In this lecture, we consider continuous random variables, which are random variables which can take on an uncountably infinite number of values. To put it in a different yet equivalent way, continuous random variables are random variables such that $F(x) = P(X \leq x)$ is a **continuous function of** x. Unlike discrete random variables, F does not have jumps.

1 Continuous Random Variables

Consider the situation in which the random variable X takes on the values of the time one has to wait for the 6 subway train at the Bleecker St station.

In principle, one could think of associating a probability $P(X = x)$ for any $x \in [0, \infty)$. However, this does not make much sense, as the probability for any given x would be infinitely small, and there is in any case no means to measure time with infinite precision.

In that context, it makes much more sense to consider

$$P(a \leq X \leq b)$$

for a time interval $[a, b]$ which may be as small as can be measured with a standard watch. This motivates the following definition.

1.1 Definition

A random variable X is continuous if there exists a function $f : x \in \mathbb{R} \mapsto f(x) \in \mathbb{R}$ such that

- $$\forall x \in \mathbb{R}\ ,\ f(x \geq 0)$$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- For any $a \in \mathbb{R}$ and $b \in \mathbb{R}$ such that $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad (1)$$

f is called the **probability density** of X, or **probability density function (p.d.f.)** of X.

1.2 Interpretation

We recall from our calculus classes that for positive functions like the p.d.f. f introduced above, $\int_a^b f(x)dx$ is the surface area of the region between the graph of f and the x-axis, between the abcissae a and b, as shown in Figure 1.

Now, remember that in the spirit of Riemann sums, one can approximate the surface area in the yellow region in Figure 1 in terms of a large number of rectangles with a small base, the approximation becoming exact as one takes the limit of the base of the rectangles going to zero.

Let us focus on one of these rectangles, with base $[c - \epsilon, c + \epsilon]$ for small ϵ, as shown in Figure 2.

For ϵ small, we may write

$$P(c - \epsilon \leq X \leq c + \epsilon) \approx 2\epsilon f(c)$$

where the approximation improves as one takes $\epsilon \to 0$. This discussion provides the most direct interpration of $f(c)$ for any c: it can be viewed as a measure of how likely it is that X will be in the neighborhood of c.

One must be careful however: this is a *relative* measure, in the sense that f is always weighted with the size of the interval before it can be viewed as a probability.

f is not necessarily bounded above: it can be arbitrarily large, unlike a probability function, as long as its integral over the real line is equal to one.

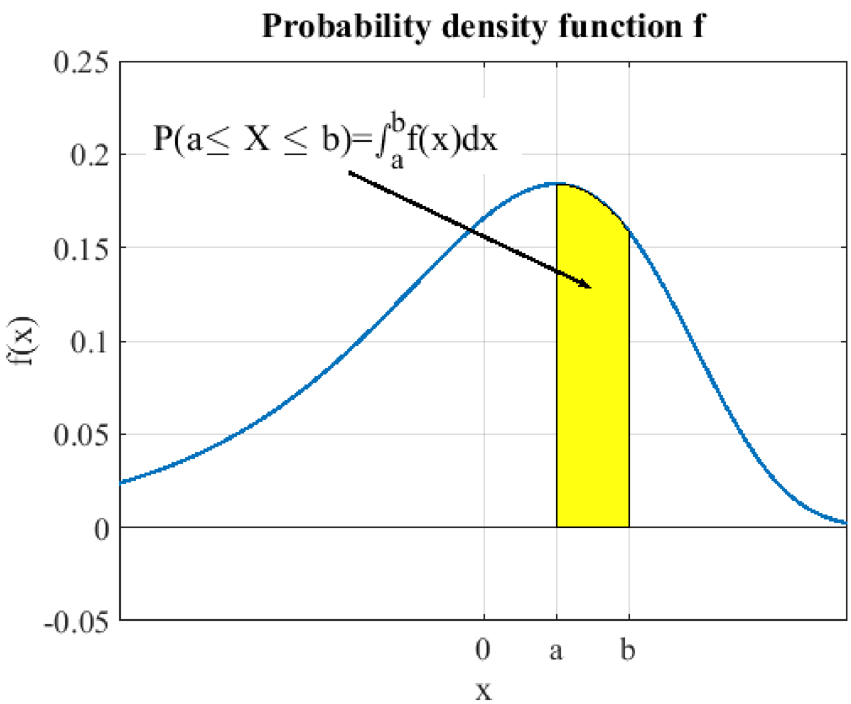

Figure 1: Interpretation of the probability $P(a \leq X \leq b)$ in terms of the area under the curve of the probability density function f between the abcissae a and b.

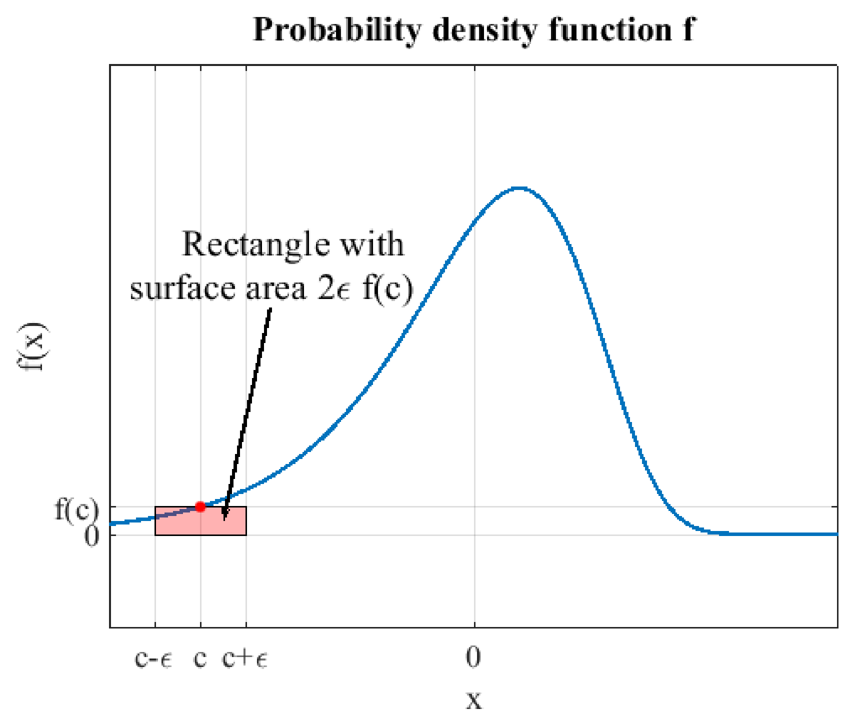

Figure 2: Small rectangle approximating the area under the curve of the probability density function f between the abcissae $c - \epsilon$ and $c + \epsilon$

Finally, note that for any $c \in \mathbb{R}$,

$$P(X = c) = \lim_{\epsilon \to 0} P(c - \epsilon \leq X \leq c + \epsilon) = \lim_{\epsilon \to 0} \int_{c-\epsilon}^{c+\epsilon} f(x)dx = 0$$

In other words,

$$P(X = c) = 0 \qquad \text{for any } c \in \mathbb{R}$$

We see the fundamental difference in the way we approach *discrete* random variables and *continuous* random variables:

- For **discrete** random variables, we think in terms of the **probability mass function (p.m.f)** defined for any $a \in \mathbb{R}$ as

$$p_X(a) = P(X = a) \qquad \text{Probability mass function (p.m.f) for discrete random variable}$$

- For **continuous random variables**, we think in terms of the **probability density function (p.d.f.)** which can be used to express the probability that **the random variable is within an interval** $[a, b]$ of $\mathbb{R}$ (as opposed to equal to a particular value):

$$P(a \leq X \leq b) = \int_a^b f(x)dx \qquad , \ f \ \text{ p.d.f.}$$

As we will now see, the two distinct approaches are reconciled when we introduce the **cumulative distribution function** already seen in Lecture 3.

1.3 Continuous random variables and distribution functions

Consider a continuous random variable X and two real numbers a and b such that $a < b$. We may write

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}$$

and the events $\{X \leq a\}$ and $\{a < X \leq b\}$ are clearly disjoint. Hence

$$P(X \leq b) = P(a < X \leq b) + P(X \leq a)$$

Introducing the cumulative distribution function (c.d.f.) as we did in Lecture 3, i.e. from its definition

$$F(x) = P(X \leq x)$$

we just proved:

$$\int_a^b f(x)dx = P(a < X \leq b) = F(b) - F(a)$$

F can be viewed as the **antiderivative** of f.

Observe that $P(a < X \leq b) = F(b) - F(a)$ holds for both discrete and continuous random variables.

Now,

$$\lim_{x \to -\infty} F(x) = 0$$

so for continuous random variables, we may write, for any real number b:

$$P(X \leq b) = F(b) = \int_{-\infty}^{b} f(x)dx \tag{2}$$

2 Important examples of continuous random variables

2.1 The uniform distribution

Definition: A continuous random variable has a **uniform distribution** on the interval $[a, b]$ if its probability density function f is given by

$$f(x) = \begin{cases} 0 \text{ if } x \notin [a,b] \\ \frac{1}{b-a} \text{ if } x \in [a,b] \end{cases} \qquad \text{Uniform distribution p.d.f.} \tag{3}$$

We usually write such a random variable $U(a, b)$, where U stands for uniform.

Let us compute the cumulative distribution function corresponding to this probability density function. For any $c \in \mathbb{R}$,

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

Now, there are three cases to consider:

1. <u>If $c < a$</u>

   Then $\forall x \in (-\infty, c]$, $f(x) = 0$. Thus

   $$F(c) = \int_{-\infty}^{c} f(x)dx = \int_{-\infty}^{c} 0dx = 0$$

2. <u>If $a \leq c \leq b$</u>

   Then in $(-\infty, c]$, f is only nonzero for $x \in [a, c]$, with value $\frac{1}{b-a}$. Thus

   $$F(c) = \int_{-\infty}^{c} f(x)dx = \int_{a}^{c} f(x)dx = \int_{a}^{c} \frac{1}{b-a}dx = \frac{c-a}{b-a}$$

3. <u>If $c > b$</u>

   Then in $(-\infty, c]$, f is only nonzero for $x \in [a, b]$, with value $\frac{1}{b-a}$. Thus

   $$F(c) = \int_{-\infty}^{c} f(x)dx = \int_{a}^{b} f(x)dx = \int_{a}^{b} \frac{1}{b-a}dx = \frac{b-a}{b-a} = 1$$

The probability density function and cumulative distribution function for $U(a, b)$ are shown in Figure 3.

2.2 The normal distribution

Definition: A continuous random variable has a **normal distribution** with parameters μ and $\sigma^2 > 0$ if its probability density function f is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{Normal distribution p.d.f.} \tag{4}$$

We usually write such a random variable $N(\mu, \sigma^2)$, where N stands for normal.

We will soon see that μ and σ^2 have a meaning: μ is the mean, and σ^2 is the variance. We plot the probability density function for $N(-2, 4)$ in Figure 4. Observe that the graph of the function is symmetric about the line $x = -2$, i.e. $x = \mu$.

Now, for any $b \in \mathbb{R}$,

$$F(b) = \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{b} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \text{Normal distribution c.d.f.} \tag{5}$$

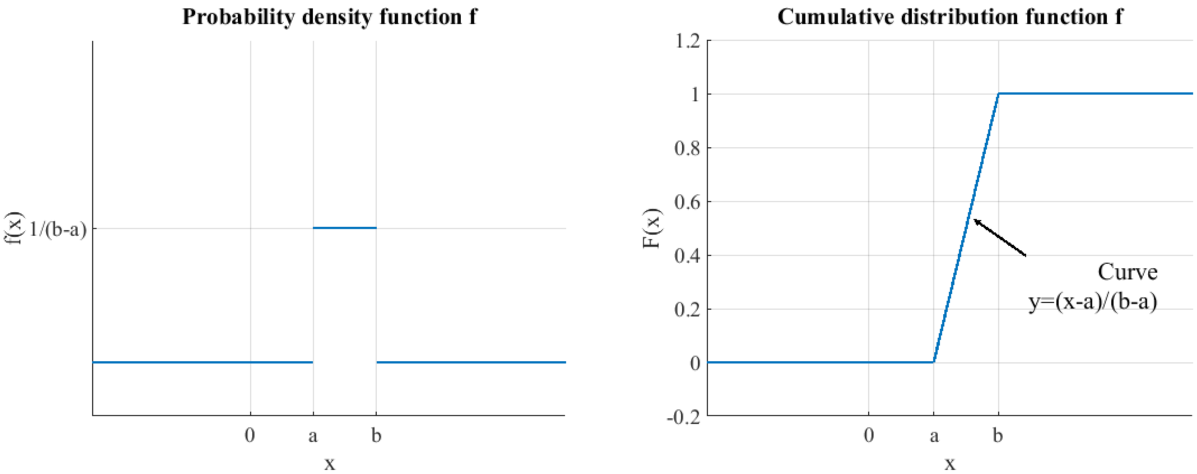

Figure 3: Probability density function and cumulative distribution function for $U(a,b)$.

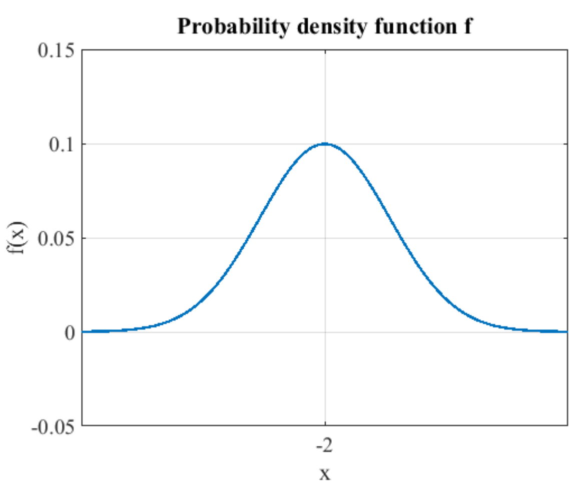

Figure 4: Probability density function for $N(-2,4)$.

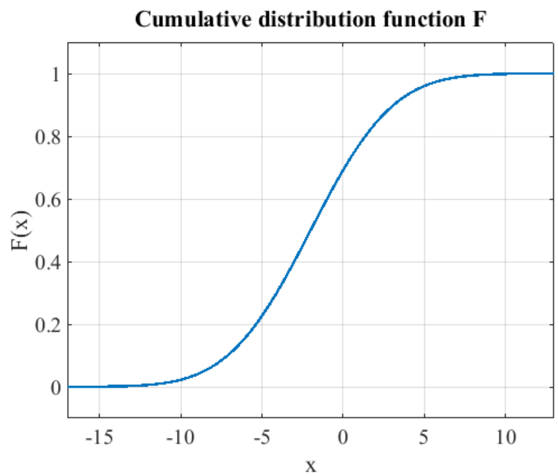

Figure 5: Cumulative distribution function for $N(-2, 4)$.

It turns out that there are no closed form expression for this integral. However, the normal distribution plays such a fundamental role in probability and statistics that $N(0, 1)$ has been tabulated: there are look-up tables in which one can read the values of the c.d.f. of $N(0, 1)$ for a range of equispaced input values.

For $N(0, 1)$, which is called the **standard normal distribution**, f is given the special letter ϕ:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The cumulative distribution function for $N(0, 1)$ is called Φ:

$$\Phi(b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-\frac{x^2}{2}} dx$$

Any cumulative distribution function for a normal random variable can be computed from the c.d.f. Φ of $N(0, 1)$. Consider the change of variable

$$u = \frac{x - \mu}{\sigma} \qquad du = \frac{dx}{\sigma}$$

in Equation (5). We obtain

$$F(b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{b-\mu}{\sigma}} e^{-\frac{u^2}{2}} du = \Phi\left(\frac{b - \mu}{\sigma}\right)$$

We will soon learn how to read these tables to be able to use normal distributions without having to resort to the numerical computation of integrals. In Figure 5, we plot the cumulative distribution function for $N(-2, 4)$.

Many situations of every day life follow the normal distribution accurately, such as the heights of people in the general population, or the scores of students on a standardized test.

Its key role in probability and statistics is due to the central limit theorem, which we will cover later in this course, and which explains why it is often a good approximation to assume that a random variable with unknown distribution is normally distributed. This is precisely why the inventor of this distribution, the famous mathematician Carl Friedrich Gauss, first used it as he tried to estimate measurement errors in astronomical observations.

Note that the normal distribution can be viewed as the limit of the binomial distribution in the limit of n large. Finally, people often colloquially call the normal distribution the "bell curve".

2.3 The log-normal distribution

Definition: A continuous random variable X has a **lognormal distribution** with parameters μ and $\sigma^2 > 0$ if $\ln(X)$ has a normal distribution with parameters μ and σ^2. In other words, $X = e^Y$, where Y has the distribution $N(\mu, \sigma^2)$.

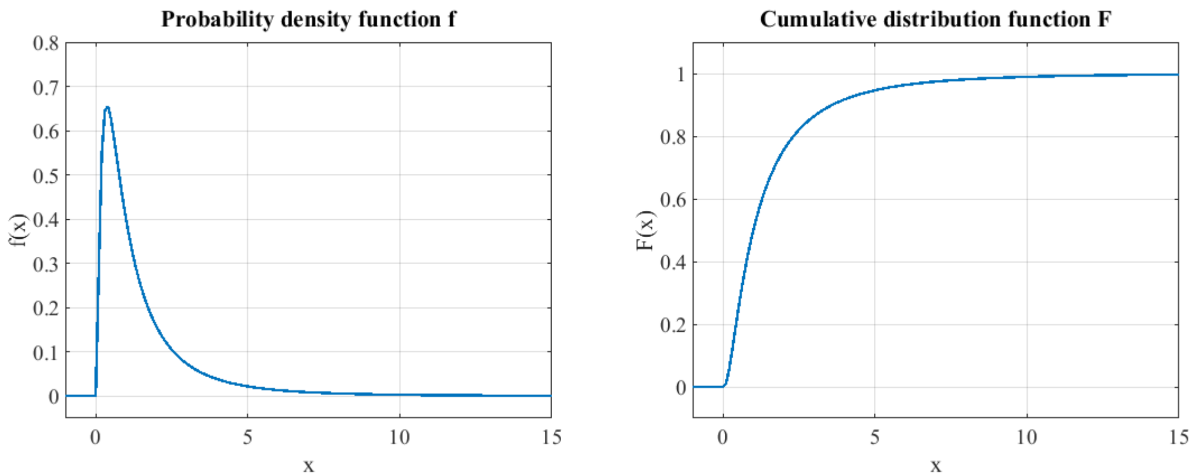

Figure 6: Probability density function and cumulative distribution function for the lognormal distribution with parameters $\mu = 0$ and $\sigma = 1$.

We can derive a closed form expression for the probability density function f using the connection between f and the c.d.f. F, as well as the expression for the probability density function for the corresponding normal distribution:

$$\begin{aligned} f(x) = \frac{d}{dx}\left(F(x)\right) &= \frac{d}{dx}\left[P(X \leq x)\right] \\ &= \frac{d}{dx}\left[P(\ln X) \leq \ln x\right] \\ &= \frac{d}{dx}\left[F_{\substack{\text{normal}\\ \text{distribution}}}(\ln x)\right] \\ &= \frac{1}{x} f_{\substack{\text{normal}\\ \text{distribution}}}(\ln x) \qquad \text{(Using the chain rule)} \end{aligned}$$

Hence

$$\underline{f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}} \qquad \text{Log-normal distribution p.d.f.}$$

Observe that by definition, a random variable which is log-normally distributed can only take *positive* real values. Consequently, it makes sense to define f for $x > 0$ only, in agreement with the presence of the term $\ln x$ in the formula.

From the derivation above, we obtain an expression for the c.d.f. F:

$$\underline{F(b) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{b} \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} dx} \qquad \text{Log-normal distribution c.d.f.}$$

As was the case for the normal distribution, there does not exist a closed form formula for this integral, so one has to rely on the use of numerical codes or tables to evaluate F. Figure 6 shows the p.d.f. and the c.d.f. for the log-normal distribution with parameters $\sigma = 1$ and $\mu = 0$.

The log-normal distribution is a good model for many random variables, such as the weight and blood pressure of people in the general population, the size of US cities, or stock prices.

To end this section, observe that there are many, many other famous and useful distributions for continuous random variables. I encourage you to look at the exponential distribution and the Pareto distribution presented in Chapter 5 of the textbook, for example.

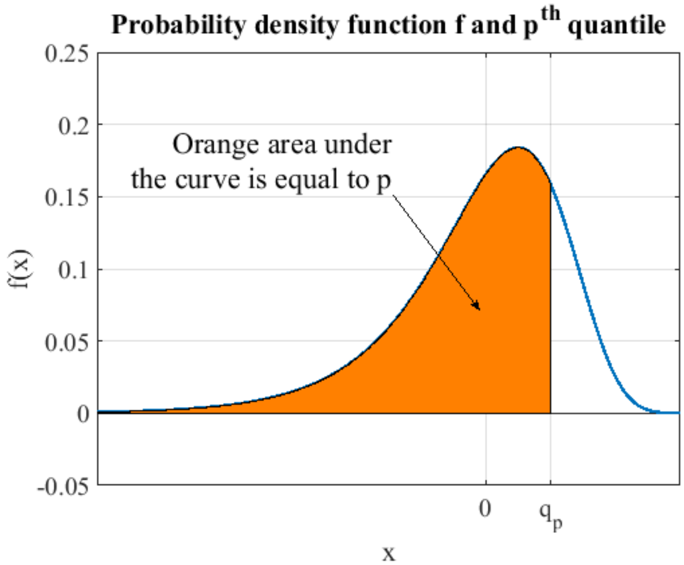

Figure 7: Graphical illustration of the definition of the p^{th} quantile q_p.

3 Quantiles

Definition: Let X be a continuous random variable, and let p be a number between 0 and 1. The p^{th} **quantile**, or $100p^{\text{th}}$ **percentile** of the distribution of X is the **smallest number** q_p such that

$$F(q_p) = P(X \leq q_p) = p \qquad \text{p}^{\text{th}} \text{ quantile} \tag{6}$$

The **median** of a distribution is its $\mathbf{50^{th}}$ **percentile**.

Graphical interpretation

From the definition of the p^{th} quantile, we see that it is the abcissa $x = q_p$ such that the area under the graph of f represents a fraction p of the total area under the graph, i.e. p. This is illustrated in Figure 7.

Example: Consider a random variable X with probability density function

$$f(x) = \begin{cases} \frac{1}{3} & \text{if } 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

What is the 25^{th} percentile of f? What is the median of f?

The 25^{th} percentile $q_{0.25}$ of f satisfies

$$\begin{aligned} \int_{-\infty}^{q_{0.25}} f(x)dx = 0.25 \quad &\Leftrightarrow \quad \int_0^{q_{0.25}} f(x)dx = 0.25 \\ &\Leftrightarrow \quad \frac{1}{3}[x]_0^{q_{0.25}} = 0.25 \\ &\Leftrightarrow \quad q_{0.25} = 0.75 \end{aligned}$$

For the median $q_{0.5}$, the same calculation leads to

$$q_{0.5} = 3 \times 0.5 = 1.5$$