3 Continuous Random Variables

3.1 Introduction

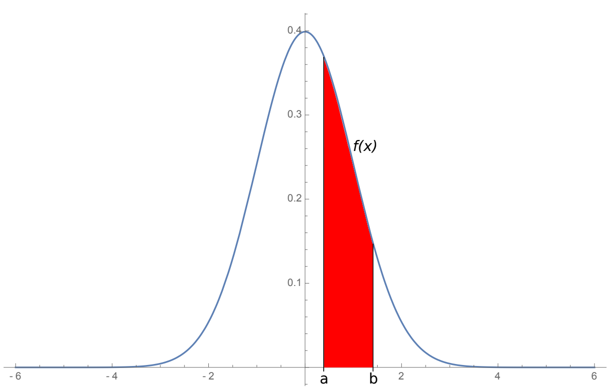
Many quantities of interest in the real world are not discrete is nature. Examples include the following:

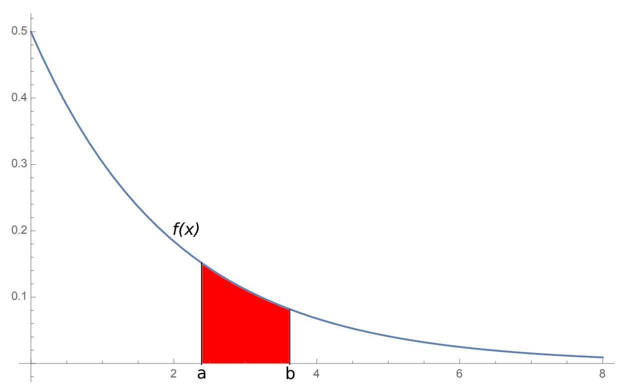
1. The amount of rainfall in one day
2. The weight of an adult chimpanzee
3. The functional lifetime of an O-ring in a jet engine

A random variable which can take on any value within a range is called a *continuous random variable*. These are fundamentally different from discrete random variables in the following way. Recall that to specify a discrete random variable, all we have to do is assign probabilities between 0 and 1 to every possible output of the random variable in such a way that all the probabilities add up to 1. This is not possible for a random variable which can take values in an interval on the real number line. We must, therefore, use a different technique to describe a continuous random variable. Here will we use calculus for (essentially) the first time in the course.

3.2 Probability Density Functions

For a continuous random variable, rather than talk about the probability that a random variable equals a particular value, we talk about the probability that a random value falls within a particular range. (In fact, as we will see, the probability that a continuous random variable equals a specific value is 0.) A continuous random variable is described by a *probability density function (pdf)*. Here are examples of pdfs:





A pdf is a nonnegative function $f(x)$ where the total area under the curve is 1 (this is analogous to the discrete case where the probabilities of all the outputs sum to 1). The probability that a random variable falls into an interval $[a, b]$ is the area under the density curve between a and b. This is illustrated in red in the above pdfs. Since we are talking about areas under curves, we need to use calculus. In particular, we need to integrate! Here is the formal definition of a pdf:

Probability Density Function (pdf)

The function $f(y)$ is a *probability density function* (pdf) for a continuous random variable Y if $f(y) \geq 0$ for all y, and

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

The probability that Y falls into the interval $[a, b]$ is the area under the density curve between a and b, i.e.

$$\mathbb{P}(a \leq Y \leq b) = \int_a^b f(y)dy$$

Before we continue, let's mention one way in which continuous and discrete distributions are very different. For a continuous distribution, the probability of any single point is 0. We can see that from the density function since for a continuous random variable Y with density y,

$$\mathbb{P}(Y = a) = \int_a^a f(y)dy = 0$$

Thus all the following probabilities are the same:

$$\mathbb{P}(a \leq Y \leq b) = \mathbb{P}(a \leq Y < b) = \mathbb{P}(a < Y \leq b) = \mathbb{P}(a < Y < b) = \int_a^b f(y)dy$$

In other words, it does not matter which inequality sign (¡ or $\leq$) we use for the endpoints of the interval since the probability of hitting the endpoints in 0. This is not at all the case for the discrete case, where individual points have positive probability. If X is the number of flips of heads in 20 tosses of a fair coin, then $\mathbb{P}(8 \leq X \leq 12)$ and $\mathbb{P}(8 < X \leq 12)$ are very different!

Example. Consider the function

$$f(y) = \begin{cases} cy^2 & 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the value of c for which $f(y)$ is a valid density function.

   First note that $f(y)$ is nonnegative, so there is nothing to worry about there. Next we need to make sure that the density function integrates to 1.

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(y)dy \\ &= \int_0^2 cy^2 dy && \text{since } f(y) \text{ is 0 outside } [0, 2] \\ &= c\frac{y^3}{3}\Big|_0^2 \\ &= \frac{8}{3}c \end{aligned}$$

   Thus, choosing $c = 3/8$, we have a valid density function.

2. What is $P(1 \le Y \le 2)$?

Integrating the density function from 1 to 2, we get:

$$\begin{aligned}
\mathbb{P}(1 \le Y \le 2) &= \int_1^2 f(y)dy \\
&= \int_1^2 \frac{3}{8}y^2 dy \\
&= \frac{3}{8}\frac{y^3}{3}\Big|_1^2 \\
&= \frac{7}{8}
\end{aligned}$$

3.3 Cumulative Distribution Functions

Another way to describe a continuous random variable is with its *cumulative distribution function* (cdf)[1].

Cumulative Distribution Function (cdf)

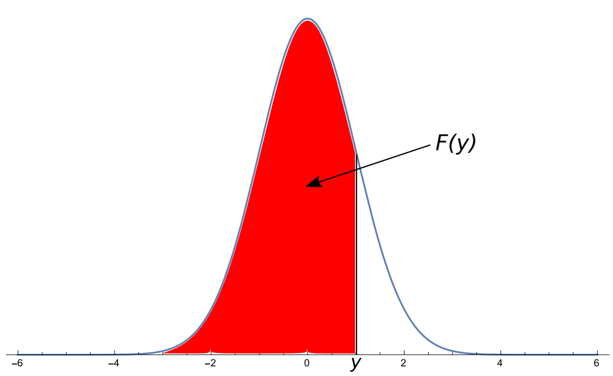
Let Y be a random variable. The the *cumulative distribution function* cdf of Y, denoted $F(y)$, is defined by

$$F(y) = \mathbb{P}(Y \le y)$$

If Y has density function $f(y)$, then

$$F(y) = \int_{-\infty}^{y} f(y)dy$$

The cdf gives the probability that our random variable Y is less than or equal to a certain value. For a continuous random variable, $F(y)$ can be visualized graphically as the area under the density curve to the left of y.



[1]Sometimes you will see this called simply a *distribution function*. I will always use the term cdf.

Note that traditionally the cdf is written with an uppercase F, while the density is written with a lowercase f. The cdf is defined for discrete as well as continuous random variables, although we will never use it in the discrete case[2]. The cdf has the following properties.

Properties of cdfs

Let Y be a random variable with cdf $F(y)$. Then:

1. $F(y)$ is a nondecreasing function.
2. $\lim_{y \to -\infty} F(y) = 0$
3. $\lim_{y \to \infty} F(y) = 1$
4. For a continuous random variable Y, the cdf $F(y)$ is a continuous function.

For a continuous random variable, the cdf and pdf are related via the fundamental theorem of calculus.

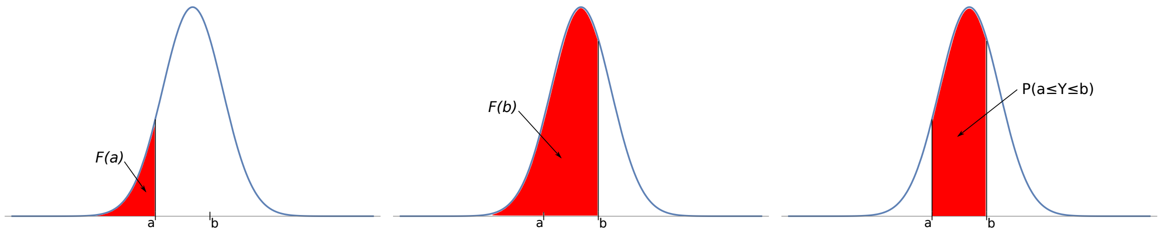
Relationship between cdf and pdf

Let Y be a continuous random variable with cdf $F(y)$ and density $f(y)$. Then we have the following relationships:

1. $$F(y) = \int_{\infty}^{y} f(y)dy$$
2. $$f(y) = \frac{DF(y)}{dy} = F'(y)$$
3. $$\mathbb{P}(a \leq Y \leq b) = \int_{a}^{b} f(y)dy = F(b) - F(a)$$

We can illustrate the third relationship above, $\mathbb{P}(a \leq Y \leq b) = F(b) - F(a)$, using the graphs below. You can see that subtracting the first area from the second area yields the third area.

[2] The cdf for a discrete random variable is always a step function, since the cdf only increases on the finite or countable set of points which have positive probabilities.



You may ask why we bother at all with cdfs since it seems easier to work with densities. There are three good reasons to consider CDFs. From a mathematical standpoint, the cdf is a more fundamental object than the pdf; every continuous random variable has a cdf, while there are continuous random variables which have no density[3]. From a practical standpoint, the normal distribution, perhaps the most important distribution is all of probability, has a density which is unwieldy, thus we will use the cdf to compute probabilities involving that distribution. Finally, the cdf is used to define the *median* and *quartiles* of a probability distribution, which are important descriptors of a probability distribution.

3.4 Median and Quartiles

The *median* is the “middle” of a probability distribution. It is the value which separates the upper half of the distribution from the lower half of the distribution. The median is more robust to outlier values than the mean, and so in some cases may be a better descriptor of a typical outcome than the mean. Mathematically, if Y is a random variable, the we can define the median of Y as the value m such that $\mathbb{P}(Y \leq m) = \mathbb{P}(Y \geq m) = 1/2$. For a continuous random variable Y with cdf $F(y)$, we can define the median using the cdf as the value m for which $F(m) = 1/2$. Similarly we can define the 1st and 3rd quartiles (the median is sometimes called the 2nd quartile).

Median and quartiles

Let Y be a continuous random variable with cdf $F(y)$. Then we define the *median* m, *first quartile* Q_1, and *third quartile* Q_3 by the following relationships:

$$\begin{aligned} F(Q_1) &= \mathbb{P}(Y \leq Q_1) = 1/4 \\ F(m) &= \mathbb{P}(Y \leq m) = 1/2 \\ F(Q_3) &= \mathbb{P}(Y \leq Q_3) = 1/4 \end{aligned}$$

Let’s revisit our example from the previous section.

Example. Let Y be a continuous random variable with density $f(y)$ defined by

$$f(y) = \begin{cases} \frac{3}{8}y^2 & 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

[3]All continuous random variables we will encounter in this course will have a density

Find the median of Y.

Let m be the median of Y. Since $F(m) = 1/2$, where $F(y)$ is the cdf of Y, we need to first find the cdf. For $y < 0$, $F(y) = 0$ and for $y > 2$, $F(y) = 1$ (do you see why this is the case?) For $0 \leq y \leq 2$, which is the region we care about, we integrate the density to get the cdf.

$$\begin{aligned} F(y) &= \int_0^y \frac{3}{8}t^2 dt \\ &= \frac{3}{8}\frac{t^3}{3}\Big|_0^y \\ &= \frac{y^3}{8} \end{aligned}$$

To find the median, we solve $F(m) = 1/2$ for m.

$$\begin{aligned} \frac{m^3}{8} &= \frac{1}{2} \\ m^3 &= 4 \\ m &= 2^{2/3} \approx 1.59 \end{aligned}$$

3.5 Expectation and Variance

Just as with discrete random variables, we can talk about the expected value and variance of a continuous random variable. They have the exact same interpretations as in the discrete case. Expectation and variance work almost exactly the same way in the continuous case as in the discrete case. In fact, we can use the exact same formulas, if we make two key changes:

1. The probability mass function $p(y)$ is replaced by the probability density function $f(y)$.
2. Summation is replaced by integration.

Making these changes, we have the following definitions for the expected value of a continuous random variable.

Expected value of a continuous random variable

Let Y be a continuous random variable with density $f(y)$. Then we define the expected value by

$$\mathbb{E}(Y) = \int_{\infty}^{\infty} yf(y)dy$$

If $g(y)$ is a real-valued function, then the expected value of $G(Y)$ is given by

$$\mathbb{E}[g(Y)] = \int_{\infty}^{\infty} g(y)f(y)dy$$

The variance of a continuous random variable is defined the same way as in the discrete case.

Variance of a continuous random variable

Let Y be a continuous random variable with density $f(y)$, and let $\mu = \mathbb{E}(Y)$. Then the variance of Y is defined byL

$$Var(Y) = \mathbb{E}[(Y-\mu)^2]$$

This is usually computed using the Magic Variance Formula, which holds for continuous random variables as well:

$$Var(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2$$

Let's compute the expected value and variance of the continuous random variable we used in the section on probability densities.

Example. Let Y be the continuous random variable defined by the pdf

$$f(y) = \begin{cases} \frac{3}{8}y^2 & 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value and variance of Y?

Using the formula for expected value,

$$\begin{aligned} \mathbb{E}(Y) = \int_{-\infty}^{\infty} yf(y)dy \\ &= \int_0^2 y\frac{3}{8}y^2 dy \\ &= \frac{3}{8}\int_0^2 y^3 dy \\ &= \frac{3}{8}\frac{y^4}{4}\Big|_0^2 \\ &= \frac{3}{8}\frac{16}{4} = 1.5 \end{aligned}$$

For the variance, we will compute $\mathbb{E}(Y^2)$ and use the Magic Variance Formula.

$$\begin{aligned}\mathbb{E}(Y^2) &= \int_{-\infty}^{\infty} y^2 f(y)dy \\ &= \int_0^2 y^2 \frac{3}{8} y^2 dy \\ &= \frac{3}{8}\int_0^2 y^4 dy \\ &= \frac{3}{8}\frac{y^5}{5}\Big|_0^2 \\ &= \frac{3}{8}\frac{32}{5} = 2.4\end{aligned}$$

Thus by the Magic Variance Formula we have:

$$Var(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = 2.4 - (1.5)^2 = 0.15$$

We will now look at three specific continuous distributions.

3.6 Continuous Uniform Distribution

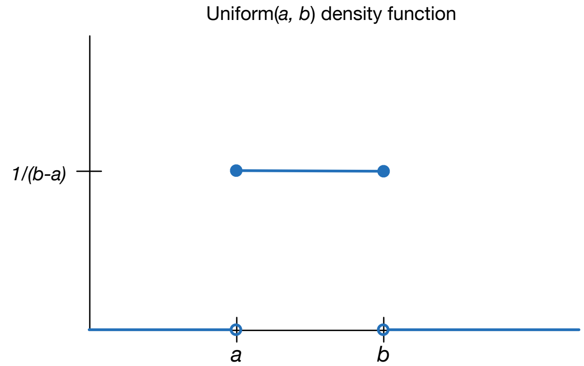
The *continuous uniform distribution*, which we shall generally just call the *uniform distribution*, describes the probability distribution on a finite interval $[a, b]$ which has the property that all subintervals of equal length are equally probable. The uniform distribution must be specified on a finite interval and is not defined for interval of infinite length. Here are some examples where we can use the uniform distribution to model a problem.

1. A RIPTA bus is scheduled to arrive at tunnel on Thayer St. at 8:00 am, but experience has shown that its arrival times vary between 8:00 am and 8:15 am. We could model this as a uniform distribution on the time interval $[0, 15]$, representing the number of minutes the bus is behind schedule.

2. Strokkur, one of the most famous geysers in Iceland, erupts approximately once every 10 minutes[4]. For a given 10-minute interval, we can model the probability that Strokkur will erupt by a uniform distribution on the interval $[0, 10]$

A uniform distribution is specified in terms of parameters a and b, which are the endpoints of the interval $[a.b]$ on which the uniform distribution is defined. What is the density function for the uniform distribution? Look at the picture below:

[4]Its neighbor Geysir, from which we get the word "geyser", hardly ever erupts these days.



The uniform density is a horizontal line between a and b and is 0 otherwise. (Does this make sense?) What is the height of the horizontal line? Since the integral of a probability density must integrate to 1, and since the uniform density is nothing more than a box, the area of the box must be 1. For that to be the case, the height of the box must be $1/(b-a)$. This is summarized below.

> *Uniform random variable*
>
> ---
>
> A continuous random variable Y has a *uniform distribution* on the interval $[a, b]$ if the pdf of Y is given by:
>
> $$f(y) = \begin{cases} \frac{1}{b-a} & a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$
>
> Y a *uniform random variable*, which we can write as $Y \sim \text{Uniform}(a, b)$.

Let's do an example.

Example. A circle has a radius which is uniformly distributed on the interval $[0, 1]$. What is the expected value of the area of the circle?

Let $Y \sim \text{Uniform}(0, 1)$. Then Y has density

$$f(y) = \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The area of a circle of radius y is given by $g(y) = \pi y^2$. Then the expected value of the area is:

$$\begin{aligned}
\mathbb{E}[g(Y)] &= \int_{-\infty}^{\infty} g(y)f(y)dy \\
&= \int_0^1 \pi y^2 \, 1dy \\
&= \pi \frac{y^3}{3}\Big|_0^1 \\
&= \frac{\pi}{3}
\end{aligned}$$

As with every other distribution, we are interested in the mean and the variance of the uniform distribution. These are given below.

Properties of the uniform distribution

Let Y have the *uniform distribution* on the interval $[a, b]$. Then

$$\begin{aligned}
\mathbb{E}(Y) &= \frac{a+b}{2} \\
Var(Y) &= \frac{(b-a)^2}{12}
\end{aligned}$$

It makes sense that the mean of the uniform distribution is halfway between the endpoints. To verify this, for $Y \sim \text{Uniform}(a, b)$ with density $f(y)$ as given above:

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} yf(y)dy \\
&= \int_a^b y\frac{1}{b-a}dy \\
&= \frac{1}{b-a}\frac{y^2}{2}\Big|_a^b \\
&= \frac{b^2-a^2}{2(b-a)} \\
&= \frac{(b+a)(b-a)}{2(b-a)} \\
&= \frac{a+b}{2}
\end{aligned}$$

The variance can be found similarly by computing $\mathbb{E}(Y^2)$ and using the Magic Variance Formula.