Lecture 3.1

Random variables and their distribution

Random variables

Many of the random experiments we have encounter had numerical measurements associated with them, e.g., the time required for search engine to complete a search.

More generally, whether an experiment yields qualitative or quantitative outcomes, methods of statistical analysis require that we focus on certain numerical aspects of the data, e.g., sample proportion, or sample mean. In other words, to do statistical analysis, we need a way to represent the outcomes of an experiment with numerical values.

The concept of a **random variable** allows us to go from the experimental outcomes themselves to their numerical representations.

Random variables

Definition: A function X assigning a real number to every outcome $\omega \in \Omega$ is called a *random variable*.

Example: We roll a die and record the face value X. The sample space is $\Omega = \{1, 2, \ldots, 6\}$ and the random variable X is the function $X(\omega) = \omega$.

Example: We toss a coin twice and record the outcome. The sample space is $\Omega = \{HH, HT, TH, TT\}$. Counting th number of heads can be represented by the random variable Y, defined as $Y(HH) = 2$, $Y(HT) = Y(TH) = 1$, and $Y(TT) = 0$.

Random variables

We want to assign probabilities to random variables taking on particular value(s). Recall that probabilities are assigned to events. So we need to describe events in terms of random variables.

For example, recalling the experiment of rolling a die and recording the face value X, the event that the die returns an even number is

$$\{\omega \in \Omega : X(\omega) \bmod 2 = 0\} \subset \Omega.$$

Events like

$$\{\omega \in \Omega : X(\omega) \leq x\} \quad \text{and} \quad \{\omega \in \Omega : X(\omega) = x\}, \quad x \in \mathbb{R},$$

are usually abbreviated to $\{X \leq x\}$ and $\{X = x\}$, respectively.

The corresponding probabilities $\mathbb{P}(\{X \leq x\})$ and $\mathbb{P}(\{X = x\})$ are further abbreviated to $\mathbb{P}(X \leq x)$ and $\mathbb{P}(X = x)$, respectively.

Cumulative distribution function

So we can construct various events from a random variable and assign probabilities to them.

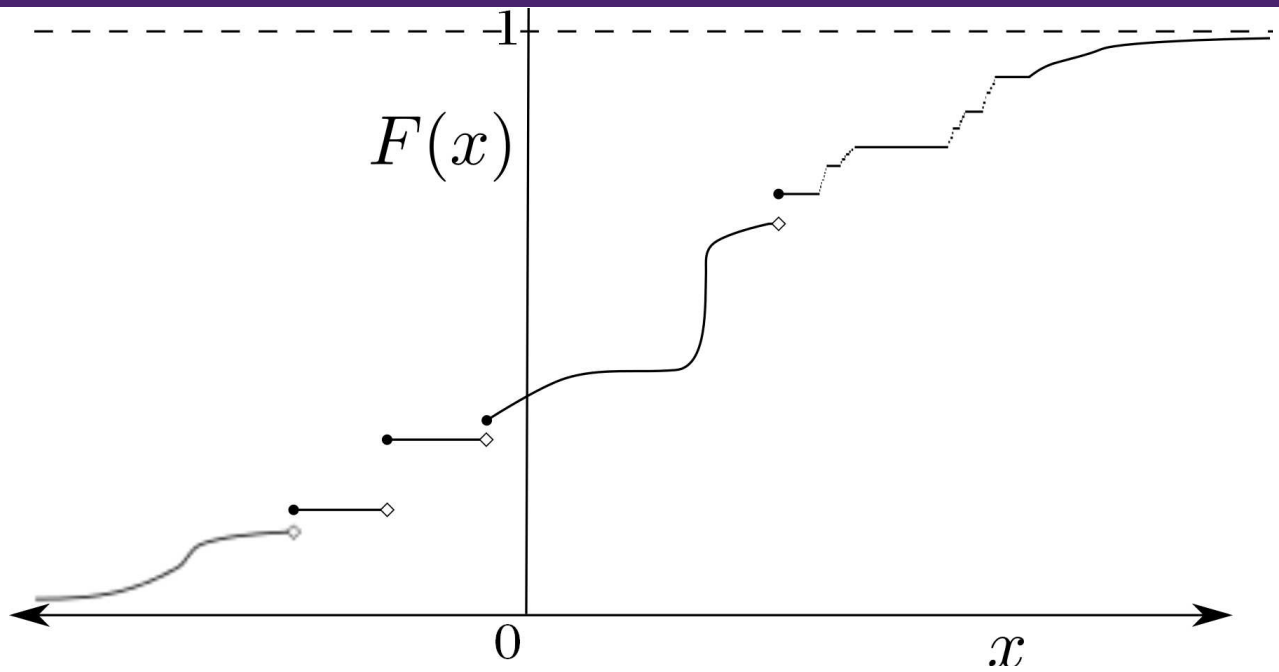
If we can specify the probability of all possible events constructed from X, we say that we have determined the **probability distribution** of X.

It turns out that it suffices to only consider the probabilities of all events of the form $\{X \leq x\}$, $x \in \mathbb{R}$.

The **cumulative distribution function (cdf)** of a random variable X is the function F defined by

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

Cumulative distribution function



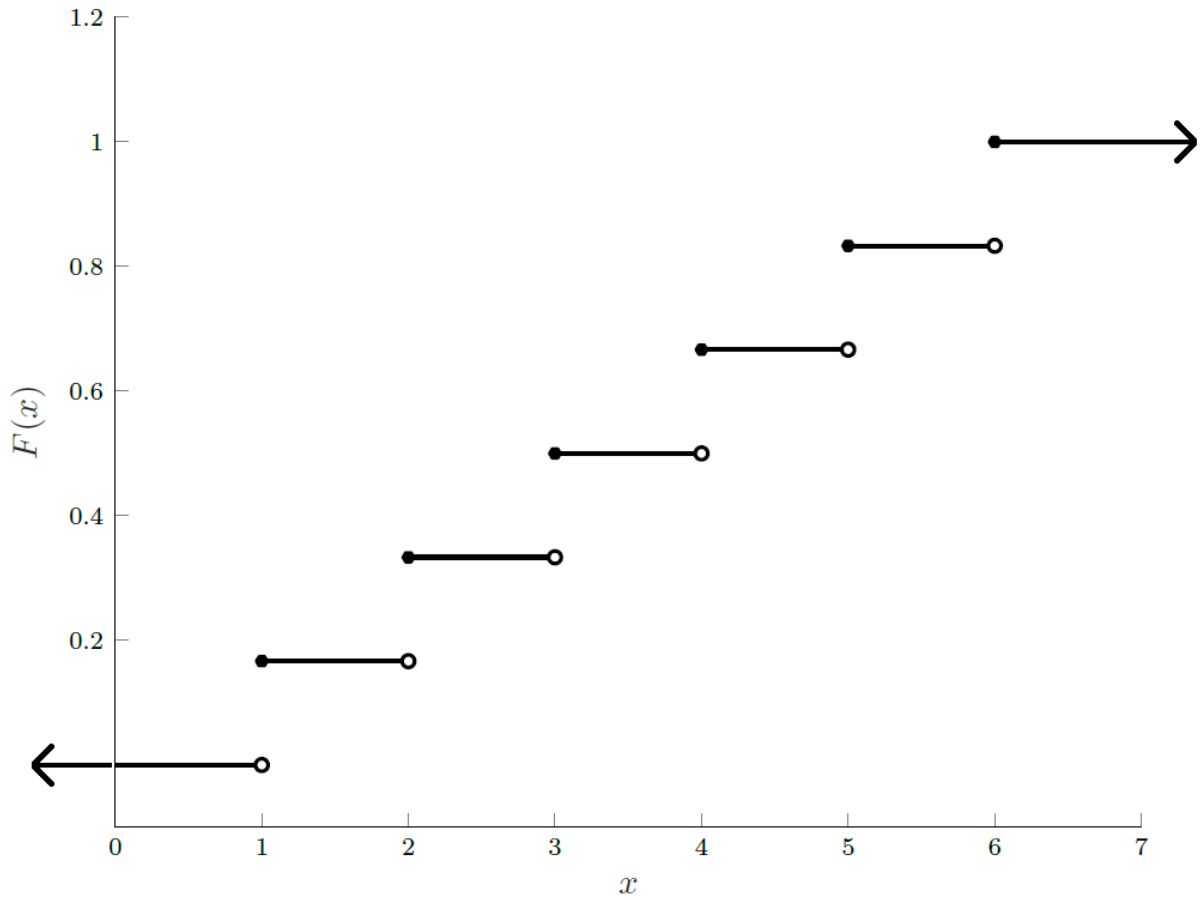
Any cdf is increasing, right-continuous[1] and lies between 0 and 1.

These properties follow from the definition of a probability measure.

[1] Roughly speaking, a function is right-continuous if no jump occurs when the limit point is approached from the right.

Example: If X is the face value of a fair, six sided die, then

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0, & x < 1 \\ \lfloor x \rfloor / 6, & x \in [1, 6) \\ 1, & x \geq 6 \end{cases}$$



Discrete and continuous random variables

Similar to our discussion of the data types in Chapter 2, we distinguish between **discrete** and **continuous** random variables.

Discrete random variables

A **discrete random variable** takes on countably many values. More precisely, X is a discrete random variable if

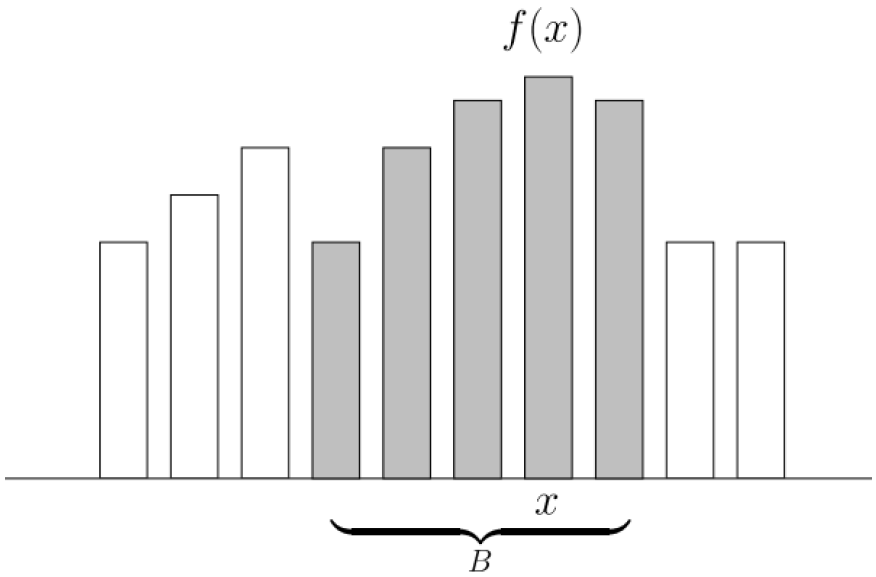
1. $X : \Omega \to \{x_1, x_2, \ldots\}$,
2. $\mathbb{P}(X = x_i) > 0$ for $i = 1, 2, \ldots$, and
3. $\sum_i \mathbb{P}(X = x_i) = 1$.

The **probability mass function (pmf)** of X is the function f defined by $f(x) = \mathbb{P}(X = x)$.

The pmf of a discrete rv says how the total probability of 1 is distributed among the various possible values of X.

Discrete random variables

Suppose X takes on values $x_1, x_2, \ldots$.



The probability that X lies in some set B is $\mathbb{P}(X \in B) = \sum_{x_i \in B} f(x_i)$.

So, cdf is easily given by the pmf as $\mathbb{P}(X \leq x) = \sum_{x_i \leq x} f(x_i)$.

Discrete random variables

Example: Roll two fair dice and let M be the largest face value showing. The pmf of M, that is $f(x) = \mathbb{P}(M = x)$, is given by

x	1	2	3	4	5	6
$f(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Continuous random variables

We saw that for a discrete random variable taking on values $x_1, x_2, \ldots$, we have $\mathbb{P}(X = x_i) > 0,\ i = 1, 2, \ldots$. This implies that the cdf of X is not continuous (in fact it would be a step function).

Indeed, we have

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(\{X = x\} \cup \{X < x\}) \\ &= \mathbb{P}(X = x) + \mathbb{P}(X < x).\end{aligned}$$

Since $\mathbb{P}(X = x) > 0$,

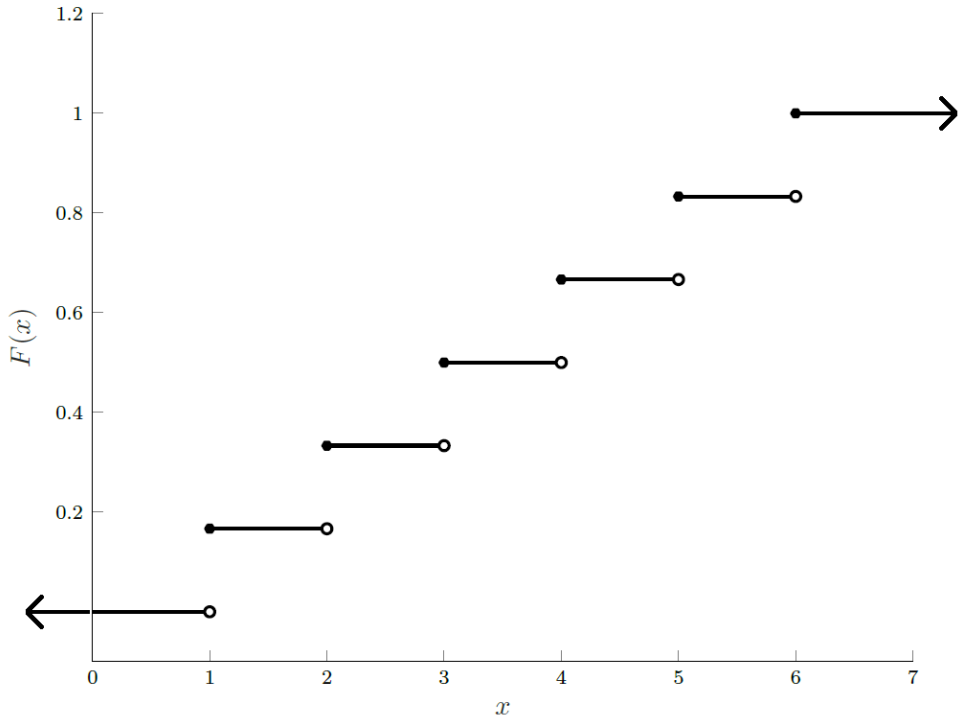
$$\mathbb{P}(X \leq x) > \mathbb{P}(X < x),$$

so,

$$F(x) > \lim_{y \uparrow x} F(y).$$

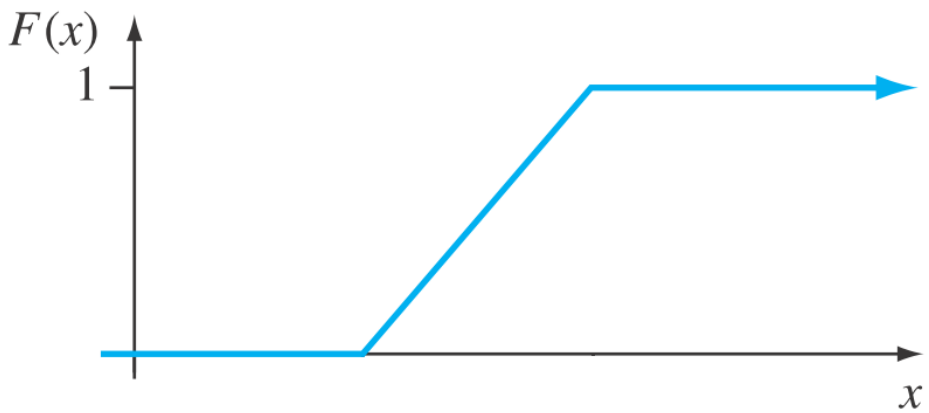
Continuous random variables

Recall: X is the face value of a fair, six sided die



Continuous random variables

We saw that the discontinuity in cdf of a dicrete random variable is due to the fact that $\mathbb{P}(X = x_i) > 0$ for $i = 1, 2, \ldots$. However, if $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$, then the X is called a **continuous random variable**. This is equivalent to its cdf being continuous since $F(x) = \lim_{y \uparrow x} F(y)$.



Of course there are random variables which are neither discrete nor continuous, i.e., their cdf is neither continuous nor a step function. But in this course we will not consider them.

Continuous random variables

Suppose a continuous random variable X takes on values in $[a, b]$ and consider its cdf F. If there exists a function $f : [a, b] \to \mathbb{R}$ such that

1. $f(x) \geq 0$ for all x, and
2. for all $a \leq c \leq d \leq b$,

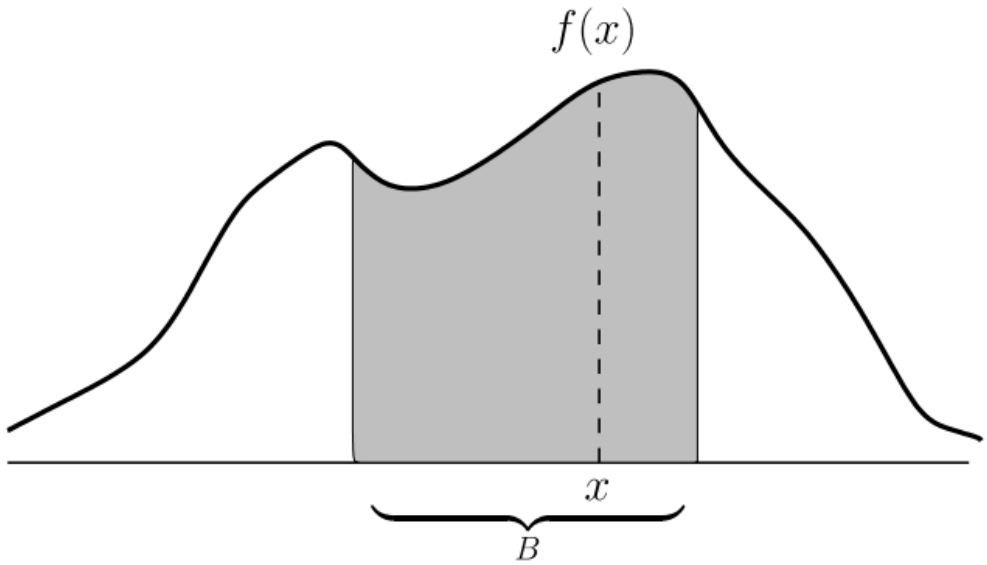
$$\mathbb{P}(c \leq X \leq d) = F(d) - F(c) = \int_c^d f(u)\, du,$$

then the function f is called the **probability density function (pdf)** of X.

At any point x where F is differentiable, we have $f(x) = F'(x)$.

Note: Not all continuous random variables have a pdf!...but most useful ones do.

Continuous random variables



In analogy to discrete random variables, we can calculate the probability that X lies in some set B as

$$\mathbb{P}(X \in B) = \int_{u \in B} f(u)du.$$

So, cdf is easily given by the pdf as $\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)du.$

Continuous random variables

Consider the function

$$f(x) = \begin{cases} c(1-x^2), & x \in (-1,1) \\ 0, & x \notin (-1,1), \end{cases}$$

where c is a constant. For what value of c is f a valid pdf?

Answer: As a pdf is non-negative at every value x, we must have $c \geq 0$. Also, the integral of a pdf over $\mathbb{R}$ must equal 1 so

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_{-1}^{1} c(1-x^2)\,dx = c(x-\tfrac{1}{3}x^3)\Big|_{-1}^{1} = c(2-\tfrac{2}{3}) = \tfrac{4}{3}c.$$

Therefore, $c = \frac{3}{4}$.

Quantile function

Recall that the p-quantile of the data $x_1, \ldots, x_n$ is the "smallest" value y that is greater than or equal to a fraction p of the data.

We now consider quantiles of a distribution.

The **quantile function** of a random variable X with cdf F is the function $Q : (0, 1) \to \mathbb{R}$ defined by

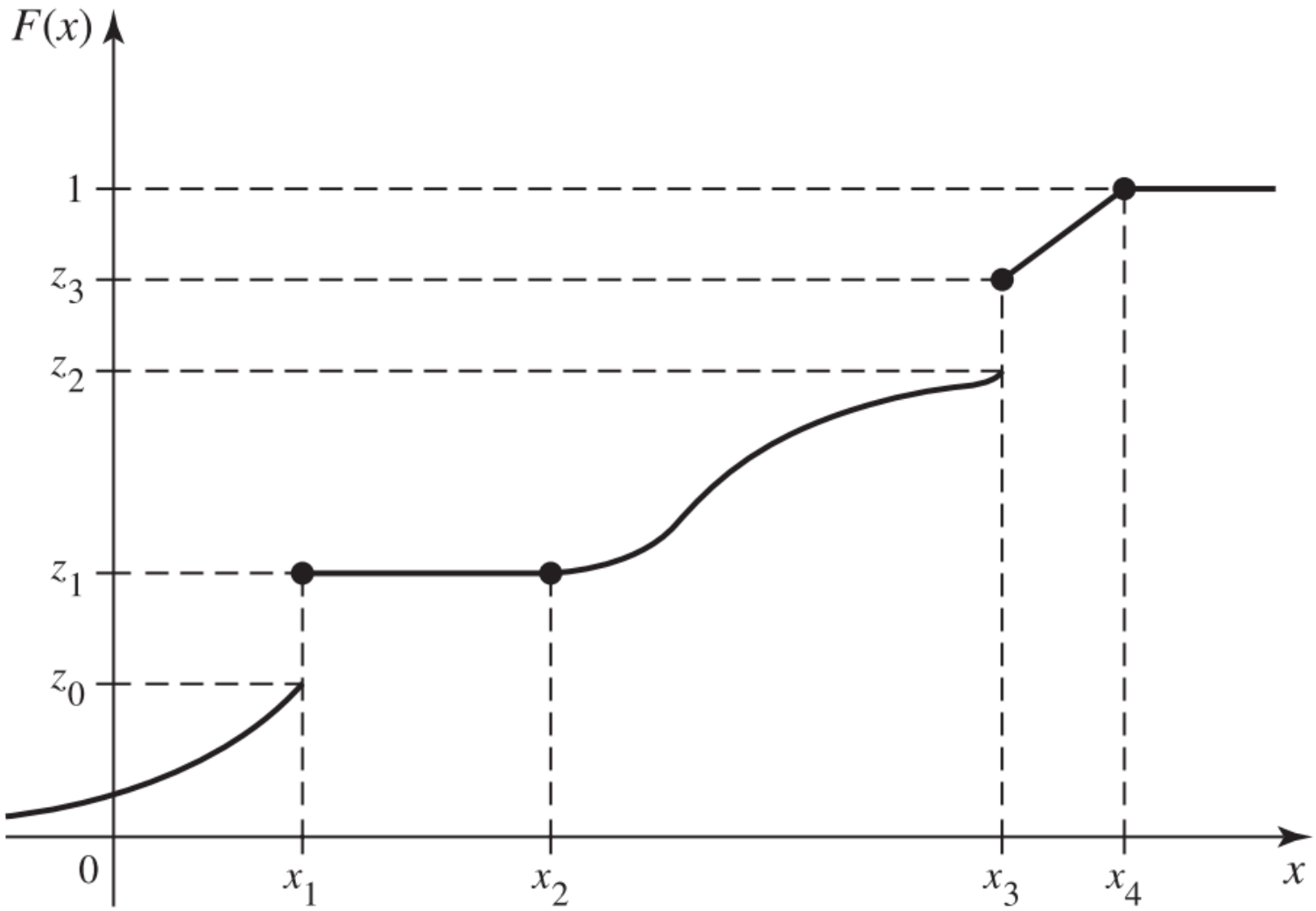
$$Q(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

In words, find all points x for which cdf $F(x)$ is larger than or equal to p, and then take the smallest of these x's.

Since distribution functions are continuous from the right, the smallest x such that $F(x) \geq p$ exists for all $0 < p < 1$.

For continuous random variables, the p-quantile of X is simply the *smallest* value x such that $F(x) = p$.

Quantile function



Quantile function

What is the quantile function of the distribution with cdf

$$F(x) = \begin{cases} 1 - e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Answer: This is a continuous distribution so we will try to find the smallest x such that $F(x) = p$

$$1 - e^{-x} = p \implies x = -\ln(1-p).$$

For every p, there is only one such x (because this cdf is strictly increasing), so $Q(p) = -\ln(1-p)$ for $p \in (0, 1)$.

This example also shows why we never talk about quantiles for $p = 0$ or $p = 1$.