Activity: Continuous Distributions

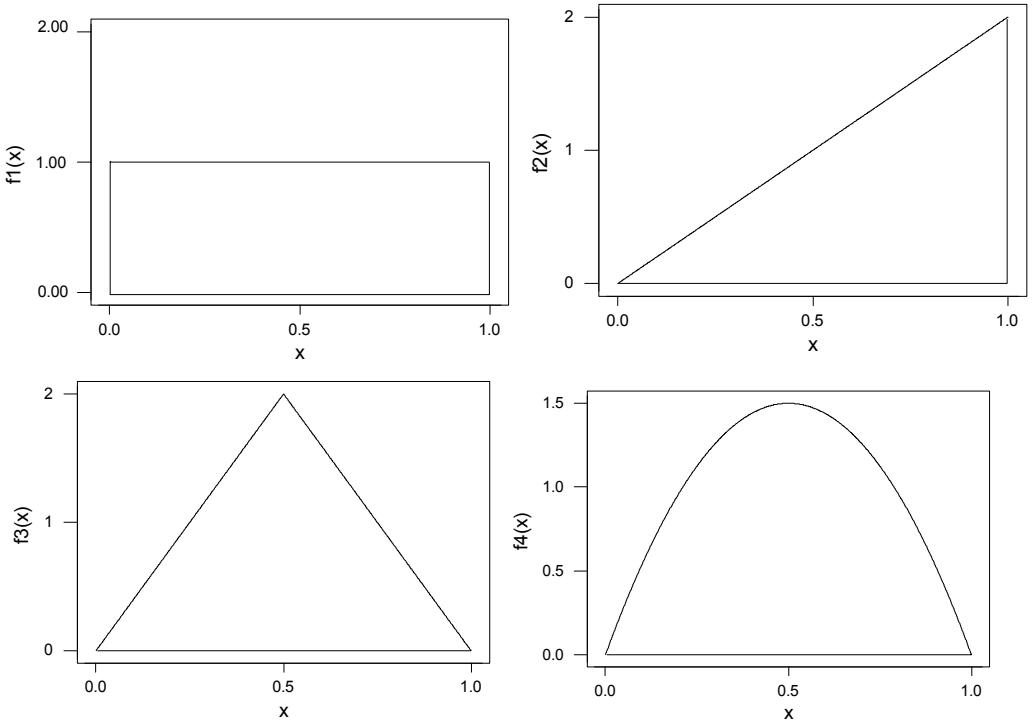
Concepts: Fundamentals of continuous distributions. Includes probability density functions (pdfs), cumulative distribution functions (cdfs), expected value, variance, computing probabilities using pdfs and cdfs, and percentiles (including the median).

Prerequisites: The student should be familiar with integral and differential calculus and with basic probability concepts from their study of discrete probability distributions.

Concepts and Definitions: A *continuous random variable* can take on *any* value in some *interval*. If a discrete random variable can take on many possible values, then it can be approximated by a continuous one.

A continuous random variable is characterized by its *probability density function* (pdf). The probability that the random variable takes on a value in any interval corresponds to the *area* under the pdf over that interval. Analytically, if $f(x)$ represents the pdf of a continuous random variable X, then $P(a<X<b) = \int_a^b f(x)dx$. Any pdf must satisfy two properties: it must be non-negative, and its integral over the entire real line must equal one: $f(x)\geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.
Please see the definition on page 165 of your text.

Scenario: Random Lunch Times
Suppose that a businessperson leaves for 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. Consider the following four probability density functions for X:



Consider the following three events:

- that the person's lunch time will begin before 12:15
- that the person's lunch time will begin after 12:45
- that the person's lunch time will begin between 12:20 and 12:40

(a) With which of the four pdf's do you think the probability will be highest, and with which do you think the probability will be smallest. Do not perform any calculations yet, but base your guesses on the appropriate areas under the curves represented by the pdf's. [Remember that X is measured in hours after noon.] Fill in your guesses (numbered 1, 2, 3, or 4) in the following table:

	highest probability	smallest probability
before 12:15		
after 12:45		
between 12:20 and 12:40		

(b) Use geometry to determine the relevant areas, and therefore probabilities, of these events for pdf #1 (pictured in the upper left):

before 12:15 after 12:45 between 12:20 and 12:40

(c) Use geometry to determine the relevant areas, and therefore probabilities, of these events for pdf #2 (pictured in the upper right):

before 12:15 after 12:45 between 12:20 and 12:40

(d) Use geometry to determine the relevant areas, and therefore probabilities, of these events for pdf #3 (pictured in the lower left). [Hint: For the last probability, find the area of the complement and then subtract from one.]

before 12:15 after 12:45 between 12:20 and 12:40

The probability density function pictured in the lower right can be expressed as:

$$f(x) = \begin{cases} cx(1-x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

(e) Determine the value of c necessary for this to be a legitimate pdf; i.e., for the total area under the curve to equal one. [*Hint*: Use calculus.]

(f) Use calculus to find the three probabilities asked about above for pdf #4:

before 12:15

after 12:45

between 12:20 and 12:40

(g) Choose any two of these four pdf's, and find $P(X \leq 1/4)$. How does it compare to $P(X<1/4)$. Explain why this makes sense.

(h) Determine $P(X=1/4)$ for any two of these pdf's. Explain why this makes sense both geometrically and with calculus.

With continuous random variables, the probability of any one specific value $P(X=k)$ equals zero. This in turn establishes that $P(X \leq k)=P(X<k)$, so with continuous random variable we need not worry about strict vs. non-strict inequalities. In particular, plugging a value into the probability density function does not provide the probability of anything.

(i) Determine functional expressions for the pdf's pictured in #1, #2, and #3 above.

You should find that functional expressions for these pdf's are:

$$f_1(x)=\begin{cases}1 & 0<x<1\\ 0 & \text{otherwise}\end{cases} \qquad f_2(x)=\begin{cases}2x & 0<x<1\\ 0 & \text{otherwise}\end{cases}$$

$$f_3(x)=\begin{cases}4x & 0<x<1/2\\ 4(1-x) & 1/2<x<1\\ 0 & \text{otherwise}\end{cases} \qquad f_4(x)=\begin{cases}6x(1-x) & 0<x<1\\ 0 & \text{otherwise}\end{cases}$$

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: $F(x)=P(X\le x)$. 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.

(a) Determine the cdf for each of these pdf's. Also draw sketches of the cdf's. [*Hints*: In all cases, be sure to specify what the cdf outputs for all real number inputs. Use geometric arguments to help make sense of the integrals, especially in case #3.]

case #1:

case #2:

case #3:

case #4:

Consider a fifth probability distribution, whose cdf is given by $F_5(x)=\begin{cases}0 & x\le 0\\ x^4 & 0<x<1\\ 1 & x\ge 1\end{cases}$.

(b) Sketch this cdf.

(c) Determine P(X<1/4), P(X>3/4), and P(1/3<X<2/3) directly from this cdf.

(d) Determine the pdf, and sketch it.

You should have found that the cdf of a continuous distribution can lead directly to probability calculations, such as P(a<X<b)=$F(b)$-$F(a)$. You should also have noted that the pdf may be found from the cdf from $f(x) = F'(x)$.

As with the discrete case, one can calculate the expected (mean) value and variance of a continuous random variable: $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$, $V(X) = E\left[(X-\mu)^2\right] = E(X^2) - [E(X)]^2$, where one find the expected value of a function of the random variable by the expression: $E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$.

(e) For the five probability distributions that you have analyzed above, make a guess for the mean (expected) value of each. Record your guesses in the table below. Then use calculus to determine these expected values, record them in the table as well, and comment on how well your guesses did.

distribution	1	2	3	4	5
guess for mean					
expected value					

The interpretation of expected (or mean) value is the same for continuous distributions as for discrete ones: the long-term average value that would result from repeating the process over and over.

(f) Judging from the variation in the graphs of the pdf's, make a guess for the relative ranking of their standard deviations. Record your guesses (by case number) in the following table.

Then calculate their standard deviations, record their values in the table, and comment on how well your guesses did.

	smallest	next smallest	middle	next largest	largest
guess					
std dev value					

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

(g) Determine the median, lower quartile, and upper quartile for the fifth probability distribution described above. Comment on how the mean and median compare.