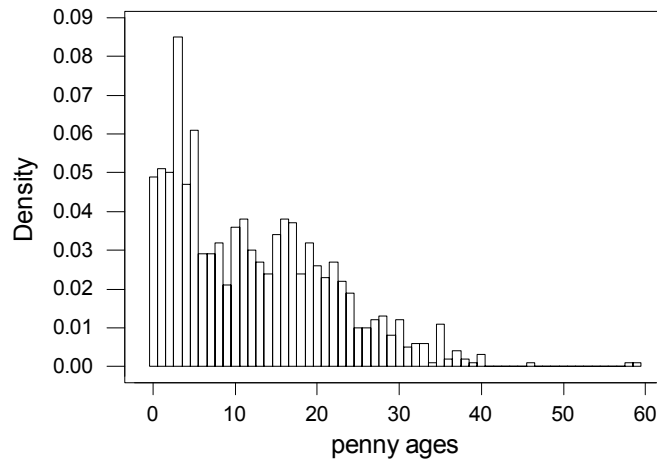


Stat 321 - Day 17, 18
Continuous Probability Distributions

Example: Penny Ages

A few years ago I collected pennies as they came into my possession, and I recorded the age (in years) of the first 1000 pennies that I encountered. These ages can be found in the Minitab worksheet `pennies.mtw`. A *density histogram* of these ages appears below:



When describing the distribution of data, it is helpful to comment on *shape*, *center*, and *spread*.

- The shape of this particular histogram is said to be *skewed to the right* (or skewed positively) because the data are clustered around smaller values and extend further along in the tail toward higher values.
- The center is typically measured by the *mean* (arithmetic average) and/or *median* (the 50th percentile, the value such that half fall above and half fall below).
- Spread is commonly measured by the *standard deviation* or *inter-quartile range*. The standard deviation is roughly (but not technically) the average deviation from the mean, while the inter-quartile range is the difference between the upper and lower quartiles (also known as 75th and 25th percentiles, respectively).

(a) Use Minitab's `describe` command to calculate these descriptive statistics for these penny age data. [Note: The upper and lower quartiles are reported as Q3 and Q1; you must subtract to find the IQR.] Record them below:

mean: median: standard deviation: IQR:

(b) Determine what proportion of these 1000 pennies are less than 10 years old. [Hint: Either use `tally c1` and count yourself, or let `c2=(c1<10)` and then `tally c2`.]

(c) Determine what proportion of these 1000 pennies are at least 20 years old.

(d) Determine what proportion of these 1000 pennies are between 5 and 15 (inclusive) years old.

Consider the random variable $X =$ age of a penny selected at random from these 1000 pennies. X is technically a discrete random variable, but since there are so many possible values (0-60 in this case) it can be approximated by a continuous random variable.

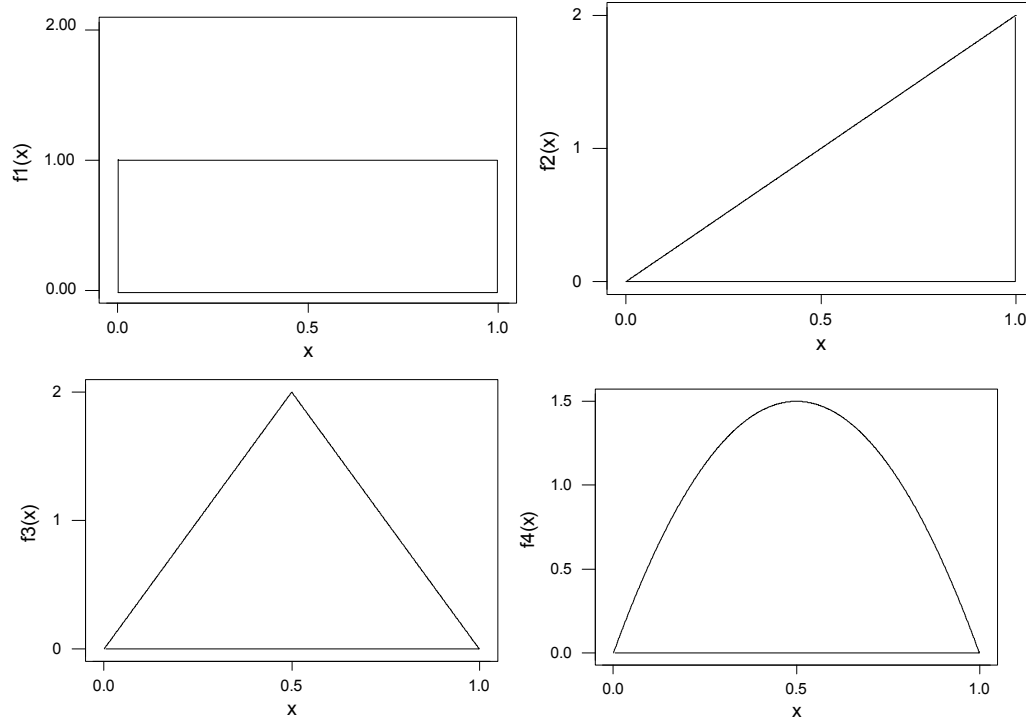
- (e) Draw a smooth curve over the density histogram above to approximate the probability distribution of penny ages.
- (f) Do you agree that your answers to (b)-(d) correspond geometrically to the areas of the histogram bars for the relevant values? Then would the areas under your smooth curve between the relevant values approximate those probabilities?

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 $\int_{-\infty}^{\infty} f(x)dx = 1$.

Example: 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

(g) 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		

(h) 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

(i) 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

(j) 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:

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

(k) 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.]

(l) 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

(m) 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.

(n) 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.

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