STAT 515 fa 2023 Lec 6

Continuous random variables

Karl Gregory

Continuous random variables

A continuous random variable X is a random variable whose support (the set of values it can take) is an interval or a union of intervals. The idea is that we cannot list, or even begin listing all of the values X can take.

Example. Suppose I show up at the bus stop to catch the Comet and I have no knowledge of when the next bus comes; I only know that it comes once every hour. Let X be the amount of time in hours I must wait for the bus. Then

$$\mathcal{X} = [0, 1].$$

Recall that the *probability distribution* of a random variable X tells us which values X can take and assigns probabilities to the values. Since a continuous random variable X takes values on an interval, we cannot tabulate its probability distribution as we could for a discrete random variable; we have different ways of writing down the probability distribution of a continuous random variable. Before coming to these, however, we must have a philosophical discussion...

Single values get zero probability for continuous r.v.s

Probability distributions of continuous random variables assign nonzero probabilities only to intervals, not to single values. That is if X is the wait time for the bus in hours, we might have

$$\begin{aligned} P(1/6 < X < 2/6) &= 1/6 \\ P(X > 1/2) &= 1/2 \\ P(X = 1/3) &= 0, \end{aligned}$$

having the interpretation

$$P(\text{wait between 10 and 20 minutes }) = 1/6$$
$$P(\text{wait more than 30 minutes}) = 1/2$$
$$P(\text{wait exactly 20 minutes}) = 0.$$

Why is this so? To understand why this makes sense, we must first understand that 20 minutes means 20.000000000 minutes, with zeroes going on forever. Now, suppose we go take the bus and experience a wait time of 20.15983145677 minutes. How can we say that $P(X = 20.15983145677) = 0$ when we have just experienced it?! To understand this, we must imagine repeating our statistical experiment over and over again. If we wait for the bus on 9 more occasions, we will not likely experience a wait time of 20.15983145677 minutes a second time. If we wait for the bus a total of 1000 times, chances are still very small that we will experience the wait time of 20.15983145677 minutes again. Even if we wait for the bus a million times, the chances that we will observe the wait time of 20.15983145677 minutes again are very very small. So, if we were to wait for the bus again and again and again and again on into eternity, we may interpret $P(X = 20.15983145677)$ as the proportion of all of our waiting times which we expect to be equal to 20.15983145677 minutes. This proportion approaches zero as we repeat our experiment over and over again. We give the result formally as:

Result: Zero point probabilities for continuous random variables

If X is a continuous random variable, then $P(X = x) = 0$ for all x.

Probability density function

The probability distribution of a continuous random variable may be characterized by a *probability density function* (pdf). The probability density function allows us to compute the probability that a random variable lies in some interval.

Definition: Probability density function

The *probability density function* (pdf) of a continuous random variable X is the function f which satisfies

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad \text{for all } a \leq b.$$

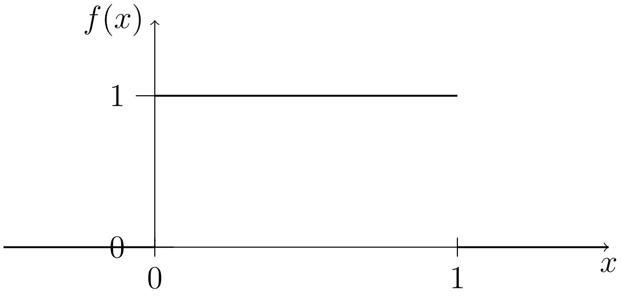
For those who have not taken calculus, the notation on the right hand side of the above is read "the integral of f from a to b", and it is equal to the area between the function f and the horizontal axis over the interval $[a, b]$. That is,

$$\int_a^b f(x)dx = \text{Area between } f \text{ and the horizontal axis on the interval } [a, b].$$

Every pdf f has the following two properties:

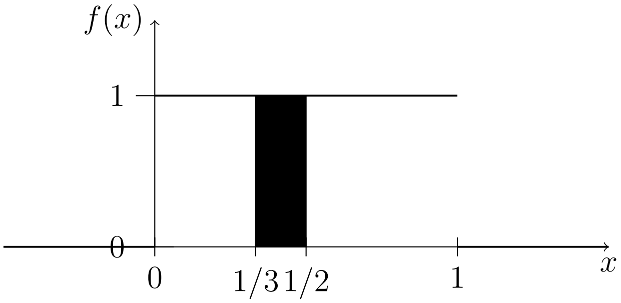
1. It is greater than or equal to zero for all x.
2. The total area between f and the horizontal axis is equal to 1. That is $\int_{-\infty}^{\infty} f(x)dx = 1$, giving $P(-\infty < X < \infty) = 1$.

Exercise. Suppose I show up at the bus stop to catch the Comet and I have no knowledge of when the next bus comes; I only know that it comes once every hour. If X is the amount of time (in hours) that I will wait for the bus, the probability density function f of the random variable X might look like



According to this probability density function, what is the probability that I must wait between 20 and 30 minutes?

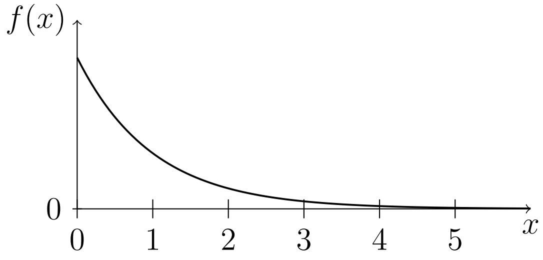
Answer: The probability that I must wait between 20 and 30 minutes is $P(1/3 \leq X \leq 1/2)$, and this is given by the area under f between 1/3 and 1/2, which we can depict as



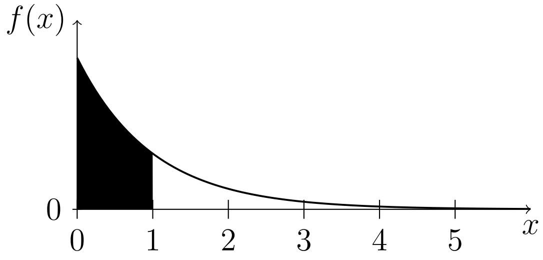
The area of the shaded region is simply $1/2 - 1/3 = 1/6$.

We can see that a probability density function assigns probability zero to single points, because the area of a line is zero!

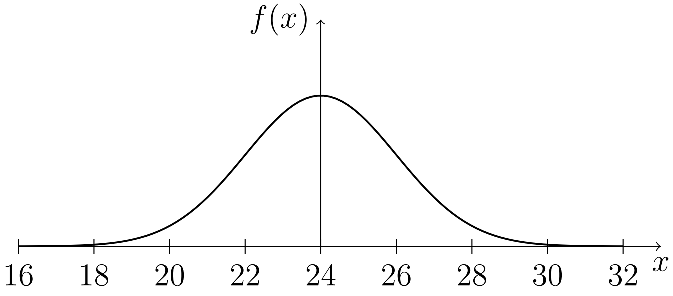
Example. Suppose I sample at random a purchaser of the latest iPhone and I follow them until their screen gets a crack in it. Let X be the amount of time (in years) that passes until the purchaser's screen gets a crack in it. The probability density function might look like



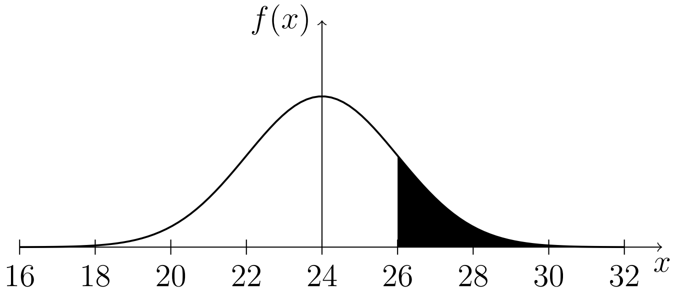
The probability that the screen gets a crack in the first year, that is $P(X \leq 1)$, would be equal to the area



Example. Let X be the mpg you get on your next tank of gas. The probability density function of X might look like



The probability that you get more than 26 miles to the gallon on your next tank, that is $P(X > 26)$, is given by the shaded area in the plot below:



Note that for any continuous random variable X, we have

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

because of the fact that $P(X = a) = P(X = b) = 0$.

Cumulative distribution function

Definition: Cumulative distribution function

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

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

The definition of the cdf applies to both continuous and discrete random variables: If X is a continuous random variable with pdf f, then the cdf of X is given by

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

in which case $F(x)$ is the area under the probability density function f over the interval $(-\infty, x]$ If X is a discrete random variable with pmf p and support $\mathcal{X}$, then the cdf of X is given by

$$F(x) = \sum_{\{t \in \mathcal{X} : t \leq x\}} p(t),$$

in which case $F(x)$ is the sum of all probabilities over the support of X which are less than or equal to x.

We will find that R has built-in functions for computing the cdf of many commonly encountered probability distributions.