Chapter 10

Continuous probability distributions

10.1 Introduction

We call \( x \) a continuous random variable in \( a \leq x \leq b \) if \( x \) can take on any value in this interval. An example of a random variable is the height of adult human male, selected randomly from a population. (This takes on values in a range \( 0.5 \leq x \leq 3 \) meters, say, so \( a = 0.5 \) and \( b = 3 \).) If we select a male subject at random from a large population, and measure his height, we might expect to get a result in the proximity of 1.7-1.8 meters most often - thus, such heights will be associated with a larger value of probability than heights in some other interval of equal length, e.g. heights in the range \( 2.7 < x < 2.8 \) meters, say.

Unlike the case of discrete probability, however, the measured height can take on any real number within the interval of interest. This leads us to redefine our idea of the probability, using a continuous function in place of the discrete bar-graph seen in the previous chapter.

10.2 Basic definitions and properties

Definition

A function \( p(x) \) is a probability density provided it satisfies the following conditions:

1. \( p(x) \geq 0 \) for all \( x \).
2. \( \int_{a}^{b} p(x) \, dx = 1 \) where the possible range of values of \( x \) is \( a \leq x \leq b \).

The probability that a random variable \( x \) takes on values in the interval \( a_1 \leq x \leq a_2 \) is defined as

\[
\int_{a_1}^{a_2} p(x) \, dx.
\]
Unlike our previous discrete probability, we will not ask “what is the probability that $x$ takes on some exact value?” Rather, we ask for the probability that $x$ is within some range of values, and this is computed by performing an integral. (Remark: the probability that $x$ is exactly equal to $b$ is the integral $\int_b^b p(x) \, dx = 0$; the value is zero, by properties of the definite integral.)

**Definition**

The cumulative distribution function $F(x)$ represents the probability that the random variable takes on any value up to $x$, i.e.

$$F(x) = \int_a^x p(s) \, ds.$$  

The cumulative distribution is simply the area under the probability density.

The above definition has several implications:

**Properties of continuous probability**

1. Since $p(x) \geq 0$, the cumulative distribution is an *increasing* function.

2. The connection between the probability density and its cumulative distribution can be written (using the Fundamental Theorem of Calculus) as

$$p(x) = F'(x).$$

3. $F(a) = 0$. This follows from the fact that

$$F(a) = \int_a^a p(s) \, ds.$$  

By a property of the definite integral, this is zero.

4. $F(b) = 1$. This follows from the fact that

$$F(b) = \int_a^b p(s) \, ds = 1$$

by property 2 of $p(x)$.

5. The probability that $x$ takes on a value in the interval $a_1 \leq x \leq a_2$ is the same as

$$F(a_2) - F(a_1).$$

This follows from the additive property of integrals:

$$\int_a^{a_2} p(s) \, ds - \int_a^{a_1} p(s) \, ds = \int_{a_1}^{a_2} p(s) \, ds$$
Finding the normalization constant

Not every function can represent a probability density. For one thing, the function must be positive everywhere. Further, the total area under its graph should be 1, by property 2 of a probability density. However, in many cases we can convert a function to a probability density by simply multiplying it by a constant, \( C \), equivalent to the reciprocal of the total area under its graph over the interval of interest. This process is called “normalization”, and the constant \( C \) is called the normalization constant.

10.2.1 Example

Consider the function \( f(x) = \sin(\pi x/6) \) for \( 0 \leq x \leq 6 \).

(a) Normalize the function so that it describes a probability density.

(b) Find the cumulative distribution function, \( F(x) \).

Solution

The function is positive in the interval \( 0 \leq x \leq 6 \), so we can define the desired probability density. Let

\[
p(x) = C \sin(\frac{\pi}{6}x).
\]

(a) We must find the normalization constant, \( C \), such that

\[
\int_0^6 p(x) \, dx = 1.
\]

Here is how we find the constant:

\[
1 = \int_0^6 C \sin\left(\frac{\pi}{6}x\right) \, dx = C \frac{6}{\pi} \left( -\cos\left(\frac{\pi}{6}x\right) \right) \bigg|_0^6
\]

\[
1 = C \frac{6}{\pi} (-\cos(\pi) + \cos(0)) = C \frac{12}{\pi}
\]

Thus, we find that the desired constant of normalization is

\[
C = \frac{\pi}{12}
\]

Once we rescale our function by this constant, we get the desired probability density,

\[
p(x) = \frac{\pi}{12} \sin\left(\frac{\pi}{6}x\right).
\]

This density has the property that the total area under its graph over the interval \( 0 \leq x \leq 6 \) is 1. A graph of this probability density function is shown in Figure 10.1(a).

(b) \[
F(x) = \int_0^x p(s) \, ds = \frac{\pi}{12} \int_0^x \sin\left(\frac{\pi}{6}s\right) \, ds
\]
\[
F(x) = \frac{\pi}{12} \left( -\frac{6}{\pi} \cos\left(\frac{\pi}{6} x\right) \right) \bigg|_0^x = \frac{1}{2} \left( 1 - \cos\left(\frac{\pi}{6} x\right) \right).
\]

This cumulative distribution function is shown in Figure 10.1(b).

Figure 10.1: (a) The probability density \( p(x) \), (left) and (b) the cumulative distribution \( F(x) \) (right) for example 1.

### 10.3 Mean and median

Recall that we have defined the mean of a distribution of grades or mass in a previous chapter. For a mass density \( \rho(x) \), the idea of the mean coincides with the center of mass of the distribution,

\[
\bar{x} = \frac{\int_a^b x \rho(x) \, dx}{\int_a^b \rho(x) \, dx}.
\]

This definition can also be applied to a probability density, but in this case the integral in the denominator is simply 1 (by property 2), i.e. \( \int_a^b p(x) \, dx = 1 \). (The simplification is analogous to an observation we made for expected value in a discrete probability distribution.)

We define the mean of a probability density as follows:

**Definition**

For a random variable in \( a \leq x \leq b \) and a probability density \( p(x) \) defined on this interval, the **mean** or **average** value of \( x \) (also called the **expected value**), denoted \( \bar{x} \) is given by

\[
\bar{x} = \int_a^b x p(x) \, dx.
\]

The idea of median encountered previously in grade distributions also has a parallel here. Simply put, the median is the value of \( x \) that splits the probability distribution into two portions whose areas are identical.
Definition

The median $x_m$ of a probability distribution is a value of $x$ in the interval $a \leq x_m \leq b$ such that

$$\int_a^{x_m} p(x) \, dx = \int_{x_m}^b p(x) \, dx = \frac{1}{2}.$$ 

It follows from this definition that the median is the value of $x$ for which the cumulative distribution satisfies

$$F(x_m) = \frac{1}{2}.$$

10.3.1 Example

Find the mean and the median of the probability distribution found in Example 10.2.1.

Solution

To find the mean we compute

$$\bar{x} = \frac{\pi}{12} \int_0^6 x \sin \left( \frac{\pi}{6} x \right) \, dx.$$ 

Integration by parts is required here. Let $u = x$, $dv = \sin \left( \frac{\pi}{6} x \right) \, dx$. Then $du = dx$, $v = -\frac{6}{\pi} \cos \left( \frac{\pi}{6} x \right)$

The result is

$$\bar{x} = \frac{\pi}{12} \left( -x \frac{6}{\pi} \cos \left( \frac{\pi}{6} x \right) \bigg|_0^6 + \frac{6}{\pi} \int_0^6 \cos \left( \frac{\pi}{6} x \right) \, dx \right)$$

$$\bar{x} = \frac{1}{2} \left( -x \cos \left( \frac{\pi}{6} x \right) \bigg|_0^6 + \frac{6}{\pi} \sin \left( \frac{\pi}{6} x \right) \bigg|_0^6 \right)$$

$$\bar{x} = \frac{1}{2} \left( -6 \cos(\pi) + \frac{6}{\pi} \sin(\pi) - \frac{6}{\pi} \sin(0) \right) = \frac{6}{2} = 3$$

To find the median, $x_m$, we set

$$F(x_m) = \frac{1}{2}.$$ 

Using the form of the cumulative distribution from example 1, we find that

$$\frac{1}{2} \left( 1 - \cos \left( \frac{\pi}{6} x_m \right) \right) = \frac{1}{2}.$$

$$1 - \cos \left( \frac{\pi}{6} x_m \right) = 1.$$
\[ \cos\left(\frac{\pi}{6}x_m\right) = 0 \]

The angles whose cosine is zero are \( \pm\pi/2, \pm3\pi/2 \) etc. We select the angle in the relevant interval, i.e. \( \pi/2 \). This leads to

\[ \frac{\pi}{6}x_m = \frac{\pi}{2} \]

so the median is

\[ x_m = 3. \]

**Remark**

A glance at the original probability distribution should convince us that it is symmetric about the value \( x = 3 \). Thus we should have anticipated that the mean and median of this distribution would both occur at the same place, i.e. at the midpoint of the interval. This will be true in general for symmetric probability distributions, just as it was for symmetric mass or grade distributions.

### 10.3.2 How is the mean different from the median?

![Figure 10.2](image)

We have seen in Example 2 that for symmetric distributions, the mean and the median are the same. Is this always the case? When are the two different, and how can we understand the distinction?

Recall that the *mean* is closely associated with the idea of a center of mass, a concept from physics that describes the location of a pivot point at which the entire “mass” would exactly balance. It is worth remembering that

\[ \text{mean of } p(x) = \text{expected value of } x = \text{average value of } x. \]

This concept is not to be confused with the average value *of a function*, which is an average value of the \( y \) coordinate.

The *median* simply indicates a place at which the “total mass” is subdivided into two equal portions. (In the case of probability density, each of those portions represents an equal area, \( A_1 = A_2 = 1/2 \) since the total area under the graph is 1 by definition.)
Figure 10.2 shows how the two concepts of median (indicated by vertical line) and mean (indicated by triangular “pivot point”) differ. At the left, for a symmetric probability density, the mean and the median coincide, just as they did in Example 2. To the right, a small portion of the distribution was moved off to the far right. This change did not affect the location of the median, since the areas to the right and to the left of the vertical line are still equal. However, the fact that part of the mass is farther away to the right leads to a shift in the mean of the distribution, to compensate for the change.

Simply put, the mean contains more information about the way that the distribution is arranged spatially. This stems from the fact that the mean of the distribution is a “sum” - i.e. integral - of terms of the form $xp(x)\Delta x$. Thus the location along the $x$ axis, $x$, not just the “mass”, $p(x)\Delta x$, affects the contribution of parts of the distribution to the value of the mean.

10.4 Radioactive decay

Radioactive decay is a probabilistic phenomenon: an atom spontaneously emits a particle and changes into a new form. We cannot predict exactly when a given atom will undergo this event, but we can study a large collection of atoms and draw some interesting conclusions.

We can define a probability density function that represents the probability that an atom would decay at time $t$. This function represents the fraction of the atoms that decay per unit time. It turns out that a good candidate for such a function is

$$p(t) = Ce^{-kt},$$

where $k$ is a constant that represents the rate of decay of the specific radioactive material. In principle, this function is defined over the interval $0 \leq t \leq \infty$; that is, it is possible that we would have to wait a “very long time” to have all of the atoms decay. This means that these integrals have to be evaluated “at infinity”, introducing a complication that we will learn how to handle in the context of this example. Using this function we can characterize the mean and median decay time for the material.

Normalization

We first find the constant of normalization, i.e. set

$$\int_{0}^{\infty} p(t) \, dt = 1.$$

We need to ensure that

$$\int_{0}^{\infty} Ce^{-kt} \, dt = 1.$$

An integral of this sort in which one of the endpoints is at infinity is called an improper integral. We must see if it makes sense by computing it as a limit, i.e. by calculating

$$I_T = \int_{0}^{T} Ce^{-kt} \, dt$$
and computing a limit:

\[ I = \lim_{T \to \infty} I_T. \]

This is shown below

\[ I_T = C \int_0^T e^{-kt} \, dt = C \left[ \frac{e^{-kt}}{-k} \right]_0^T = \frac{1}{k} C (1 - e^{-kT}). \]

Now recall that the exponential function decays to zero so that

\[ \lim_{T \to \infty} e^{-kT} = 0. \]

Thus, the second term in braces will vanish as \( T \to \infty \) so that the values of the improper integral will be

\[ I = \lim_{T \to \infty} I_T = \frac{1}{k} C. \]

(We will discuss improper integrals more fully in a later chapter.) To find the constant of normalization \( C \) we set this equal to 1, i.e. \( \frac{1}{k} C = 1 \), which means that

\[ C = k. \]

Thus the probability density for the decay is

\[ p(t) = ke^{-kt}. \]

This means that the fraction of atoms that decay between time \( t_1 \) and \( t_2 \) is

\[ k \int_{t_1}^{t_2} e^{-kt} \, dt. \]

**Cumulative decays**

The fraction of the atoms that decay between time 0 and time \( t \) (i.e. “by time \( t \)” ) is

\[ F(t) = \int_0^t p(s) \, ds = k \int_0^t e^{-ks} \, ds. \]

We can simplify this expression:

\[ F(t) = k \left[ \frac{e^{-ks}}{-k} \right]_0^t = - [e^{-kt} - e^0] = 1 - e^{-kt}. \]

Thus, the probability of the atoms decaying by time \( t \) (which means anytime up to time \( t \)) is

\[ F(t) = 1 - e^{-kt}. \]

We note that \( F(0) = 0 \) and \( F(\infty) = 1 \), as expected for a cumulative distribution function.
Median decay time

We can use the cumulative distribution function to help determine the median decay time, $t_m$. To determine $t_m$, the time at which half of the atoms have decayed, we set $F(t_m) = 0.5$, giving us

$$F(t_m) = 1 - e^{-kt_m} = \frac{1}{2}$$

we get

$$e^{-kt_m} = \frac{1}{2}$$
$$e^{kt_m} = 2$$
$$kt_m = \ln 2$$

So we find that

$$t_m = \frac{\ln 2}{k}.$$ 

Thus half of the atoms have decayed by this time. (Remark: this is easily recognized as the half life of the radioactive process from previous familiarity with exponentially decaying functions.)

Mean decay time

The mean time of decay $ar{t}$ is given by

$$\bar{t} = \int_0^\infty tp(t) \, dt.$$ 

We compute this integral again as an improper integral by taking a limit as the top endpoint increases to infinity, i.e. we first find

$$I_T = \int_0^T tp(t) \, dt,$$ 

and then set

$$\bar{t} = \lim_{T \to \infty} I_T.$$ 

To compute $I_T$ we use integration by parts:

$$I_T = \int_0^T tk e^{-kt} \, dt = k \int_0^T t e^{-kt} \, dt.$$ 

Let $u = t, dv = e^{-kt} \, dt$. Then $du = dt, v = e^{-kt}/(-k)$, so that

$$I_T = k \left[ t \frac{e^{-kt}}{-k} - \int_0^T \frac{e^{-kt}}{-k} \, dt \right]_0^T$$

$$I_T = \left[ -te^{-kt} + \int_0^T e^{-kt} \, dt \right]_0^T = \left[ -te^{-kt} - \frac{e^{-kt}}{k} \right]_0^T.$$
\[ I_T = \left[ -T e^{-kT} - \frac{e^{-kT}}{k} + \frac{1}{k} \right]. \]

Now as \( T \to \infty \), we have
\[ T e^{-kT} \to 0, \quad e^{-kT} \to 0 \]
so that
\[ \bar{t} = \lim_{T \to \infty} I_T = \frac{1}{k}. \]
Thus the mean or expected decay time is
\[ \bar{t} = \frac{1}{k}. \]

### 10.5 Discrete versus continuous probability

In an earlier chapter, we compared the treatment of two types of mass distributions. We first explored a set of discrete masses strung along a “thin wire”. Later, we considered a single “bar” with a continuous distribution of density along its length. In the first case, there was an unambiguous meaning to the concept of “mass at a point”. In the second case, we could assign a mass to some section of the bar between, say \( x = a \) and \( x = b \). (To do so we had to integrate the mass density on the interval \( a \leq x \leq b \).) In the first case, we talked about the mass of the objects, whereas in the latter case, we were interested in the idea of density (mass per unit distance: Note that the units of mass density are not the same as the units of mass.)

The same dichotomy exists in the topic of probability. In an earlier chapter, we were concerned with the probability of discrete events whose outcome belongs to some finite set of possibilities (e.g. Head or Tail for a coin toss, allele A or a in genetics). But many random processes lead to an infinite, continuous set of possible outcomes. We need the notion of continuous probability to deal with such cases. In continuous probability, we consider the probability density - analogous to mass density. We can assign a probability to some range of values of the outcome between \( x = a \) and \( x = b \). (To do so we have to integrate the probability density on the interval \( a \leq x \leq b \).)

The examples below provide some further insight to the connection between continuous and discrete probability. In particular, we will see that one can arrive at the idea of probability density by refining a set of measurements and making the appropriate scaling. We explore this connection in more detail below.

#### 10.5.1 Example: Student heights

Suppose we measure the heights of all UBC students. This would produce about 30,000 data values. We could make a graph and show how these heights are distributed. For example, we could subdivide the student body into those students between 0 and 1.5m, and those between 1.5 and 3 meters. Our bar graph would contain two bars, with the number of students in each height category represented by the heights of the bars, as shown in Figure 10.3(a).

Suppose we want to keep more detail. We could divide the population into smaller groups by shrinking the size of the interval or “bin” into which height is subdivided. (An example is shown
Figure 10.3: Refining a histogram by increasing the number of bins leads (eventually) to the idea of a continuous probability density.

in Figure 10.3(b)). Here, by a “bin” we mean a little interval of width $\Delta h$ where $h$ is height, i.e. a height interval. For example, we could keep track of the heights in increments of 50 cm. If we were to plot the number of students in each height category, then as the size of the bins gets smaller, so would the height of the bar: there would be fewer students in each category if we increase the number of categories.

To keep the bar height from shrinking, we might reorganize the data slightly. Instead of plotting the number of students in each bin, we might plot the number of students in bin $\Delta h$.

If we do this, then both numerator and denominator decrease as the size of the bins is made smaller, so that the shape of the distribution is preserved (i.e. it does not get flatter).

We observe that in this case, the number of students in a given height category is represented by the area of the bar corresponding to that category:

$$\text{Area of bin} = \Delta h \left( \frac{\text{number of students in bin}}{\Delta h} \right) = \text{number of students in bin}.$$

The important point to consider is that the height of each bar in the plot represents the number of students per unit height.

This type of plot is precisely what leads us to the idea of a density distribution. As $\Delta h$ shrinks, we get a continuous graph. If we “normalize”, i.e. divide by the total area under the graph, we get a probability density, $p(h)$ for the height of the population. As noted, $p(h)$ represents the fraction of students per unit height whose height is $h$. It is thus a density, and has the appropriate units.

More generally,

$$p(x) \Delta x$$

represents the fraction of individuals whose height is in the range $x \leq h \leq x + \Delta x$.

10.5.2 Examples: Age dependent mortality

Let $p(a)$ be a probability density for the probability of mortality of a female Canadian non-smoker at age $a$, where $0 \leq a \leq 120$. (We have chosen an upper endpoint of age 120 since practically no Cana-
dian female lives past this age at present.) Let $F(a)$ be the cumulative distribution corresponding to this probability density.

(a) What is the probability of dying by age $a$?

(b) What is the probability of surviving to age $a$?

(c) Suppose that we are told that $F(75) = 0.8$ and that $F(80)$ differs from $F(75)$ by 0.11. What is the probability of surviving to age 80? Which is larger, $F(75)$ or $F(80)$?

(d) Use the information in part (c) to estimate the probability of dying between the ages of 75 and 80 years old. Further, estimate $p(80)$ from this information.

Solution

(a) The probability of dying by age $a$ is the same as the probability of dying any time up to age $a$, i.e. it is

$$F(a) = \int_0^a p(s) \, ds,$$

i.e. it is the cumulative distribution for this probability density. That, precisely, is the interpretation of the cumulative function.

(b) The probability of surviving to age $a$ is the same as the probability of not dying before age $a$. By the elementary properties of probability discussed in the previous chapter, this is

$$1 - F(a).$$

(c) From the properties of probability, we know that the cumulative distribution is an increasing function, and thus it must be true that $F(80) > F(75)$. Then $F(80) = F(75) + 0.11 = 0.8 + 0.11 = 0.91$. Thus the probability of surviving to age 80 is $1 - 0.91 = 0.09$. This means that 9% of the population will make it to their 80'th birthday, according to this analysis.

(d) The probability of dying between the ages of 75 and 80 years old is exactly

$$\int_{75}^{80} p(x) \, dx.$$

However, we can also state this in terms of the cumulative function, since

$$\int_{75}^{80} p(x) \, dx = \int_0^{80} p(x) \, dx - \int_0^{75} p(x) \, dx = F(80) - F(75) = 0.11$$

Thus the probability of death between the ages of 75 and 80 is 0.11.

To estimate $p(80)$, we use the connection between the probability density and the cumulative distribution:

$$p(x) = F'(x).$$

Then it is approximately true that

$$p(x) \approx \frac{F(x) - F(x - \Delta x)}{\Delta x}.$$
(Recall the definition of the derivative - the limit of the slope of the secant line as the width increments $\Delta x$ approach 0.)

Thus

\[ p(80) \approx \frac{F(80) - F(75)}{5} = \frac{0.11}{5} = 0.022 \text{ per year} \]

### 10.5.3 Example: Raindrop size distribution

During a Vancouver rainstorm, the density function which describes the radii of raindrops is constant over the range $0 \leq r \leq 4$ (where $r$ is measured in mm) and zero for larger $r$.

(a) What is the density function $p(r)$?

(b) What is the cumulative distribution $F(r)$?

(c) In terms of the volume, what is the cumulative distribution $F(V)$?

(d) In terms of the volume, what is the density function $p(V)$?

(e) What is the average volume of a raindrop?

**Solution**

This problem is challenging because one may be tempted to think that the uniform distribution of drop radii should give a uniform distribution of drop volumes. This is not the case, as the following argument shows! The sequence of steps is illustrated in Figure 10.4.

![Figure 10.4: Probability densities for raindrop radius and raindrop volume (left panels) and for the cumulative distributions (right) of each.](image-url)
(a) The density function is \( p(r) = \frac{1}{4} \), \( 0 \leq r \leq 4 \). This means that the probability \textit{per unit radius} of finding a drop of size \( r \) is the same for all radii in \( 0 \leq r \leq 4 \). Some of these drops will correspond to small volumes, and others to very large volumes. We will see that the probability \textit{per unit volume} of finding a drop of given volume will be quite different.

(b) The cumulative distribution function is

\[
F(r) = \int_{0}^{r} \frac{1}{4} ds = \frac{r}{4}, \quad 0 \leq r \leq 4.
\]

(c) The cumulative distribution function is proportional to the radius of the drop. We use the connection between radii and volume of spheres to rewrite that function in terms of the volume of the drop: Since

\[
V = \frac{4}{3} \pi r^3
\]

we have

\[
r = \left( \frac{3}{4 \pi} \right)^{1/3} V^{1/3}
\]

and so

\[
F(V) = \frac{r}{4} = \frac{1}{4} \left( \frac{3}{4 \pi} \right)^{1/3} V^{1/3}.
\]

We find the range of values of \( V \) by substituting \( r = 0, 4 \) into the equation \( V = \frac{4}{3} \pi r^3 \), to get \( V = 0, \frac{4}{3} \pi 4^3 \). Therefore the interval is \( 0 \leq V \leq \frac{4}{3} \pi 4^3 = (256/3) \pi \).

(d) We now use the connection between the probability density and the cumulative distribution, namely that \( p \) is the derivative of \( F \). Now that the variable has been converted to volume, that derivative is a little more “interesting”:

\[
p(V) = F'(V)
\]

Therefore,

\[
p(V) = \frac{1}{4} \left( \frac{3}{4 \pi} \right)^{1/3} \frac{1}{3} V^{-2/3}.
\]

Thus the probability \textit{per unit volume} of finding a drop of volume \( V \) in \( 0 \leq V \leq \frac{4}{3} \pi 4^3 \) is not at all uniform. This results from the fact that the differential quantity \( dr \) behaves very differently from \( dV \), and reinforces the fact that we are dealing with density, not with a probability per se. We note that this distribution has smaller values at larger values of \( V \).

(e) The range of values of \( V \) is

\[
0 \leq V \leq \frac{4}{3} \pi 4^3 = \frac{256\pi}{3}
\]
and therefore the mean volume is
\[
\bar{V} = \int_{0}^{256/3} V p(V) dV
\]
\[
= \frac{1}{12} \left( \frac{3}{4\pi} \right)^{1/3} \int_{0}^{256/3} V \cdot V^{-2/3} dV
\]
\[
= \frac{1}{12} \left( \frac{3}{4\pi} \right)^{1/3} \int_{0}^{256/3} V^{1/3} dV
\]
\[
= \frac{1}{12} \left( \frac{3}{4\pi} \right)^{1/3} \left[ \frac{4}{3} V^{4/3} \right]_{0}^{256/3}
\]
\[
= \frac{1}{16} \left( \frac{3}{4\pi} \right)^{1/3} \left( \frac{256\pi}{3} \right)^{4/3}
\]
\[
= \frac{64\pi}{3} \approx 67 \text{mm}^3.
\]

10.6 Moments of a probability distribution

We are now familiar with some of the properties of probability distributions. On this page we will introduce a set of numbers that describe various properties of such distributions. Some of these have already been encountered in our previous discussion, but now we will see that these fit into a pattern of quantities called moments of the distribution.

Moments

Let \( f(x) \) be any function which is defined and positive on an interval \([a, b]\). We might refer to the function as a distribution, whether or not we consider it to be a probability density distribution. Then we will define the following moments of this function:

\[
\begin{align*}
\text{zero'th moment} & \quad M_0 = \int_{a}^{b} f(x) \, dx \\
\text{first moment} & \quad M_1 = \int_{a}^{b} x \, f(x) \, dx \\
\text{second moment} & \quad M_2 = \int_{a}^{b} x^2 \, f(x) \, dx \\
\vdots & \\
\text{n'th moment} & \quad M_n = \int_{a}^{b} x^n \, f(x) \, dx.
\end{align*}
\]

Observe that moments of any order are defined by integrating the distribution \( f(x) \) with a suitable power of \( x \) over the interval \([a, b]\). However, in practice we will see that usually moments
up to the second are usefully employed to describe common attributes of a distribution.

### 10.6.1 Moments of a probability density distribution

In the particular case that the distribution is a probability density, \( p(x) \), defined on the interval \( a \leq x \leq b \), we have already established the following:

\[
M_0 = \int_a^b p(x) \, dx = 1
\]

(This follows from the basic property of a probability density.) Thus

*The zero’th moment of any probability density is 1.*

Further

\[
M_1 = \int_a^b x \, p(x) \, dx = \bar{x} = \mu.
\]

That is,

*The first moment of a probability density is the same as the mean (i.e. expected value) of that probability density.*

So far, we have used the symbol \( \bar{x} \) to represent the mean or average value of \( x \) but often the symbol \( \mu \) is also used to denote the mean.

The second moment, of a probability density also has a useful interpretation. From above definitions, the second moment of \( p(x) \) over the interval \( a \leq x \leq b \) is

\[
M_2 = \int_a^b x^2 \, p(x) \, dx
\]

We will shortly see that the second moment helps describe the way that density is distributed about the mean. For this purpose, we must describe the notion of *variance or standard deviation.*

**Variance and standard deviation**

Two kids of approximately the same size can balance on a teeter-totter by sitting very close to the point at which the beam pivots. They can also achieve a balance by sitting at the very ends of the beam, equally far away. In both cases, the center of mass of the distribution is at the same place: precisely at the pivot point. However, the mass is distributed very differently in these two cases. In the first case, the mass is clustered close to the center, whereas in the second, it is distributed further away. We may want to be able to describe this distinction, and we could do so by considering higher moments of the mass distribution.

Similarly, if we want to describe how a probability density distribution is distributed about its mean, we consider moments higher than the first. We use the idea of the *variance* to describe whether the distribution is clustered close to its mean, or spread out over a great distance from the mean.
Variance

The variance is defined as the average value of the quantity \((distance \ from \ mean)^2\), where the average is taken over the whole distribution. (The reason for the square is that we would not like values to the left and right of the mean to cancel out.)

For discrete probability with mean, \(\mu\) we define variance by

\[
V = \sum (x_i - \mu)^2 p_i
\]

For a continuous probability density, with mean \(\mu\), we define the variance by

\[
V = \int_a^b (x - \mu)^2 p(x) \, dx
\]

The standard deviation

The standard deviation is defined as

\[
\sigma = \sqrt{V}
\]

Let us see what this implies about the connection between the variance and the moments of the distribution.

Relationship of variance to second moment

From the equation for variance we calculate that

\[
V = \int_a^b (x - \mu)^2 p(x) \, dx = \int_a^b (x^2 - 2\mu x + \mu^2) p(x) \, dx.
\]

Expanding the integral leads to:

\[
V = \int_a^b x^2 p(x) \, dx - \int_a^b 2\mu x p(x) \, dx + \int_a^b \mu^2 p(x) \, dx
\]

\[
= \int_a^b x^2 p(x) \, dx - 2\mu \int_a^b x p(x) \, dx + \mu^2 \int_a^b p(x) \, dx.
\]

We recognize the integrals in the above expression, since they are simply moments of the probability distribution. Plugging in these facts, we arrive at
Thus

\[ V = M_2 - \mu^2 \]

Thus the variance is clearly related to the second moment and to the mean of the distribution.

**Relationship of variance to second moment**

> From the definitions given above, we find that

\[ \sigma = \sqrt{V} = \sqrt{M_2 - \mu^2} \]

### 10.6.2 Example

Consider the continuous distribution, in which the probability is constant, \( p(x) = C \), for values of \( x \) in the interval \([a, b]\) and zero for values outside this interval. Such a distribution is called a uniform distribution. (It has the shape of a rectangular band of height \( C \) and base \((b - a)\).) It is easy to see that the value of the constant \( C \) should be \( 1/(b - a) \) so that the area under this rectangular band will be 1, in keeping with the property of a probability distribution. Thus the equation of this probability is \( p(x) = \frac{1}{b - a} \). We compute the first moments of this probability density

\[ M_0 = \int_a^b p(x) \, dx = \frac{1}{b - a} \int_a^b 1 \, dx = 1. \]

(This was already known, since we have determined that the zeroth moment of any probability density is 1.) We also find that

\[ M_1 = \int_a^b x \, p(x) \, dx = \frac{1}{b - a} \int_a^b x \, dx = \frac{1}{b - a} \frac{b^2 - a^2}{2} = \frac{b + a}{2}. \]

This last expression can be simplified by factoring, leading to

\[ \mu = M_1 = \frac{(b - a)(b + a)}{2(b - a)} = \frac{b + a}{2} \]

Thus we have found that the mean \( \mu \) is in the center of the interval \([a, b]\), as expected. The median would be at the same place by a simple symmetry argument: half the area is to the left and half the area is to the right of this point.
To find the variance we might first calculate the second moment,

\[ M_2 = \int_a^b x^2 \, p(x) \, dx = \frac{1}{b - a} \int_a^b x^2 \, dx \]

It can be shown by simple integration that this yields the result

\[ M_2 = \frac{b^3 - a^3}{3(b - a)}, \]

which can be simplified to

\[ M_2 = \frac{(b - a)(b^2 + ab + a^2)}{3(b - a)} = \frac{b^2 + ab + a^2}{3}. \]

We would then compute the variance

\[ V = M_2 - \mu^2 = \frac{b^2 + ab + a^2}{3} - \frac{(b + a)^2}{4}. \]

After simplification, we get

\[ V = \frac{b^2 - 2ab + a^2}{12} = \frac{(b - a)^2}{12}. \]

The standard deviation is then

\[ \sigma = \frac{(b - a)}{2 \sqrt{3}}. \]