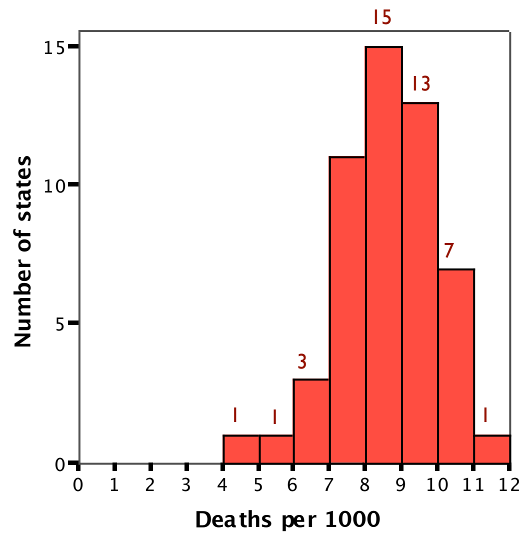
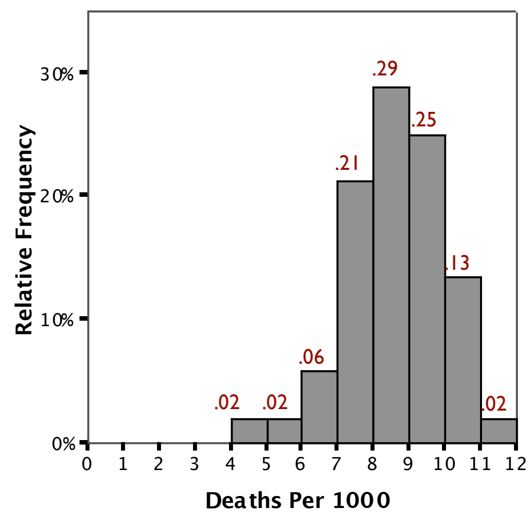
Activity #9:  Continuous Random Variables

Recall: **Discrete random variables** can take on a countable number of values.
**Continuous random variables** can take on *any* value within a specified region.

Situation: The following histograms show the 2001 death rates for each state in the U.S.  The values represent the number of deaths per 1000 people.
[*National Vital Statistics Report* (vol. 51) produced by the *National Center for Health Statistics*]




1) From the frequency histogram (on the left), how would you calculate the proportion of states that have death rates less than 7 per 1000?  How could you calculate the same value from the relative frequency histogram (on the right)?  If we select a state at random, what's the probability that we will select a state with a death rate of 7 or less?
*Note:  The data includes Puerto Rico and Washington, DC.*

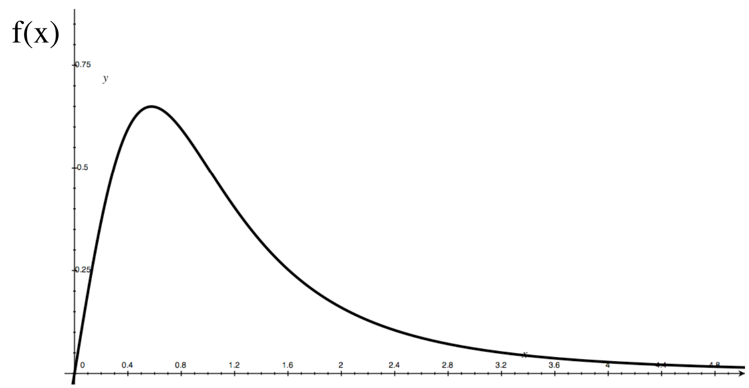
2) Remember that when the bars of a histogram have a width of one unit, we can calculate probabilities by finding the areas under the curve.  Verify this.

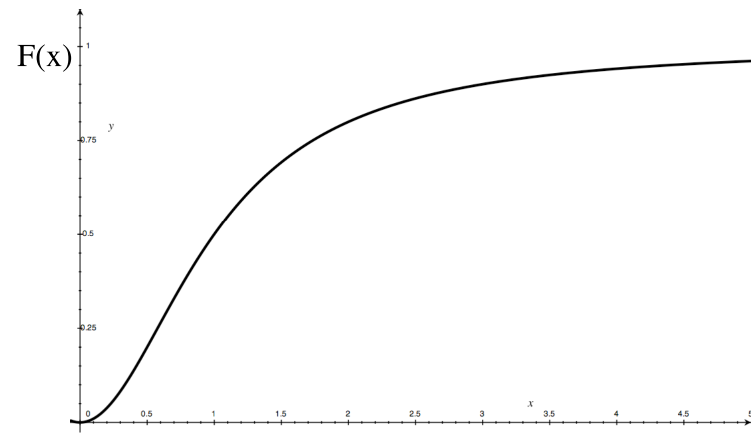
3) Is "death rate" a discrete or continuous random variable?  Draw a smooth curve over the relative frequency histogram.  Suppose we were interested in finding the probability of selecting a state with a death rate of 6.284 or less.  How could we calculate this probability?

When we have a continuous random variable, its probability model can be displayed via a function f(x). This function can be displayed as a formula or a distribution (curve).

> **IMPORTANT**: f(x) is not a probability! We must integrate f(x) if we want to calculate probabilities (because probability is equivalent to the area under a curve)

The probability that a product will last X years is modeled by the following probability mass function: $f(x) = \dfrac{2x}{(1+x^2)^2}$. Graphically, the pmf looks like this




> In order to be a valid pmf, two conditions must be met:
>
> A) The pmf must be positive for all valid values of x: f(x) ≥ 0 for all x
>
> B) The total area under the curve must be 1.0: $\int_{-\infty}^{\infty} f(x)dx = 1$

Let's verify condition B. We use u-substitution to integrate this function:

Let $u = 1 + x^2$. Then $du = 2xdx$ and $dx = \dfrac{1}{2x}du$. We substitute to get $\int_0^{\infty} \dfrac{2x}{u^2}\left(\dfrac{1}{2x}\right)du = \int_0^{\infty} u^{-2}du = \lim_{a\to\infty} \dfrac{-1}{1+x^2}\Big|_0^a = 1$.

The integral of f(x), denoted as F(x), is what we use to calculate probabilities. I.

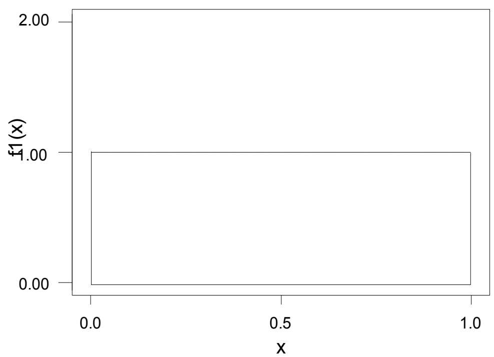
Scenario: Suppose a college student eats lunch at a time between noon and 1:00pm that varies from day to day. Let the random variable X=time (in hours) after noon that the person leaves for lunch. The following page displays four potential probability functions for X.

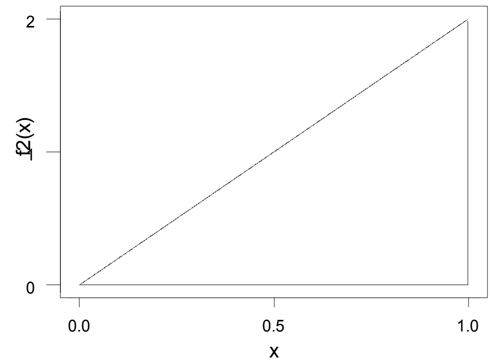
We're interested in the following three events:

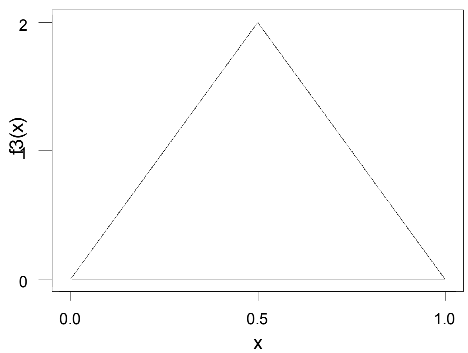
1) The student's lunch time will begin begore 12:15

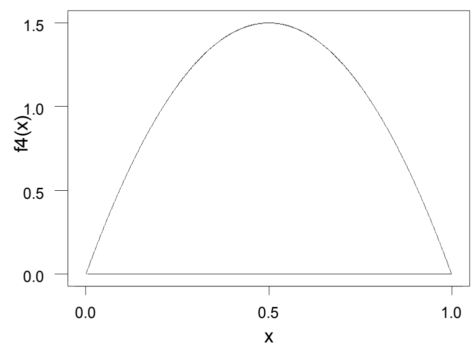
2) The student's lunch time will begin after 12:45

3) The student's lunch time will begin between 12:20 and 12:40




4) Looking at the distributions, decide which probability functions (1, 2, 3, or 4) represents the highest and lowest probability for each event. Complete the table:

| | Highest Probability | Smallest Probability |
|---|---|---|
| Before 12:15 | | |
| After 12:45 | | |
| Between 12:20 and 12:40 | | |

5) Using geometry, determine the relevant probabilities of these events for the 1st (top left) probability function.

| | Probability (area under curve) |
|---|---|
| Before 12:15 | |
| After 12:45 | |
| Between 12:20 and 12:40 | |

6) Using geometry, determine the relevant probabilities of these events for the 2nd (top right) probability function.

| | Probability (area under curve) |
|---|---|
| Before 12:15 | |
| After 12:45 | |
| Between 12:20 and 12:40 | |

7) Using geometry, determine the relevant probabilities of these events for the 3rd (bottom left) probability function.

| | Probability (area under curve) |
|---|---|
| Before 12:15 | |
| After 12:45 | |
| Between 12:20 and 12:40 | |

8) Calculating areas under the final pdf are more difficult. Suppose we know that the pdf can be expressed by: $f(x) = \begin{cases} cx(1-x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$.
Determine the value of C that makes this a valid pdf.

9) Our pdf, once we input the proper value of c and calculate its integral, simplifies to: $\left[3x^2 - 2x^3\right]_{x=a}^{x=b}$ Calculate the probabilities using this pdf.

| | Probability (area under curve) |
|---|---|
| Before 12:15 | |
| After 12:45 | |
| Between 12:20 and 12:40 | |

10) Using the same pdf, find the probability that the student leaves for lunch at exactly 12:15.

With continuous random variables, the probability of any one specific value P(X=k) is always zero.

11) The other pdfs can be expressed by the following:

$$f_1(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(x) = \begin{cases} 4x & 0 < x < 1/2 \\ 4(1-x) & 1/2 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_4(x) = \begin{cases} 6x(1-x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Another way to characterize a continuous probability distribution is with a **cumulative distribution function** (cdf).
This function is defined just as it was with discrete distributions: $F(x)=P(X\leq x)$.

Because probabilities correspond to integrals with continuous probability distributions, this function can also be written as $F(x) = \int_{-\infty}^{x} f(t)dt$,

where $f(x)$ is the pdf and $t$ is a dummy variable of integration.

12) Determine the cdf's for each of the four probability functions. Sketch these cdf's:

| |
|---|
| The cdf can lead directly to probability calculations, such as: $P(a<X<b) = F(b) - F(a)$. |

Calculating the expected value and variance of continuous random variables is rather straightforward:

| |
|---|
| **Definition:** The *mean* or *expected value* is defined to be $\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx$. |
| **Rules for Expected Values:** $E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x)dx$ When *h*(X) is linear: E(aX+b)= aE(X) + b |
| **Variance:** $V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$ or the same calculation shortcut: $V(X) = E(X^2) - [E(X)]^2$ |

13) Find the expected value of the first three distributions

| |
|---|
| The (100*p*)th *percentile* of a probability distribution is the value (call it $k_p$) such that $P(X \le k_p) = p$. In particular, the *median* is the 50th percentile, the *lower quartile* is the 25th percentile, and the *upper quartile* is the 75th percentile. |

14) Determine the median, lower quartile, and upper quartile for the 2nd distribution.