# Continuous Random Variables Notes

Lt Col Ken Horton  Professor Bradley Warner

23 June, 2020

## Objectives

1) Define the terms probability density function (pdf) and cumulative distribution function (cdf).
2) Given a continuous random variable, describe probability using the pdf and cdf.
3) Find the mean and variance of a continuous random variable.

### Continuous Random Variables

In the last lesson, we introduced random variables, and explored discrete random variables. In this lesson, we will move into continuous random variables, their properties, their distribution functions, and how they differ from discrete random variables.

Recall that a continuous random variable has a domain that is a continuous interval (or possibly a group of intervals). For example, let $Y$ be the random variable corresponding to the height of a randomly selected individual. While our measurement will necessitate "discretizing" height to some degree, technically, height is a continuous random variable since a person could measure 67.3 inches or 67.4 inches or anything in between.

#### Continuous Distribution Functions

So how do we describe the randomness of continuous random variables? In the case of discrete random variables, the probability mass function (pmf) and the cumulative distribution function (cdf) are used to describe randomness. However, recall that the pmf is a function that returns the probability that the random variable takes the inputted value. Due to the nature of continuous random variables, the probability that a continuous random variable takes on any one individual value is technically 0. Thus, a pmf cannot apply to a continuous random variable.

Rather, we describe the randomness of continuous random variables with the *probability density function* (pdf) and the *cumulative distribution function* (cdf). Note that the cdf has the same interpretation and application as in the discrete case.

### Probability Density Function

Let $X$ be a continuous random variable. The probability density function (pdf) of $X$, given by $f_X(x)$ is a function that describes the behavior of $X$. It is important to note that in the continuous case, $f_X(x) \neq \mathrm{P}(X = x)$, as the probability of $X$ taking any one individual value is 0.

The pdf is a *function*. The input of a pdf is any real number. The output is known as the density. The pdf has three main properties:

1) $f_X(x) \geq 0$
2) $\int_{S_X} f_X(x)\,\mathrm{d}x = 1$
3) $\mathrm{P}(X \in A) = \int_{x \in A} f_X(x)\,\mathrm{d}x$

Properties 2) and 3) imply that the area underneath a pdf represents probability.

### Cumulative Distribution Function

The cumulative distribution function (cdf) of a continuous random variable has the same interpretation as it does for a discrete random variable. It is a *function*. The input of a cdf is any real number, and the output is the probability that the random variable takes a value less than or equal to the inputted value. It is denoted as $F$ and is given by:

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_x(t)\,\mathrm{d}t$$

> *Example*:
> Let $X$ be a continuous random variable with $f_X(x) = 2x$ where $0 \leq x \leq 1$. Verify that $f$ is a valid pdf. Find the cdf of $X$. Also, find the following probabilities: $\mathrm{P}(X < 0.5)$, $\mathrm{P}(X > 0.5)$, and $\mathrm{P}(0.1 \leq X < 0.75)$. Finally, find the median of $X$.

To verify that $f$ is a valid pdf, we simply note that $f_X(x) \geq 0$ on the range $0 \leq x \leq 1$. Also, we note that $\int_0^1 2x\,\mathrm{d}x = x^2\Big|_0^1 = 1$.

Using `R`, we find

```
integrate(function(x)2*x,0,1)
```

```
## 1 with absolute error < 1.1e-14
```

Or we can use the `mosaicCalc` package to find the anit-derivative. If the package is not installed, you can use the `Packages` tab in `RStudio` or type `install.packages("mosaicCalc")` at the command prompt. Load the library.

```
library(mosaicCalc)
```

```
(Fx<-antiD(2*x~x))
```

```
## function (x, C = 0)
## 1 * x^2 + C
```

```
Fx(1)-Fx(0)
```

```
## [1] 1
```

Graphically, the pdf is displayed below:

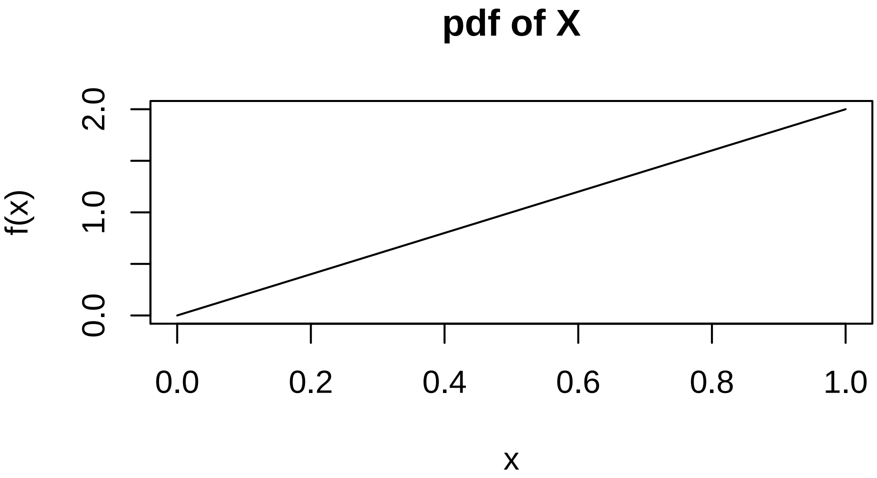


Figure 1: pdf of $X$

The cdf of $X$ is found by

$$\int_0^x 2t\,\mathrm{d}t = t^2\Big|_0^x = x^2$$

This is `antiD` found from the calculations above.

So,

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x^2, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

The plot of the cdf of $X$ is shown below:

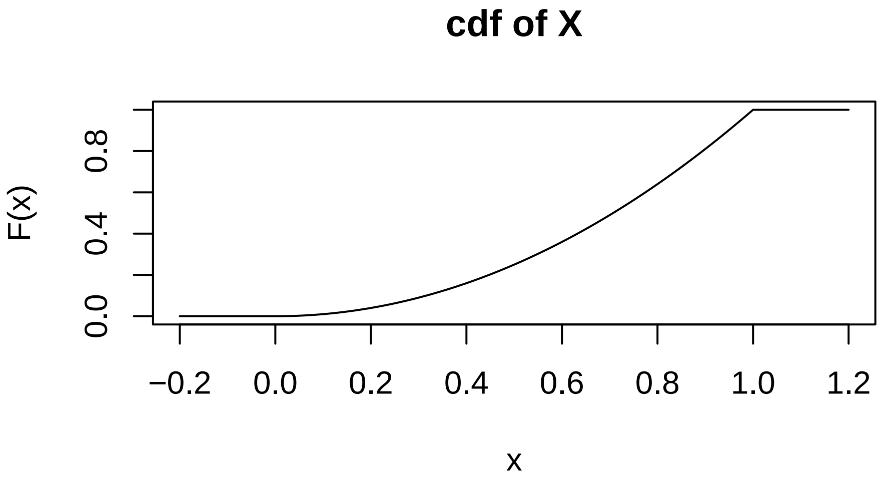


Figure 2: cdf of $X$

Probabilities are found either by integrating the pdf or using the cdf:

$\mathrm{P}(X < 0.5) = \mathrm{P}(X \leq 0.5) = F_X(0.5) = 0.5^2 = 0.25$

$\mathrm{P}(X > 0.5) = 1 - \mathrm{P}(X \leq 0.5) = 1 - 0.25 = 0.75$

$\mathrm{P}(0.1 \leq X < 0.75) = \int_{0.1}^{0.75} 2x\,\mathrm{d}x = 0.75^2 - 0.1^2 = 0.553$

```r
integrate(function(x)2*x,.1,.75)
```

```
## 0.5525 with absolute error < 6.1e-15
```

Alternatively, $\mathrm{P}(0.1 \leq X < 0.75) = F(0.75) - F(0.1) = 0.75^2 - 0.1^2 = 0.553$

```r
Fx(0.75)-Fx(0.1)
```

```
## [1] 0.5525
```

The median of $X$ is the value $x$ such that $\mathrm{P}(X \leq x) = 0.5$. So we simply solve $x^2 = 0.5$ for $x$. Thus, the median of $X$ is $\sqrt{0.5} = 0.707$.

Or using `R`

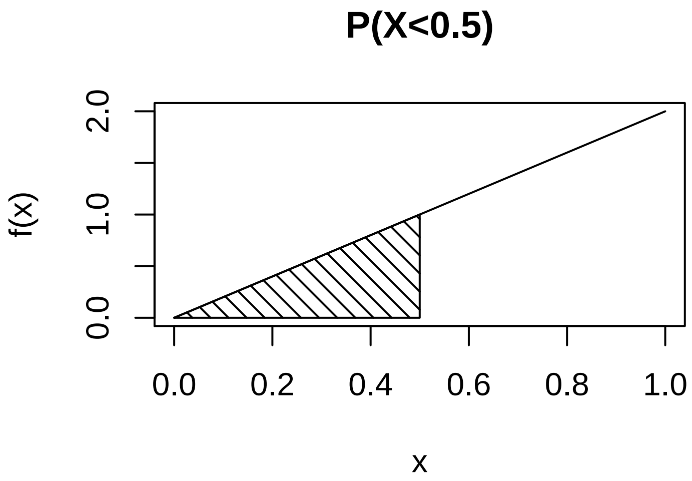


Figure 3: Probability represented by shaded area

```
uniroot(function(x)(Fx(x)-.5),c(0,1))$root
```

```
## [1] 0.7071067
```

**Simulation**

As in the case of the discrete random variable, we can simulate a continuous random variable if we have an inverse for the cdf. The range of the cdf is $[0, 1]$, so we generate a random number in this interval and then apply the inverse cdf to obtain a random variable. In a similar manner, for a continuous random variable, we use the following pseudo code:
1. Generate a random number in the interval $[0, 1]$, $U$.
2. Find the random variable $X$ from $F_X^{-1}(U)$.
In `R` for our example, this looks like the following.

```
sqrt(runif(1))
```

```
## [1] 0.3817585
```

```
results <- do(10000)*sqrt(runif(1))
```

```
inspect(results)
```

```
##
## quantitative variables:
##   name   class       min        Q1    median        Q3       max      mean
## 1 sqrt numeric 0.0129881 0.5044009 0.7121372 0.8678138 0.9998806 0.6696693
##         sd     n missing
## 1 0.235766 10000       0
```

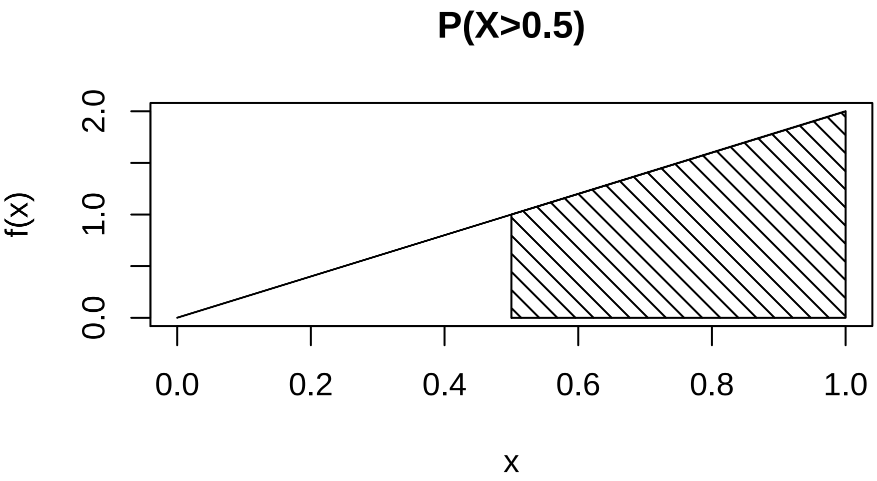
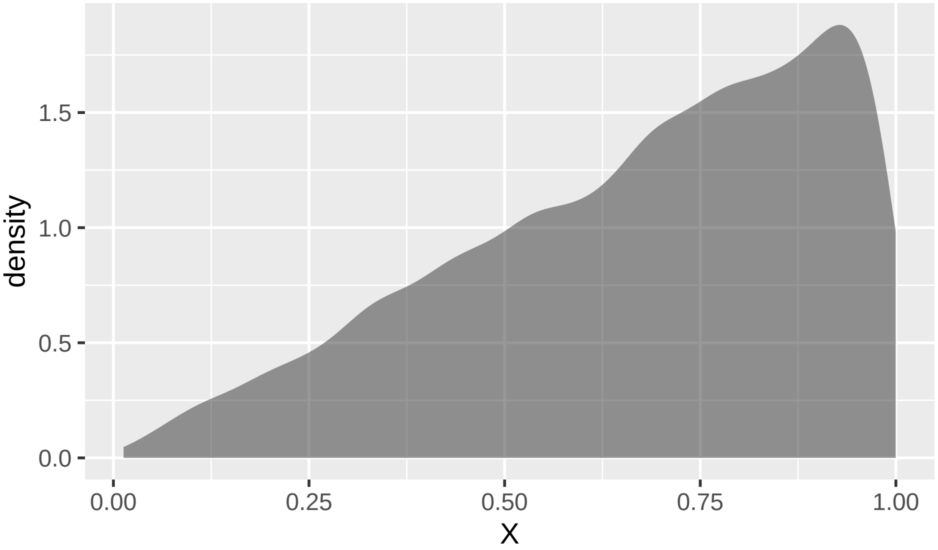


Figure 4: Probability represented by shaded area

```
results %>%
  gf_density(~sqrt,xlab="X")
```



## Moments

As with discrete random variables, moments can be calculated to summarize characteristics such as center and spread. In the discrete case, expectation is found by multiplying each possible value by its associated probability and summing across the domain ($\mathrm{E}(X) = \sum_x x \cdot f_X(x)$). In the continuous case, the domain of $X$ consists of an infinite set of values. From your calculus days, recall that the sum across an infinite domain is represented by an integral.

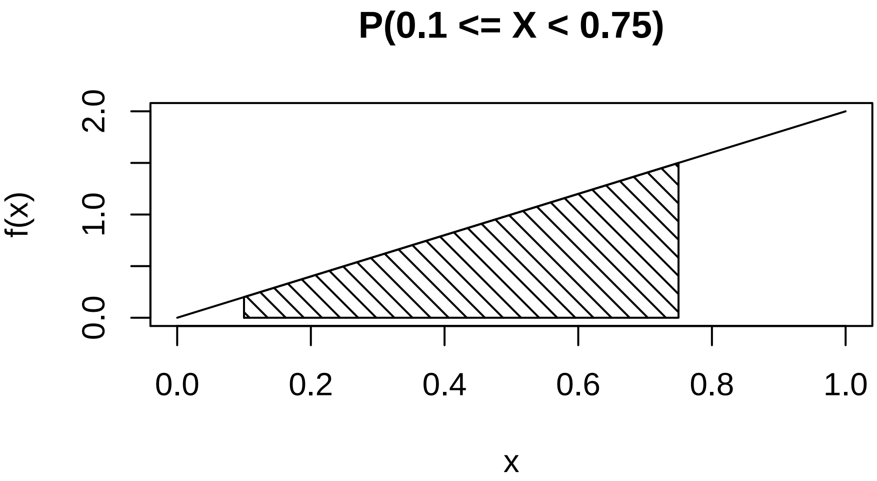


Figure 5: Probability represented by shaded area

Let $g(X)$ be any function of $X$. The expectation of $g(X)$ is found by:

$$\mathrm{E}(g(X)) = \int_{S_X} g(x) f_X(x)\,\mathrm{d}x$$

**Mean and Variance**

Let $X$ be a continuous random variable. The mean of $X$, or $\mu_X$, is simply $\mathrm{E}(X)$. Thus,

$$\mathrm{E}(X) = \int_{S_X} x \cdot f_X(x)\,\mathrm{d}x$$

As in the discrete case, the variance of $X$ is the expected squared difference from the mean, or $\mathrm{E}[(X-\mu_X)^2]$. Thus,

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2] = \int_{S_X} (x-\mu_X)^2 \cdot f_X(x)\,\mathrm{d}x$$

Recall Application problem 3 from last lesson. In this problem, you showed that $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$. Thus,

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \int_{S_X} x^2 \cdot f_X(x)\,\mathrm{d}x - \mu_X^2$$

> *Example*:
> Consider the random variable $X$ from above. Find the mean and variance of $X$.

$$\mu_X = \mathrm{E}(X) = \int_0^1 x \cdot 2x\,\mathrm{d}x = \frac{2x^3}{3}\bigg|_0^1 = \frac{2}{3} = 0.667$$

Side note: Since the mean of $X$ is smaller than the median of $X$, we say that $X$ is skewed to the left, or negatively skewed.

Using `R`.

```
integrate(function(x)x*2*x,0,1)
```

```
## 0.6666667 with absolute error < 7.4e-15
```

Or using `antiD()`

```
Ex<-antiD(2*x^2~x)
Ex(1)-Ex(0)
```

```
## [1] 0.6666667
```

Using our simulation.

```
mean(~sqrt,data=results)
```

```
## [1] 0.6696693
```

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \int_0^1 x^2 \cdot 2x\,\mathrm{d}x - \left(\frac{2}{3}\right)^2 = \frac{2x^4}{4}\bigg|_0^1 - \frac{4}{9} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} = 0.056$$

```
integrate(function(x)x^2*2*x,0,1)$value-(2/3)^2
```

```
## [1] 0.05555556
```

or

```
Vx<-antiD(x^2*2*x~x)
Vx(1)-Vx(0)-(2/3)^2
```

```
## [1] 0.05555556
```

```
var(~sqrt,data=results)*9999/10000
```

```
## [1] 0.05558005
```

And finally, the standard deviation of $X$ is $\sigma_X = \sqrt{\sigma_X^2} = \sqrt{1/18} = 0.236$.

### File Creation Information

- File creation date: 2020-06-23
- Windows version: Windows 10 x64 (build 17763)
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.6.0
- `tidyverse` package version: 1.3.0
- `mosaicCalc` package version: 0.5.1