In the last activity, we explored special discrete distributions: binomial, hypergeometric, geometric, negative binomial, and poisson. In this activity, we'll investigate continuous random variables – variables for which we cannot even begin to count all the possible outcomes.

1. The following histogram displays the distribution of Unit #1 test scores for students in Fall 2012. Using this data from the 32 students in the class, how would you estimate the probability of a student scoring 10 or less?
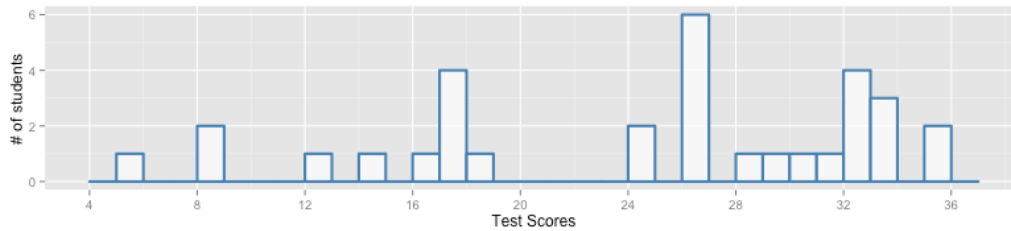
2. You can't tell from the histogram, but some students earned partial credit on some questions (and, therefore, earned scores such as 5.5 and 8.25). If we let X = a score from this test, does X represent a <u>discrete</u> or <u>continuous</u> random variable? Explain.

3. Suppose I want to generalize these test scores beyond the 32 students who took the test. I want to imagine the distributions of scores I'd get if an infinite number of students took this 36-point test.

   Since students can earn partial-credit on test questions, it might make sense to treat X as a continuous random variable. If so, I can draw a smooth curve to represent the theoretical distribution I'd get if I could test an infinite number of students*. This distribution implies it's possible for students to get <u>any</u> score from 0-36.
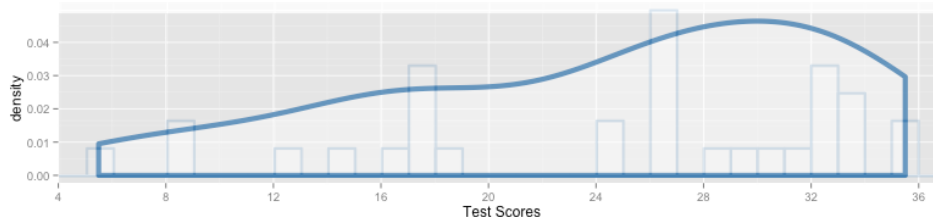
   This is a kernel density plot for the data displayed in the histogram. We'll learn about density plots in Activity #13.

   Using this continuous distribution (which we will call a pdf – a *probability density function*), how could we estimate the probability that a student scores 10 or less?

4. What does the <u>height of this curve</u> represent? What does the <u>area under this curve</u> represent? If I calculate the area under this curve from 0 to 36, what will I get? Why?

   Height of curve:

   Area under curve:

   Area under this curve from 0 to 36 = $\int_{x=0}^{36} f(x)dx =$

5. The pdf to the right models the <u>number of hours past noon</u> <u>that a student leaves for lunch</u>. What does this pdf tell us about the student's lunch habits?

$$f(x) = cx(1-x)$$

0.00   0.25   0.50   0.75   1.00
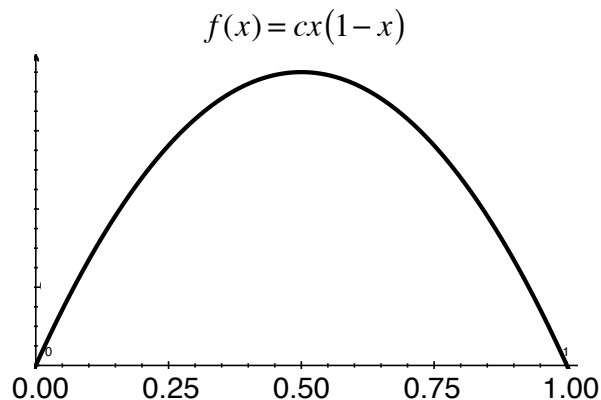
6. With this function, f(0.50) = 1.5. Does that represent the probability that the student leaves for lunch at 12:30?

7. You already know that in order to be a valid pdf, the area under the entire distribution has to sum to 1.00. Find the value of *c* that makes this a valid pdf. How do we know, in this scenario, that x can only range from 0 to 1?

8. In order to find probabilities (areas under the curve), we're going to need to integrate our pdf. Recall from Calculus that when we integrate, we're accumulating area under the curve. We write the integral of f(x) as F(x) and call it the cdf (the *cumulative density function*).

$$F(x) = \int f(x)\,dx$$

0.00   0.25   0.50   0.75   1.00
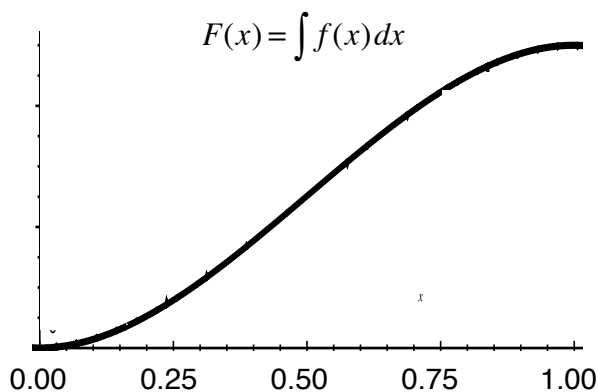
Then, to find the probability under this curve from a to b, we can simply evaluate the cdf at the endpoints and find the difference:

$$P(a < x < b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

9. Using our pdf and/or cdf, find the probability that the student leaves for lunch <u>before</u> 12:15.

P(X < 0.25) =

10. Find the probability that the student leaves for lunch <u>at or before</u> 12:15.

   $P(X \leq 0.25) =$

11. Find the probability that the student leaves for lunch <u>after</u> 12:45.

12. Find the probability that the student leaves for lunch <u>between</u> 12:15 and 12:45.

13. As you may already know, the <u>median</u> (or *50th percentile*) represents the value of X such that $P(X < \text{median}) = 0.50$. Find the median for X in this scenario.

14. Find the 25th percentile for X. What does it represent in this scenario?

15. Recall the formula for the expected value of a discrete random variable: $E[x] = \sum_{i=1}^{n} x_i P(x_i)$
   Why can't we use this formula for a continuous random variable?
   How could we transform this formula to apply it to continuous distributions?

16. Calculate the expected value of X in this scenario.

17. Suppose you see this student at 12:30 and realize the student has not already left for lunch. Given this information, calculate the probability that the student will leave for lunch before 12:45.

18. In actuarial science, a simple model used to describe mortality is: $f(x) = cx^2(100 - x)^2$, $0 \le x \le 100$
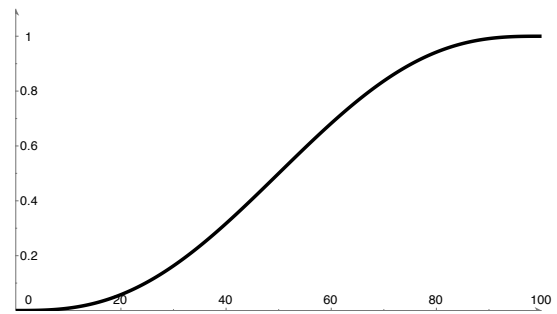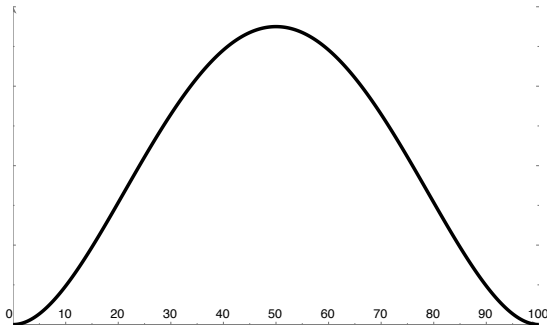
f(x) = _____

F(x) = _____

a) Find the value of c that makes this a valid pdf. Note that $cx^2(100 - x)^2 = cx^4 - 200cx^3 + 10{,}000cx^2$
   You can use wolframalpha.com to speed things up.

b) According to this model, what's the probability a person lives past the age of 80?

c) Given someone has lived to age 60, what's the probability that person lives past the age of 80?

d) Find and interpret the 90th percentile of this distribution.

19. Let's take one last look at the pdf for the time a student leaves for lunch. We know that P(X = a) = 0 for any value of a.

$$f(x) = 6x(1-x)$$

This means the probability a student leaves for lunch at exactly 12:30 is 0.00.

When asked for a probability such as P(X = 0.50), some students are tempted to substitute x = 0.50 into f(x) to find: f(0.50) = 6(0.5)(1 –0.5) = 1.5
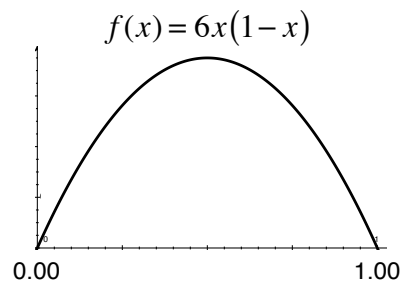
Obviously, this value is not a probability, since it's greater than 1.0. But what is it? What does f(a) represent?

Rather than calculating P(X=a), let's calculate the probability that X is very close to a. In other words, for very small, positive values of ϵ, let's find:
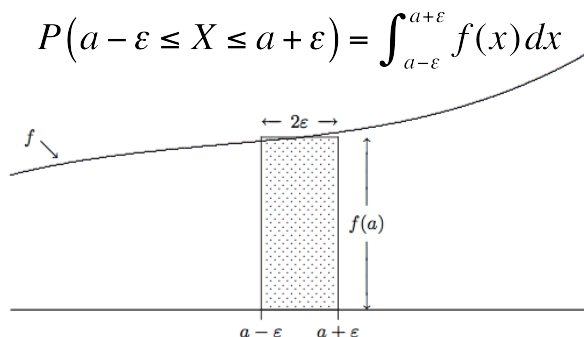
$$P\big(a - \varepsilon \le X \le a + \varepsilon\big) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,dx$$

As the above picture demonstrates, this integral is approximately equal to:

$$P\big(a - \varepsilon \le X \le a + \varepsilon\big) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,dx \approx 2\varepsilon f(a)$$

So we can think of f(a) as a <u>relative measure of the likelihood that X will be near a</u>. However, we cannot think of f(a) as a probability, since f(a) can be larger than 1.0.

So, in our example, f(0.50) = 1.5 tells us the relative likelihood that X is near 0.50.

Continuous Random Variables

pdf: $f(x)$ cdf: $F(x) = \int_{a}^{b} f(x)\,dx$

Expected value: $E[x] = \int_{a}^{b} xf(x)\,dx$ Median: Find b: $0.50 = \int_{min}^{b} xf(x)\,dx$

Hand calculator: $P(a \le X \le b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a)$

R: (We'll use an example where $f(x) = 3x^2$ and we want to know P(0 ≤ x ≤ 0.5)

```
> library(mosaic)              # Load the mosaic package
> f <- makeFun(3 *x^2 ~ x)     # Define the pdf
> F <- antiD( f(x) ~ x )       # Find antiderivative
> F(0.5) – F(0)                # Find probability
```

Other: http://www.wolframalpha.com can find definite and indefinite integrals easily