Probability & Statistics

Dongjiao Ge

djge@cityu.edu.mo

Session 06

3.6 Continuous random variables

continuous random variable: Its target set is the set of real numbers, or perhaps the non-negative real numbers or just an interval.

- The crucial property is that, for any real number a, we have $P(X = a) = 0$;
  (the probability that the height of a random student, or the time I have to wait for a bus, is precisely a, is zero.)

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. The c.d.f. of the random variable X is the function defined by F_X

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

Note for the cdf:

- The name of the function is F_X; the lower case x refers to the argument of the function, the number which is substituted into the function. Note that $F_X(y)$ is the same function written in terms of the variable y instead of x, whereas $F_Y\ (x)$ is the c.d.f. of the random variable Y (which might be a completely different function.)
- 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 $P(X\ <\ x)$.

Proposition 3.5 *The c.d.f. is an increasing function (this means that* $F_X(x) \leq F_X(y)$ *if* $x < y$*), and approaches the limits* 0 *as* $x \to -\infty$ *and* 1 *as* $x \to \infty$.

The function is increasing because, if $x < y$, then

$$F_X(y) - F_X(x) = P(X \leq y) - P(X \leq x) = P(x < X \leq y) \geq 0.$$

Also $F_X(\infty) = 1$ because X must certainly take some finite value; and $F_X(-\infty) = 0$ because no value is smaller than $-\infty$.

Another important function is the **probability density function** f_X. It is obtained by **differentiating the c.d.f**.:

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

- $f_X(x)$ 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

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

- We also have

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

Most facts about continuous random variables are obtained by replacing the p.m.f. by the p.d.f. and replacing sums by integrals. Thus, the *expected value* and *variance* of X are given by

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

$$\mathrm{Var}(X) = E(X^2) - E(X)^2,$$

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

It is also true that $Var(X) = E((X - \mu)^2)$, where $\mu = E(X)$.

Support: The ***support*** of a continuous random variable is the smallest interval containing all values of x where $f_X(x) > 0$.

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

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

The support of X is the interval [0,1].

We check the integral:

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

The cumulative distribution function of X is

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

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x = \int_0^1 2x^2\mathrm{d}x = \frac{2}{3}, \\ E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)\mathrm{d}x = \int_0^1 2x^3\mathrm{d}x = \frac{1}{2}, \\ \mathrm{Var}(X) &= \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}. \end{aligned}$$

3.7 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 $E(X) = 2/3$. The median of X is the value of m for which $F_X(m) = 1/2$. Since $F_X(x) = x^2$ for $0 \leq x \leq 1$, we see that $m = 1/\sqrt{2}$.

- **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

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

- 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 $n\%$.

Reminder If the c.d.f. of X is $F_X(x)$ and the p.d.f. is $f_X(x)$, then

- differentiate F_X to get f_X, and integrate f_X to get F_X;
- use f_X to calculate $E(X)$ and $\mathrm{Var}(X)$;
- use F_X to calculate $P(a \leq X \leq b)$ (this is $F_X(b) - F_X(a)$), and the median and percentiles of X.

3.8 Some continuous random variables

Uniform random variable $U(a, b)$

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 a 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 $c = 1/(b - a)$.

So, the p.d.f. of the random variable $X \sim U(a, b)$ is given by

$$f_X(x) = \begin{cases} 1/(b-a) & \text{if } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

By integration, the c.d.f. is

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

The expected value and the median of X are both given by $(a + b)/2$ (the midpoint of the interval), while $Var(X) = (b - a)^2/12$.

Exponential random variable $Exp(\lambda)$

- **The exponential random variable arises in the same situation as the Poisson**
  We have events which occur randomly but at a constant average rate of λ per unit time (e.g. radioactive decays, 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. Not that it takes non-negative real numbers as values.

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

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

By integration, we find the c.d.f. to be

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

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

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

Normal random variable $N(\mu, \sigma^2)$

The normal random variable is the commonest of all in applications, and the most important.

There is a theorem called the ***central limit theorem (we will have a look at it later on)*** which says that, for virtually any random variable X which is not too bizarre, if you take the sum (or the average) of n independent random variables with the same distribution as X, the result will be approximately normal, and will become more and more like a normal variable as n grows.

------*This partly explains why a random variable affected by many independent factors, like a man's height, has an approximately normal distribution.*

If n is large, then a $Bin(n,p)$ random variable is well approximated by a normal random variable with the same expected value np and the same variance npq.

Theorem * De Moivre-Laplace Central Limit Theorem.

> *Let S_n be Binomial(n,p), where p is fixed and n is large. Then, $\frac{S_n - np}{\sqrt{np(1-p)}} \approx N(0,1)$; more precisely,*
>
> $$P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) \to \Phi(x)$$
>
> *as $n \to \infty$, for every real number x.*

We should also note that the above theorem is an analytical statement; it says that

$$\sum_{k:0\leq k\leq np+x\sqrt{np(1-p)}} \binom{n}{k} p^k (1-p)^{n-k} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{s^2}{2}}\, ds$$

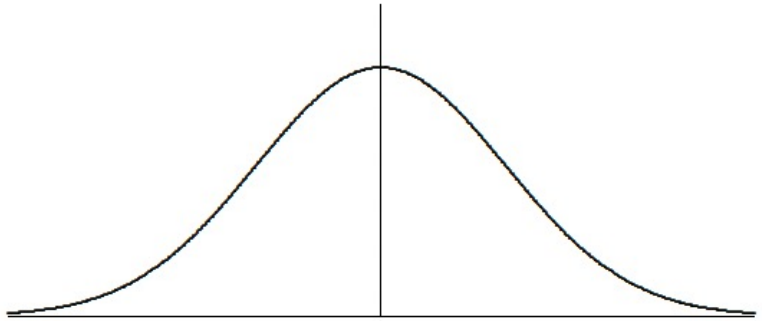
as $n \to \infty$, for every $x \in \mathbb{R}$. Indeed it can be, and originally was, proved this way, with a lot of computational work.

(If you are approximating any discrete random variable by a continuous one, you should make a "*continuity correction*" see session 3.9.)

The p.d.f. of the random variable $X \sim N(\mu, \sigma^2)$ is given by the formula

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \, e^{-(x-\mu)^2/2\sigma^2}.$$

We have $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. The picture below shows the graph of this function for $\mu = 0$, the familiar 'bell-shaped curve'.



The c.d.f. of X is obtained as usual by integrating the p.d.f. *However, it is not possible to write the integral of this function (which, stripped of its constants, is e^{-x^2}) in terms of 'standard' functions*. So there is no alternative but to make tables of its values.

Proposition 3.6 *If* $X \sim N(\mu, \sigma^2)$, *and* $Y = (X - \mu)/\sigma$, *then* $Y \sim N(0, 1)$.

So, we only need tables of the c.d.f. for N(0,1) – this is the so-called standard normal random variable – and we can find the c.d.f. of any normal random variable. We don't have to tabulate the function for all values of μ and σ.

For example, suppose that $X \sim N(6,25)$.What is the probability that $X \leq 8$? Putting $Y = (X-6)/5$, so that $Y \sim N(0,1)$, we find that $X \leq 8$ if and only if $Y \leq (8-6)/5 = 0.4$. From the tables, the probability of this is $\Phi(0.4) = 0.6554$.

The p.d.f. of a standard normal r.v. Y is symmetric about zero. This means that, for any positive number c,

$$\Phi(-c) = P(Y \leq -c) = P(Y \geq c) = 1 - P(Y \leq c) = 1 - \Phi(c).$$

So it is only necessary to tabulate the function for positive values of its argument.

So, if $X \sim N(6,25)$ and $Y = (X-6)/5$ as before, then

$$P(X \leq 3) = P(Y \leq -0.6) = 1 - P(Y \leq 0.6) = 1 - 0.7257 = 0.2743.$$

3.9 On using tables

Interpolation

Any table is limited in the number of entries it contains. Tabulating something with the input given to one extra decimal place would make the table ten times as bulky!

Interpolation can be used to extend the range of values tabulated.

Suppose that some function F is tabulated with the input given to three places of decimals. It is probably true that F is changing at a roughly constant rate between, say, 0.28 and 0.29. So $F(0.283)$ will be about three-tenths of the way between $F(0.28)$ and $F(0.29)$.

For example, if Φ is the c.d.f. of the normal distribution, then $\Phi(0.28) = 0.6103$ and $\Phi(0.29) = 0.6141$, so $\Phi(0.283) = 0.6114$. (Three-tenths of 0.0038 is 0.0011.)

Using tables in reverse

This means, if you have a table of values of F, use it to find x such that $F(x)$ is a given value c. Usually, c won't be in the table and we have to interpolate between values x_1 and x_2, where $F(x_1)$ is just less than c and $F(x_2)$ is just greater.

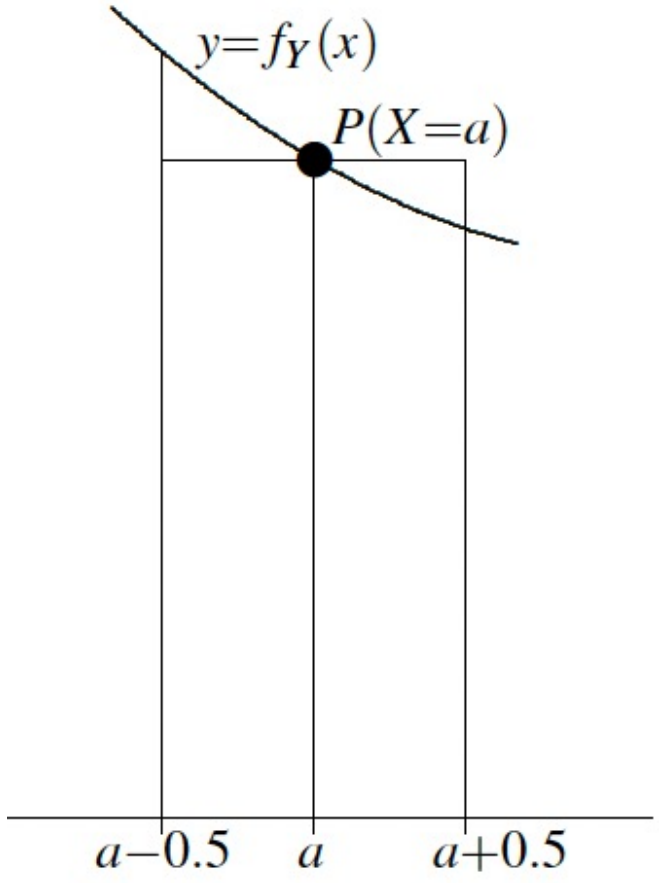
For example, if Φ is the c.d.f. of the normal distribution, and we want the upper quartile, then we find from tables $\Phi(0.67) = 0.7486$ and $\Phi(0.68) = 0.7517$, so the required value is about 0.6745 (since 0.0014/0.0031 = 0.45).

Continuity correction

Suppose we know that a **discrete random variable** X is well **approximated by a continuous random variable** Y. We are given a table of the c.d.f. of Y and want to find information about X.

For example, suppose that X takes integer values and we want to find $P(a \leq X \leq b)$, where a and b are integers. This probability is equal to

$$P(X = a) + P(x = a + 1) + \cdots + P(X = b).$$



To say that X can be approximated by Y means that, for example, $P(X = a)$ is approximately equal to $f_Y(a)$, where f_Y is the p.d.f. of Y. This is equal to the area of a rectangle of height $f_Y(a)$ and base 1 (from $a - 0.5$ to $a + 0.5$).

This in turn is, to a good approximation, the area under the curve $y = f_Y(x)$ from $x = a - 0.5$ to $x = a + 0.5$, since the pieces of the curve above and below the rectangle on either side of $x = a$ will approximately cancel. Similarly for the other values.

Adding all these pieces. we find that $P(a \leq X \leq b)$ is approximately equal to the area under the curve $y = f_Y(x)$ from $x = a - 0.5$ to $x = b + 0.5$. This area is given by $F_Y(b + 0.5) - F_Y(a - 0.5)$, since F_Y is the integral of f_Y. Said otherwise, this is $P(a - 0.5 \leq Y \leq b + 0.5)$.

continuity correction:

Suppose that the discrete random variable X, taking integer values, is approximated by the continuous random variable Y. Then

$$P(a \leq X \leq b) \approx P(a-0.5 \leq Y \leq b+0.5) = F_Y(b+0.5) - F_Y(a-0.5).$$

(Here, $\approx$ means "approximately equal".) Similarly, for example, $P(X \leq b) \approx P(Y \leq b+0.5)$, and $P(X \geq a) \approx P(Y \geq a-0.5)$.

Example The probability that a light bulb will fail in a year is 0.75, and light bulbs fail independently. If 192 bulbs are installed, what is the probability that the number which fail in a year lies between 140 and 150 inclusive?

Solution Let X be the number of light bulbs which fail in a year. Then $X \sim \text{Bin}(192, 3/4)$, and so $E(X) = 144$, $\text{Var}(X) = 36$. So X is approximated by $Y \sim N(144, 36)$, and

$$P(140 \leq X \leq 150) \approx P(139.5 \leq Y \leq 150.5)$$

by the continuity correction.

Let $Z = (Y - 144)/6$. Then $Z \sim N(0, 1)$, and

$$\begin{aligned} P(139.5 \leq Y \leq 150.5) &= P\left(\frac{139.5 - 144}{6} \leq Z \leq \frac{150.5 - 144}{6}\right) \\ &= P(-0.75 \leq Z \leq 1.083) \\ &= 0.8606 - 0.2268 \qquad \text{(from tables)} \\ &= 0.6338. \end{aligned}$$

Exercises

EX1 An archer shoots an arrow at a target. The distance of the arrow from the centre of the target is a random variable X whose p.d.f. is given by

$$f_X(x) = \begin{cases} (3+2x-x^2)/9 & \text{if } x \leq 3, \\ 0 & \text{if } x > 3. \end{cases}$$

The archer's score is determined as follows:

Distance	$X < 0.5$	$0.5 \leq X < 1$	$1 \leq X < 1.5$	$1.5 \leq X < 2$	$X \geq 2$
Score	10	7	4	1	0

Construct the probability mass function for the archer's score, and find the archer's expected score.

EX2 Let T be the lifetime in years of new bus engines. Suppose that T is continuous with probability density function

$$f_T(x) = \begin{cases} 0 & \text{for } x < 1 \\ \dfrac{d}{x^3} & \text{for } x > 1 \end{cases}$$

for some constant d.

(a) Find the value of d.

(b) Find the mean and median of T.

(c) Suppose that 240 new bus engines are installed at the same time, and that their lifetimes are independent. By making an appropriate approximation, find the probability that at most 10 of the engines last for 4 years or more.

EX3 Assume that a lightbulb lasts on average 100 hours. Assuming exponential distribution, compute the probability that it lasts more than 200 hours and the probability that it lasts less than 50 hours.

EX4 How many times do you need to toss a fair coin to get at least 100 heads with probability at least 90%?

EX5 The lifetime of a machine part has a continuous distribution on the interval $(0, 40)$ with probability density function f, where $f\ (x)$ is proportional to $(10\ +\ x)^{-2}$. Calculate the probability that the lifetime of the machine part is less than five.

EX6 The working lifetime, in years, of a particular model of bread maker is normally distributed with mean 10 and variance 4. Calculate the 12th percentile of the working lifetime, in years.