Continuous random variables

CE 311S

What was the difference between discrete and continuous random variables?

The possible outcomes of a discrete random variable (finite or infinite) can be listed out; the possible outcomes of a continuous random variable cannot. These usually take the form of a real interval.

A few examples:

- The high temperature on a given day
- The time an engine can run before needing repair
- The stream flow of the Colorado River as it enters Lake Buchanan

But is anything *truly* continuous?

You might object that all of these examples might really be discrete, for several reasons:

- You might feel that nature is fundamentally discrete (atoms, quantum physics, etc.)

But is anything *truly* continuous?

You might object that all of these examples might really be discrete, for several reasons:

- You might feel that nature is fundamentally discrete (atoms, quantum physics, etc.)
- Our *measurements* are necessarily discrete (limited precision)

These arguments are largely philosophical. If there are many possible values, discrete distributions can quickly become unwieldy. Continuous distributions let us harness the power of the calculus.

Which is easier to calculate?

$$\sum_{i=1}^{1000} i^2$$

$$\int_{i=1}^{1000} x^2 dx$$

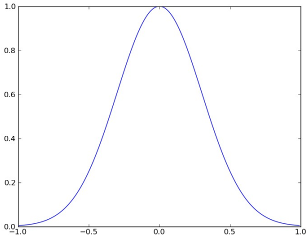
The sum is 333,833,500. The integral is 333,333,333. The difference is well under 1%, and the integral was much easier to calculate.

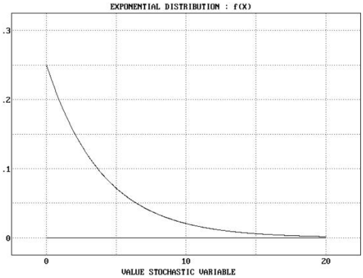
Likewise, continuous distributions may be easier to work with, and a good approximation to, a truly discrete process.

PROBABILITY DENSITY FUNCTIONS

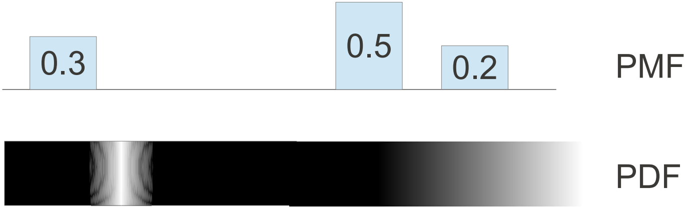
For discrete random variables, the **probability mass function** was the fundamental concept. We used the pmf to calculate probabilities, expected values, standard deviations, and so forth.

For continuous random variables, we have a **probability density function** (pdf) $f_X(x)$ which will play a similar role.





The words “mass” and “density” should be suggestive. With discrete random variables, there are individual points which have mass of their own. With continuous random variables, think about a solid object whose density varies throughout.



The main way a probability density function is used is as follows.

If f_X is a probability density function for the random variable X, then for any a and b with $a \leq b$,

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

Note that this is analogous to the discrete case, where $P(a \leq X \leq b) = \sum_{x=a}^{b} p(x)$. As before, we can omit the subscript X if it is obvious which random variable we are talking about.

From our probability axioms, we have the following properties of any pdf:

$$P(A) \geq 0 \Rightarrow f(x) \geq 0 \text{ for all } x$$

$$P(\mathcal{S}) = 1 \Rightarrow P(-\infty < x < \infty) = 1 \Rightarrow \int_{-\infty}^{\infty} f(x)dx = 1$$

Two (possibly surprising) implications

1. It *is* possible for $f(x)$ to be greater than 1. It is the *area under* any part of $f(x)$ which can't exceed 1.
2. The probability that X takes *any* specific value is 0.

Be sure to understand the implications of both of these.

Example

I want to describe a pdf to describe the amount of time I wait for the bus in the morning, assuming that (a) I don't look at the schedule and leave the house at a random time (b) buses arrive every 10 minutes. What is this pdf, and what is the probability that I have to wait more than 8 minutes?

With these assumptions, my possible waiting time is $[0, 10]$, and all of these are equally likely.

The only way to represent "all of these are equally likely" is to choose a constant pdf, so $f(x) = k$. What value should k be?

$$\int_0^{10} k dx = 1$$

$$10k = 1$$

$k = 1/10$, so $f(x) = \begin{cases} 1/10 & 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$

It is important to specify where a pdf is defined. The answer $f(x) = 1/10$ is **wrong** without specifying the range where it is valid.

(After all, $\int_{-\infty}^{\infty}(1/10)dx = \infty$!)

What is the probability I have to wait more than 8 minutes?

$$\int_8^\infty f(x)\ dx$$

$$= \int_8^{10} (1/10)\ dx$$

This is 0.2, so there is a 20% probability of having to wait more than 8 minutes.

This is actually the **uniform distribution**, the first continuous distribution family we'll see.

If all possible values of X are equally likely and fall in the range $[a, b]$, X has a uniform distribution and its pdf is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Example

The length of my lectures is given by a continuous probability distribution of the form kx^2, where x lies between 0 and 75 minutes. What is k? What is the probability my lecture is done in less than 60 minutes?

$$\int_0^{75} kx^2 dx = 1 \iff \left[\frac{k}{3}x^3\right]_0^{75} = \frac{k}{3}(75^3) = 1 \iff k = 7.11 \times 10^{-6}$$

Therefore, $f(x) = \begin{cases} 7.11 \times 10^{-6} x^2 & 0 \leq x \leq 75 \\ 0 & \text{otherwise} \end{cases}$

$$\int_0^{60} (7.11 \times 10^{-6}) x^2 dx = 0.512$$

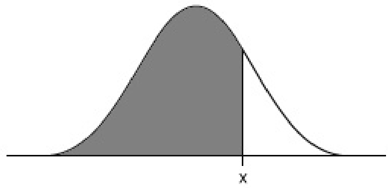
CUMULATIVE DISTRIBUTION FUNCTIONS

The cumulative distribution function $F(x)$ is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Again, this is the same as the definition of the cumulative mass function in Chapter 3, but with an integral instead of a sum.

$F(x)$ can be thought of as the area under the pdf to the left of x.



Clearly, F is nondecreasing, $\lim_{x\to-\infty} F(x) = 0$, and $\lim_{x\to\infty} F(x) = 1$

These are useful because it simplifies probability calculations. If we know F, then

$$P(a \leq X \leq b) = F(b) - F(a)$$

without having to calculate any integrals.

An important special case is

$$P(X \geq a) = 1 - F(a)$$

Example

Last class, we said that the pdf of me waiting for the bus was given by

$$f(x) = \begin{cases} 1/10 & 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

What's the probability that I have to wait more than 5 minutes?

Approach: Derive the CDF F, and calculate $1 - F(5)$

Solution

$$f(x) = \begin{cases} 1/10 & 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

For the CDF, there are three cases:

Case I, $x \leq 0$: $\int_{-\infty}^{x} f(y)dy = \int_{-\infty}^{x} 0dy = 0$

Case II, $0 < x \leq 10$: $\int_{-\infty}^{x} f(y)dy = \int_{-\infty}^{0} 0dx + \int_{0}^{x} (1/10)dy = x/10$

Case III, $x > 10$: $\int_{-\infty}^{x} f(y)dy = F(10) + \int_{10}^{x} 0dy = 1$

so

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{10} & 0 \leq x \leq 10 \\ 1 & x > 10 \end{cases}$$

Solution

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{10} & 0 \le x \le 10 \\ 1 & x > 10 \end{cases}$$

Therefore $P(X > 5) = 1 - F(5) = 1/2$.

This was a fairly simple case; when the integral of f is trickier, having the CDF handy can be more of a time saver.

If we know the CDF, can we obtain the PDF?

Example

The length of my lectures can be described by the pdf

$$f(x) = \begin{cases} 7.11 \times 10^{-6} x^2 & 0 \leq x \leq 75 \\ 0 & \text{otherwise} \end{cases}$$

Use the CDF to find the probability that:

1. My lecture lasts at least 70 minutes?
2. My lecture lasts as least 45 minutes?
3. My lecture is less than 60 minutes long?
4. My lecture is between 45 and 60 minutes long?

The CDF is also used to describe other properties related to continuous random variables.

Let $0 < p < 1$. The $(100p)$**th percentile** $\eta(p)$ is the value such that $F(\eta(p)) = p$

The **median** of a continuous random variable is the 50th percentile.

Example

The length of time I wait for the bus is given by the CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{10} & 0 \leq x \leq 10 \\ 1 & x > 10 \end{cases}$$

1. The 10th percentile is 1, because $F(1) = 0.10$
2. The 45th percentile is 4.5, because $F(4.5) = 0.45$
3. The median is 5, because $F(5) = 0.50$
4. The 99th percentile is 9.9, because $F(9.9) = 0.99$

The following formulas shouldn't be surprising:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$$

$$\sigma_X^2 = V[X] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx$$

$$V[X] = E[X^2] - \mu_X^2$$

$$\sigma_X = \sqrt{V[X]}$$

The *mode* of X is the point where the PDF takes the greatest value (if more than one such point, all are modes).

The length of time I wait for the bus is given by the pdf

$$f(x) = \begin{cases} 1/10 & 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{10} (x/10) dx$$

(Why?)

$$= \left[\frac{x^2}{20} \right]_0^{10}$$

$$= 5$$

The length of time I wait for the bus is given by the pdf

$$f(x) = \begin{cases} 1/10 & 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

$$V[X] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx = \int_0^{10} \frac{(x-5)^2}{10} dx$$

$$= \left[\frac{(x-5)^3}{30} \right]_0^{10} = 8.33$$

Or, using the shortcut formula:

$$E[X^2] = \int_0^{10} (x^2/10) dx = \left[\frac{x^3}{30} \right]_0^{10} = 33.33$$

so

$$V[X] = 33.33 - 5^2 = 8.33$$

The standard deviation is $\sqrt{8.33} = 2.89$

Example

The pdf for the length of class is $f(x) = \begin{cases} 7.11 \times 10^{-6} x^2 & 0 \leq x \leq 75 \\ 0 & \text{otherwise} \end{cases}$

1. What is the mean?
2. What is the median?
3. What is the mode?
4. What is the standard deviation?
5. What is the 95th percentile?

HOW DO WE COME UP WITH A PDF?

How can we find a pdf for a given random variable X? Unlike the pmf, we can't directly find the probability that X takes each possible value (since these are zero).

There are two main approaches:

- Find the cdf, and take the derivative.
- Method of transformations, using a pdf that we already know.

Method 1: Find the cdf

Sometimes, it is easier to find the cdf.

In the bus example, if every waiting time is equally likely between 0 and 10 minutes, then the probability of waiting less than x minutes must be $x/10$ (if $0 \leq x \leq 10$).

Therefore the pdf is the derivative of this, or $1/10$ for $0 \leq x \leq 10$, and 0 otherwise.

(We will see another example of this with the exponential distribution.)

Method 2: Transformations

If there is a direct formula for the random variable in terms of another one, sometimes we can find the pdf directly, using the method of transformations.

This is easiest if this formula (say $Y = g(X)$) is strictly increasing or decreasing; in this case

$$f_Y(y) = f_X(x_1)/|g'(x_1)|$$

where x_1 is the value of X corresponding to y; if none such value exists, $f_Y(y) = 0$

Example

In a certain group of the population, assume that the average time it takes to run a mile is uniformly distributed between 5 and 10 minutes. What is the average *speed* of this population, in miles per minute?

The average time it takes to run a mile is 7.5 minutes; but because the relationship between speed and time is nonlinear, we cannot say that the average speed is $1/7.5$ miles/minute, but need to use LOTUS.

To use LOTUS, we need the PDF for the the speed of the population. Call the time needed to run a mile X, and the speed Y; these are related by $Y = 1/X$, so $g(X) = 1/X$.

Since X can take any value between 5 and 10, Y can take any value between $1/10$ and $1/5$; for all of these values the equation $y = 1/x_1$ has a solution.

$$f_Y(y) = f_X(x_1)/|g'(x_1)|$$

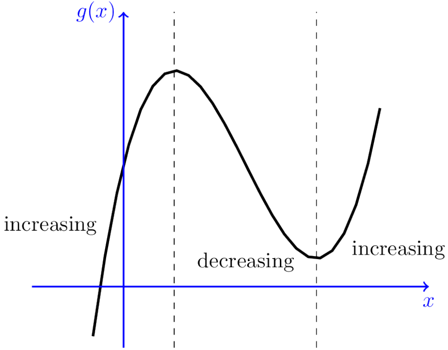
where x_1 is the value of X corresponding to y; if none such value exists, $f_Y(y) = 0$

The pdf f_X is $1/5$ for $5 \le x \le 10$; and $|g'(x_1)| = 1/x_1^2 = y^2$ since $y = 1/x_1$.

So $f_Y(y) = 1/5y^2$ for $1/10 \le y \le 1/5$, and 0 otherwise.

So $E[f_Y(y)] = \int_{1/10}^{1/5} \frac{1}{5y^2} y \; dy = 0.139 \approx 1/7.21$ miles per minute.

This formula only works if g is strictly increasing or strictly decreasing. If not, we can usually divide it into pieces which are strictly increasing and decreasing. Now the equation $g(x) = y$ may have multiple solutions; call these solutions $x_1, \ldots, x_n$ and add up the previous formula for each of these solutions.



$$f_Y(y) = \sum_{i=1}^{n} \frac{f_X(x_i)}{|g'(x_i)|}$$