RVs and their probability distributions

Continuous Random Variables

Recall that a sample space is said to be continuous if its outcomes can fall anywhere in some interval on the real line. (The interval may be infinite in one or both directions.) The same language is used for random variables:

A **continuous random variable** is one whose possible values comprise an interval on the real line.

Background to the Theory in the Continuous Case - for those who are interested

There is a considerable problem in extending the definition of probability distribution to sample spaces or random variables which are continuous. Namely, the number of points in any interval (even a bounded one) is not only infinite, it is uncountable. This was proved by Georg Cantor. As a result, even for a bounded interval, we cannot have every single number in that interval have the same (nonzero) probability because of an even worse version of the same reason as in the infinite discrete case!

So we take another approach, which comes from looking at a simple way of defining a probability that accords with the intuition of "equally likely" in some ways. One example of this is as follows:

Suppose there is a stretch of highway 10 miles long, and an accident is equally likely to occur at any point on this stretch of highway. If an accident occurs, what is the probability that it happens in the first 3 miles of this stretch of highway?

Since an accident is equally likely anywhere on this stretch of highway, it makes sense that the probability should be $\frac{3}{10}$, as the first 3 miles take up $\frac{3}{10}$ of the total length.

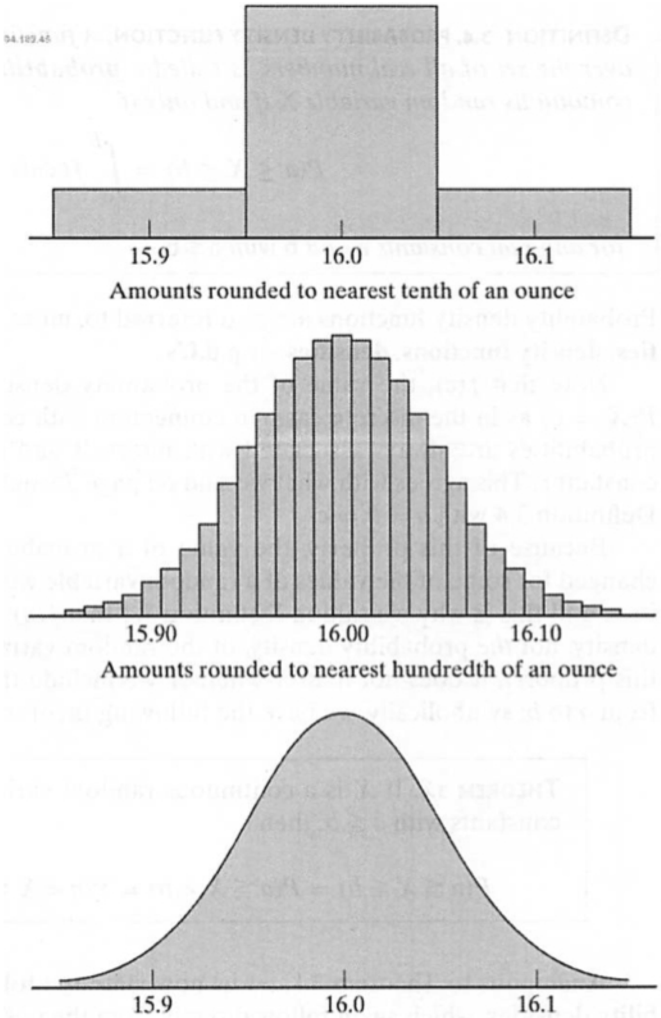
And similarly, if an accident occurs, the probability that it happened on any particular piece of this stretch of highway should be the ratio of the length of that piece to the total length, 10 miles.

However, what this line of reasoning implies is that the probability that the accident happened at any particular point (say, the 4-mile mark) will be 0, since the length of a point is 0!

What this example demonstrates is that if the sample space is a bounded interval (of finite length, that is) then it is fairly easy to define the probability of a subinterval, and that lets us compute the probability of any set in the sample space that is built by forming unions or intersections or complements or any combination of these operations, just by drawing pictures (not by using any mathematical rules). All of the probabilities work as they should according to our previous theory. The only problem is that we do not know how to define the probability of a single point (that is, a single real number). But if we are willing to give that up temporarily, we can make a probability model that makes some sense and accords fairly well with our intuition.

But what if we do not want to have "equally likely" probability distributions (or if the interval we need to use is infinite)? How could we extend this way of defining probabilities on intervals?

Here is one approach: If we look at the graph (histogram) of a pdf for a finite random variable, it forms a step function. Usually, in the important examples, these are also fairly symmetric "bell-shaped" graphs such as we have seen before. I will give you an example from another textbook. These graphs are meant to represent finite (or discrete) RVs which are approximations to a certain continuous RV, namely the actual amount of soft drink in a can which is supposed to have 16 ounces in it. (In reality the can cannot be filled so as to have exactly 16 ounces, so the actual amount varies and in principle can be any amount in some interval around 16.)



Look at these graphs carefully. The first one represents the probability distribution for a RV which is the actual amount of soft drink in a can (a measured value) measured to the nearest tenth of an ounce. Measuring to a fixed degree of precision turns this continuous RV to a discrete one, and its probability distribution is a step function as pictured. (For now you do not need to worry about how the probabilities were computed: the point is that the graph will be some step function).

The second graph represents another way of approximating the continuous RV even better, by measuring to the nearest hundredth of an ounce. This RV is still discrete but has even more possible values than we got by measuring to the nearest tenth, and they are closer together. So the step graph has bars which are narrower, and the overall shape is still more or less the same, but smoother.

Now imagine measuring to higher and higher degrees of precision. This would lead to a series of step graphs which would be better and better approximations to the continuous RV, and their graphs would presumably get smoother and smoother. In fact, this should remind you of the way in which the Riemann integral is defined, and this can be seen as a way to approach a nice continuous graph. Since the areas of the bars in the step graphs represent probabilities, we can see that it makes sense to define probabilities for a continuous RV by choosing or finding some nice curve (graph of a function) and then declaring that the probabilities will be areas under the curve.

Putting these two lines of thinking together, we conclude the following:

For a continuous RV,

1. We cannot hope to define probabilities for individual points (numbers), but a good theory could be based on defining probabilities of intervals of values.
2. We can then hope to extend the computation of probabilities to any (finite or countable) unions of intervals, (finite or countable) intersections of intervals, or complements of these.
3. One way to do this might involve finding the area under a curve by Riemann integration of a suitable function.

Probability on Continuous RVs via Riemann Integrals

Probability Density:
A probability density (function) for a continuous random variable X is a function $f(x)$ which is defined for each possible value x of X, and which satisfies the following conditions:

- $f(x)$ is non-negative
- $f(x)$ is integrable
- $\int f(x) = 1$, where the integral is taken over all possible values of X

The probability that X takes its value in some interval (a, b) is then defined as

$$P(a < X < b) = \int_a^b f(x)dx$$

Note: It follows from this definition of probability that **the probability of any individual number a is 0:** that is because

$$P(X = a) = \int_a^a f(x)dx = 0$$

for any integrable function $f(x)$.

And it also follows from the properties of Riemann integrals that it does matter whether or not we include the endpoints of the intervals over which we integrate. There is no practical difference in the continuous case between "less than" and "less than or equal to", and similarly there is no practical difference between "greater than" and "greater than or equal to".

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_a^b f(x)dx$$

Example:

Let $f(x) = ke^{-3x}$ for $x > 0$ and $f(x) = 0$ elsewhere. We will find k so that this is a probability density function and use it to compute some probabilities.

In order to make sure this is a pdf, we need to check that $f(x)$ is non-negative everywhere in $x > 0$. This will be true as long as $k > 0$ since the exponential is always positive. We also need to check that $f(x)$ is integrable and has the total integral 1. This is not so clear since the domain of X is infinite. There will be an indefinite integral and we need to check that it converges (and to the right thing!).

$$\int_0^\infty ke^{-3x}dx = k\int_0^\infty e^{-3x}dx \text{ since k is constant.}$$

Now use substitution via the "Reverse Chain Rule for Exponentials" $\int e^u du = e^u$ with $u = -3x$ and $du = -3dx$:

$$k\int_0^\infty e^{-3x}dx = \frac{k}{-3}\int_0^\infty e^u du = \frac{k}{-3}[e^u]_{x=0}^\infty = \frac{k}{-3}\left(\lim_{b\to\infty} e^{-3b} - e^0\right) = \frac{k}{-3}(0-1) = \frac{k}{3}$$

So we see that if $k = 3$ the total integral will be 1.

Note: By a similar line of reasoning we can see that $f(x) = ke^{-kx}$ on $x > 0$ will always be a good pdf for any $k > 0$.

How to use this density function $f(x) = 3e^{-3x}$ to compute probabilities:

- What is the probability that X is between 0.5 and 1? (As mentioned before, since X is continuous it does not matter whether or not we include the endpoints so we can write either $0.5 < X < 1$ or $0.5 \le X \le 1$ and the result will be the same.)

  Compute $P(0.5 < X < 1) = \int_{0.5}^1 3e^{-3x}dx = [-e^{-3x}]_{0.5}^1 = -e^{-3} - (-e^{-3(0.5)}) \approx 0.173$.

- What is the probability that X is greater than 10?
  Compute

  $$P(X > 10) = \int_{10}^\infty 3e^{-3x}dx = [-e^{-3x}]_{10}^\infty = \lim_{b\to\infty}(-e^{-3b}) - (-e^{-30}) = e^{-30} \approx 9.4\times 10^{-14}$$

  (a very small number)

- What is the probability that X is not more than $\frac{1}{2}$?
  Compute

  $$P(X \le 0.5) = \int_0^{0.5} 3e^{-3x}dx = [-e^{-3x}]_0^{0.5} = (-e^{-3(0.5)}) - (-e^0) = 1 - e^{-1.5} \approx 0.78$$

  (So we see that this density function concentrates most of its weight close to 0.)