# Lecture 7

# Continuous Random Variables

In the last two lectures we studied discrete random variables. The range of such random variables is always a finite or a countably infinite set. Now we consider continuous random variables whose range is always an uncountable set and uses tools from differential/integral calculus, instead of sequence/series. The results are, however, quite analogous to those of discrete random variables. Often the change involves replacing the summation symbol by the integral symbol.

## 7.1 Empirical Approximation of a Density

If we know exactly what the probability space of a random experiment is, we may be able to analytically find the density of the random variable, as we did in earlier lectures for binomial, geometric, hypergeometric and negative binomial random variables.

A sharp reader must have noticed that earlier we did not derive the Poisson density analytically. Poisson density was justified through its close approximation to various empirically observed data. Its underlying probability space was not specified. The empirical evidence consisted of observing the random variable (data) and constructing the corresponding relative frequencies which were then checked to see if they well approximated the density.

For continuous random variables an analytic derivation of their densities are usually difficult[1] since the unerlying probability space is often difficult to model. Instead, we observe the values of the random variable in a large number of repetitions of the experiment. The resulting collection of observed values of the random variable, namely the empirical distribution, is usually presented as a relative frequency table, called a histogram. One can use a mathematical curve (a probability density) that closely resembles the shape of the histogram as a model for the unknown actual density of the random variable. Happily it turns out that certain

[1] Consider Example 7.1.3, and Exercises 7.4.26 for a few exceptions.





shapes of histograms are distinctly different in different types of experiments, giving rise to some typical shapes of the densities. Consider the following example.

**Example - 7.1.1** - **(Exponentially declining & bell shaped densities)** Figure 7.1 shows relative frequency distributions (histograms) of two real experiments. The left histogram is that of the fill weights of 200 cans of lemonade drink mix. Each can was advertised to have 20 ounces of fill weight. The automatic filling machine was set to pour 20 ounces in each can. The fluctuations in the fill weights are modeled by the superimposed bell shaped curve.

The histogram on the right side in Figure 7.1 is that of the life spans (in hours) of 200 light bulbs. Each bulb was advertised to have average working life of 500 hours. The variation in the life spans is captured by the superimposed curve which is exponentially decaying.

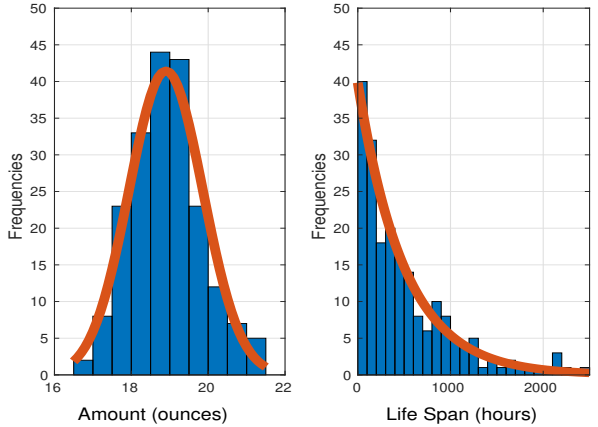


Figure 7.1: Densities Approximating Empirical Data.

When measuring weights, lengths, areas, volumes or other such characteristics of mass produced items, typically produce a relative frequency distribution (histogram) which is well approximated by a bell shaped curve (density). On the other hand, measuring waiting times until an event of interest occurs (such as waiting till a bulb fails, or a transistor burns out, or a window glass cracks), leads to a relative frequency distribution whose shape is well approximated by an exponentially decaying curve (density). Other types of shapes arise while modeling various other types of real life phenomena, leading to more varieties of densities.

In this and the next lectures our aim is to catalog a wide variety of such mathematical curves (densities) that have been found to well approximate histograms of random a wide variety of random phenomenon. The official definition of a continuous random variable and its density is as follows.

**Definition - 7.1.1** - **(Continuous random variable)** *A random variable, $X$,*

*with cumulative distribution function (cdf) $F(t)$, is called a continuous random variable, and the cdf is called absolutely continuous, if $F(t)$ obeys the following properties:*

*(i) $F(t)$ is a nondecreasing function of $t$,*
*(ii) $F(t)$ is a continuous function of $t$,*
*(iii) $F(-\infty) = 0$, $F(\infty) = 1$,*
*(iv) $\frac{d}{dt}F(t) = f(t)$ exists for almost all $t$, so that*

$$\int_{-\infty}^{t} f(x)\,dx = F(t), \qquad \text{for all } t \in \mathbb{R},$$

*In this case, $f$ is called the density of $X$. Hence, a density of a continuous random variable is a nonnegative function whose integral is one. Sometimes we denote the density and cdf as $f_X(x)$, $F_X(t)$ to remind ourselves that these are the density and cdf of the random variable $X$.*

**Example - 7.1.2** - **(Life spans of windshields)** The length of time, $X$, (measured in years) it takes for a car windshield to develop a crack has density and cdf

$$f(x) := \begin{cases} 0 & \text{if } x \le 0, \\ 0.01e^{-0.01x} & \text{if } x > 0, \end{cases} \qquad F(t) := \begin{cases} 0 & \text{if } t \le 0, \\ 1 - e^{-0.01t} & \text{if } t > 0. \end{cases}$$

The probability that a new car will develop a crack within its first three years is

$$F(3) = \mathbb{P}(X \le 3) = \int_0^3 f(x)\,dx = 0.01\int_0^3 e^{-0.01x}\,dx = 1 - e^{-3(0.01)} \approx 0.0206.$$

The random variable $X$ measures the duration of time until the first crack develops on a randomly selected car. The relevent areas under its density provide the corresponding probabilities. The number 0.01 is called its parameter and it represents the intensity of car window breaking opportunities in the region in which the car will be exposed to. The larger the parameter the more often car windows will develop cracks. Note that the cdf, $F$, is a differentiable function of $t$ except at $t = 0$, and the derivative of $F$ is $\frac{d}{dt}F(t) = f(t)$, obeying the fundamental theorem of calculus (except at one point $t = 0$ which carries zero probability and hence negligible).

**(Radioactivity)** As another example, consider a random variable $X$ to model the radioactivity of a radioactive substance. Now $X$ measures the waiting time in seconds till an $\alpha$-particle is ejected by the substance. Here is a model for its density

$$f(x) := \begin{cases} 0 & \text{if } x \notin (0, \infty], \\ 13e^{-13x} & \text{if } x \in (0, \infty). \end{cases}$$

The corresponding cdf of $X$ is

$$F(t) := \begin{cases} 0 & \text{if } t \le 0, \\ \int_0^t 13e^{-13x}dx & \text{if } t > 0 \end{cases} = \begin{cases} 0 & \text{if } t \le 0, \\ 1 - e^{-13t} & \text{if } t > 0. \end{cases}$$





The parameter of this density is 13 indicating about 13 distintegrations per second, or in other words, on average every $\frac{1}{13}$-th second a disintegration takes place.

In both of the above examples the model for the random variables was the same however the parameter(s) were different. In general the parameter of an exponential model is denoted by $\lambda$ and we succinctly express the model by the notation $X \sim Exp(\lambda)$. In the first example $X \sim Exp(0.01)$ with unit of time being 1 year while in the second example $X \sim Exp(13)$ with unit of time being 1 second.

**Example - 7.1.3** - **(Mathematically deriving a density)** We pick a point at random from the interval $[3, 14]$, giving the sample space $S = [3, 14]$. In this case,

$$\mathbb{P}(A) = \frac{\text{length of } A}{\text{length of } S} = \frac{\text{length of } A}{14 - 3} = \frac{\text{length of } A}{11}.$$

The random variable $X$ being the identity $X(s) = s$. It has the cdf

$$F(t) = \mathbb{P}(X \le t) = \begin{cases} 0 & \text{if } t < 3, \\ \frac{t-3}{14-3} & \text{if } t \in [3, 14], \\ 1 & \text{if } t > 14. \end{cases}$$

By differentiating we get the density $f(x) = \frac{1}{11}$ if $x \in (3, 14)$, and $f(x) = 0$, for $x < 3$ or $x > 14$. At $x = 3, 14$ we may take $f(x) = 0$ as well. Here $X$ is a continuous random variable since its cdf $F$ obeys all the requirements of Definition 7.1.1. $X$ represents the outcome of "a randomly selected point from the interval $[3, 14]$". Its density is called a uniform density since its graph is a flat line over the range of $X$.

**Remark - 7.1.1** - **(Partition introduced by a continuous random variable)** For continuous random variables, the range being too large, we often are unable to express the partition or functional aspect of a random variable defined on its underlying probability space. Therefore, typically continuous random variables are studied through their densities (or cdfs) or through their empirical population perspectives. When the cdf is hard or impossible to express in a closed mathematical form, the discussion then gets further restricted around the density only.

**Example - 7.1.4** - **(Density and cdf without specifying the sample space)** Consider the function

$$f(x) = \begin{cases} c\,(1 + x)^3 & \text{if } |x| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find the constant $c$ so that $f$ is a density of a continuous random variable $X$. (ii) Find $\mathbb{P}(|X| \ge \frac{1}{2})$. (iii) Find the cdf $F(t)$ of the random variable $X$.

(i) For $f$ to be a density we must have

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

Therefore, by choosing $c = \frac{1}{4}$ makes $f$ a density. (ii) Then we have

$$\mathbb{P}(|X| \ge \tfrac{1}{2}) = 1 - \mathbb{P}(|X| < \tfrac{1}{2}) = 1 - \int_{-1/2}^{1/2} f(x)\,dx = 1 - \frac{1}{4}\int_{-1/2}^{1/2} (1+x)^3\,dx = \frac{176}{256}.$$

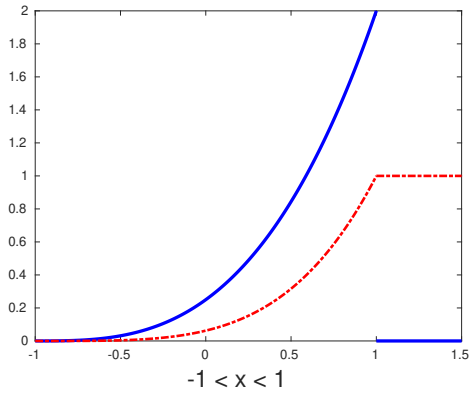


Figure 7.2: Density (solid line) and CDF (dashed line) of a RV.

(iii) For any $t < -1$¡ we have $F(t) = 0$ and $F(t) = 1$ for any $t > 1$. For any $t \in [-1, 1]$, we have

$$F(t) \;=\; \int_{-\infty}^{t} f(x)\,dx \;=\; \frac{1}{4}\int_{-1}^{t} (1+x)^3\,dx \;=\; \frac{(1+t)^4}{16}.$$

The density and the cdf are shown in Figure 7.2.

## 7.2 Comparison of Discrete & Continuous RV's

The issue is how are our earlier discrete random variables different from what we have called continuous random variables? Think of discreteness and continuity as the gender/sex of the random variable. We identify the gender/sex of the rv first before we deal with it any further. Here are some differences that help us identify the gender/sex of an rv.

- **(Look at the range of $X$).** One obvious difference is that the range of a discrete random variable is always a finite or countably infinite set (with perhaps a negligible exception) and the range of our newly defined random variable must be an uncountable set, usually an interval. So if the range is a finite or countable infinite set then the rv must be discrete. All the random variables of the last lecture fell in this category — please look back and verify this.
- **(Check out the smoothness of the cdf $F$).** The cdf of our continuous random variable must obey the fundamental theorem of calculus. That is, if $\frac{d}{dt}F(t) = f(t)$, then for any $s < t$,

$$\boxed{\mathbb{P}(s < X \leq t) \;=\; F(t) - F(s) \;=\; \int_{s}^{t} f(x)\,dx.}$$





As stated in Definition 7.1.1, this property of the cdf $F(t)$ is called absolute continuity Instead, sometimes we casually say that "$F(t)$ is a cdf of a continuous random variable".

- **(Look at the jumps, f any, of the cdf $F$).** Note that cdf $F$ of a continuous rv cannot have any jumps. In the discrete case, the density and the cdf are related as

$$\mathbb{P}(X = a) \;=\; f(a) \;=\; F(a) - F(a^-)$$

and represents the probability that $X = a$. When $X$ is a continuous r.v., $\mathbb{P}(X = a) = 0$ for any $a$, (since the integral over a single point is always zero) and $f(a)$ does not represent the probability of anything. Even more, sometimes the density of a continuous random variable becomes $f(x) > 1$, which can never happen in the discrete case.

In the continuous case the density $f(a)$ does have a link with probability. By the (integral) mean value theorem, for any small interval $(a, a+\delta]$, there exists an $\xi \in [a, a+\delta]$ such that

$$\delta\, f(\xi) \;=\; \int_{a}^{a+\delta} f(u)\,du \;=\; \mathbb{P}(a < X \leq a+\delta) \;=\; F(a+\delta) - F(a).$$

Hence, if the interval $(a, a+\delta]$ is very small, then $f(a)\delta$ is approximately the probability that $X$ lies in the interval $(a, a+\delta]$. The word "density" comes from Physics. When matter is spread out over an interval, $f(a)$ is the mass density at $a$.

- **(Density of a continuous rv is not unique)** Since the Riemann integral is not affected by changing the integrand at a point, it should be remembered that the density, $f$, of a continuous random variable is not unique. It can be changed on a countably many points without effecting anything. However, it does not matter to the probabilities of events which version of $f$ we use. So, we often use that version of the density which has fewer points of discontinuity if any.

A random variable can have both a discrete part and a continuous part. Such an $X$ (or its cdf $F$) is then called a mixed random variable (or a mixed cdf). Over the values of its range where it has a discrete part it is treated like a discrete random variable and over its continuous part it is treated like a continuous random variable. Over the discrete part the cdf $F$ has jumps, and over the continuous part the cdf $F$ obeys the fundamental theorem of calculus.

There is yet one more variety, for which $F(t)$ has no points of jump but it does not obey the fundamental theorem of calculus. Such distributions are called singular. We will not deal with them since they require a little higher level tools of calculus.

**Example - 7.2.1 - (Distribution of deferred annuity – a mixed cdf)** The present value of a deferred annuity is a random variable whose distribution is plotted in Figure 7.3. This distribution has a discrete point at 0, where it has a jump of size $q$. It also has a component with positive derivative. The sum of the jump