\mathbb{P}(-3 \leqslant Z \leqslant 3) = 0.9973$.

General normal random variables

A random variable X is *normal* if it is given by $X = aZ + b$ where Z is a standard normal random variable and a and b are constants with $a \neq 0$.

Proposition Let $X = aZ + b$, where $Z \sim N(0,1)$ and $a \neq 0$. Then

(i) $\mathbb{E}(X) = b$;

(ii) $\operatorname{Var}(X) = a^2$;

(iii) $f_X(x) = \dfrac{1}{|a|\sqrt{2\pi}}\,\mathrm{e}^{-(x-b)^2/2a^2}$.

Proof (i) $\mathbb{E}(X) = \mathbb{E}(aZ+b) = a\mathbb{E}(Z) + b = b$.

(ii) $\operatorname{Var}(X) = \operatorname{Var}(aZ+b) = \operatorname{Var}(aZ) = a^2\operatorname{Var}(Z) = a^2$.

(iii) The Corollary to Theorem 7 gives
$$f_X(x) = f_Z\left(\frac{x-b}{a}\right) \times \frac{1}{|a|} = \frac{1}{|a|\sqrt{2\pi}}\,\mathrm{e}^{-(x-b)^2/2a^2}. \quad \blacksquare$$

The p.d.f. is symmetrical about b, so the expected value and median are both equal to b. The support of X is the whole real line.

We write $X \sim N(b, a^2)$.

Since it is not possible to write the integral of this p.d.f. in a nice form, we need to convert general normal random variables into the standard normal random variable. If $X \sim N(b, a^2)$, and $Z = (X - b)/a$, then $Z \sim N(0,1)$. So we only need tables of the c.d.f. Φ for the standard normal random variable $N(0,1)$.

Here are some examples.

(a) If $X \sim N(0, 100)$ then $\dfrac{X}{10} \sim N(0,1)$, so

$$\begin{aligned} \mathbb{P}(X \leqslant 12.5) &= \mathbb{P}\left(\frac{X}{10} \leqslant 1.25\right) \\ &= \mathbb{P}\left(Z \leqslant 1.25\right), \qquad \text{where } Z \sim N(0,1), \\ &= \Phi(1.25) \\ &= 0.8944. \end{aligned}$$

(b) If $X \sim N(5,1)$ then $X - 5 \sim N(0,1)$, so

$$\begin{aligned} \mathbb{P}(X \leqslant 8) &= \mathbb{P}(X - 5 \leqslant 3) \\ &= \mathbb{P}\left(Z \leqslant 3\right), \qquad \text{where } Z \sim N(0,1), \\ &= \Phi(3) \\ &= 0.9987. \end{aligned}$$

(c) If $X \sim N(70, 25)$ then $\dfrac{X - 70}{5} \sim N(0,1)$, so

$$\begin{aligned} \mathbb{P}(X \leqslant 80) &= \mathbb{P}\left(\frac{X - 70}{5} \leqslant \frac{80 - 70}{5}\right) \\ &= \mathbb{P}\left(\frac{X - 70}{5} \leqslant 2\right) \\ &= \mathbb{P}\left(Z \leqslant 2\right), \qquad \text{where } Z \sim N(0,1), \\ &= \Phi(2) \\ &= 0.9772. \end{aligned}$$

[1] D. V. Lindley and W. F. Scott, *New Cambridge Statistical Tables*, Cambridge University Press.