RANDOM VARIABLES

A key idea in dealing with uncertainty is the idea of *random variable*. The time it takes the balloon to fall can be considered as a random variable.

One definition (often given in probability textbooks) of a random variable is "A real-valued function defined on a sample space." This is technically correct, but what is often more helpful in applications is to think of a random variable as *a variable that depends on a random process*. Here are some examples to help explain the concepts involved:

1. Toss a die and look at what number is on the side that lands up. *Tossing the die is an example of a random process; the number on top is the random variable.*

2. Toss two dice and take the sum of the numbers that land up. *Tossing the dice is the random process; the sum is the random variable.*

3. Toss two dice and take the product of the numbers that land up. *Tossing the dice is the random process; the product is the random variable.*
Examples 2 and 3 together show that *the same random process can be involved in two different random variables.*

4. Randomly pick (in a way that gives each student an equal chance of being chosen) a UT student and measure their height. *Picking the student is the random process; their height is the random variable.*

5. Randomly pick (in a way that gives each student an equal chance of being chosen) a student in this class and measure their height. *Picking the student is the random process; their height is the random variable.*
Examples 4 and 5 illustrate that *using the same variable (height) but different random processes (in this case, choosing from different populations) gives different random variables.*

6. Measure the height of the third student who walks into this class. *What is the random process?*
In Example 5, the random process was done deliberately; in Example 6, the random process is one that occurs naturally. *Can you explain how the different random processes make these two random variables different?*

7. Toss a coin and see whether it comes up heads or tails. *Tossing the coin is the random process; the variable is heads or tails.*
This example shows that *a random variable doesn't necessarily have to take on numerical values.*

8. The time it takes for an IF shuttle bus to get from 45th and Speedway to the Dean Keaton stop is a random variable.
Whoa, you may say -- where's the random process? I've given this example this way precisely because random variables are often defined in this way. The random process here is "implicit" (at least, for those used to defining random variables in this way*). What is really meant is: "Randomly pick an IF shuttle bus run and measure the time it takes to get from 45th and Speedway to the Dean Keaton stop." So the random process is picking the shuttle bus run, and the random variable is the time measured.*

9. a. The height (t minutes after its release) of an object tossed straight up from initial height h_0 with initial velocity v_0.

b. The height we measure (t minutes after its release) of an object tossed straight up from initial height h_0 with initial velocity v_0.

How are these two random variables different? What is the random process involved in each?

CAUTION: If you look in a dictionary, you may find that the first definition of "random" is something like, "Having no specific pattern or objective; haphazard." *This is NOT the technical meaning of random that is used in probability and statistics.*

Here are some examples of processes that *are* random in the technical sense:

A. We consider a process such a tossing a die or a coin to be random.

B. When we talk about randomly picking a UT student or randomly picking an IF shuttle run, we mean using a process that gives each possible UT student (or IF shuttle run) an equal chance of being chosen. We can imagine (but only imagine), for example, numbering all UT students 1, 2, 3, etc. and having a huge die with as many sides as there are UT students. Tossing the die and taking the student whose number came up would be a way of randomly picking a UT student. In practice, random selections such as this are made by replacing our imaginary huge die by a computer program (called a pseudo-random number generator, or random number generator for short) that is designed to give essentially the same result.

ADDITIONAL CAUTION: Examples, however, might lead one to believe that a random process always has to give every outcome and equal chance of happening. This is not the case. We could imagine, for example, a "loaded" or "biased" die that was made so that one of the sides came up more frequently than the others. We would still consider tossing this die to be a random process.

One important aspect of a random process is that *although there may be (and usually is) a pattern in the long run, there is no way of knowing in advance the result of one occurrence of the process.* In other words, *the result of one occurrence of a random process is uncertain, but we can (at least in theory) say something about the long-term behavior (that is, what happens over many, many occurrences) of the process.*

Examples of random variables that people are interested in studying include the following (*using the convention mentioned in example 8*):

- The time a certain type of light bulb will last until it burns out.
- The yield per acre of a field of wheat.
- The birth weight of a child.
- The length of time a person lives.
- The Dow-Jones Index.
- The concentration of ozone in the air.

Discrete and continuous random variables

A numerical-valued random variable is called *discrete* if the set of values it can take on can be labeled by some subset of the integers. It is called *continuous* if the set of values it can take on form one or more intervals of real numbers. (There are some random variables which are "mixed" – their set of values might, for example, consist of an interval and some finite number of values outside the interval.)

Exercise: 1. Classify the random variables in examples 1 – 6 and 8-9 above as discrete or continuous.

Notational convention: I will usually use capital letters to denote a random variable and lower case letters to refer to values the random variable may take on. (Analogy: f refers to a function, and f(x) refers to the value of that function at x.)

The values of a random variable that have actually been measured are sometimes called *observed values* or *realizations*

Probability Distributions

Every numerical-valued random variable has an associated *probability distribution*. The intuitive idea is that the probability distribution of the random variable X tells us (via a picture or function or table some other means) which values the X can takes on, *and* which values or ranges of values are more or less likely to occur. In particular, it can give us an idea of the pattern and variation associated with the random variable.

Example: If a fair die is thrown, we could describe the probability distribution of the outcomes by a table:

Outcome	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

A loaded die would have different probabilities in the second line.
X – for example,

Outcome	1	2	3	4	5	6
Probability	1/3	1/3	1/12	1/12	1/12	1/12

Two different random variables might have the same distribution – if they arise from different random processes or the variables have different meaning, they are considered different. Can you think of an example with the same distribution as the outcomes of the fair die?

Probability Density Functions

These are a common and useful way to describe probability distributions. The *probability density function of the random variable X* (*pdf* for short) is usually called f_X. It is defined slightly differently depending on whether the random variable is discrete or continuous.

For a *discrete* random variable, the probability density function is also called the *probability mass function*. It is defined by $f_X(x) = P(X = x)$, the probability that the random variable takes on the value x. The probability distribution can be pictured by drawing a vertical line segment above x with length equal to $f_X(x)$

Example: Example 1 above (toss a die; X = the number showing)

Probability mass function: $f_X(1) = f_X(2) = f_X(3) = f_X(4) = f_X(5) = f_X(6) = 1/6$

Probability distribution:

Exercise: 2. Find the probability mass function and sketch the probability distribution of the random variable in Example 2 above. (Toss two dice; X = the sum of the values showing.)

For a <u>continuous</u> random variable, the graph of f_X pictures the distribution of X. To make this more precise: f_X is the function with the following property:

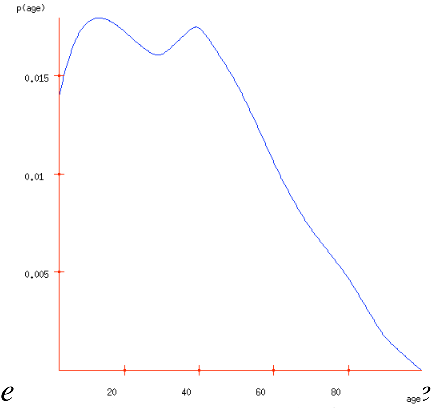
For all intervals on the real line, the probability that X is in that interval is the area under the graph of f_X above that interval. i.e., for every interval I in the real line (I could be of the form (a,b), (a, ∞), (-∞, b), or (-∞,∞)),

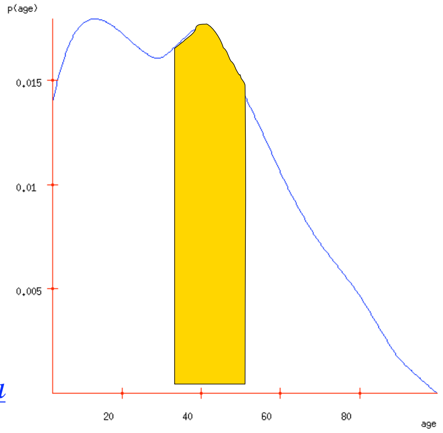
P(X ∈ I) = the area above I and below the graph of f_X = $\int_I f_X(x)dx$

Note:
1. "X ∈ I" is short for "the value of X is in I"
2. The usual convention is that f_X is defined for all real numbers. That just means that its value is zero on intervals where no values of X occur.

Example: American ages in 2000.

P(30 ≤ Age ≤ 50) = $\int_{30}^{50} f_X(x)dx$



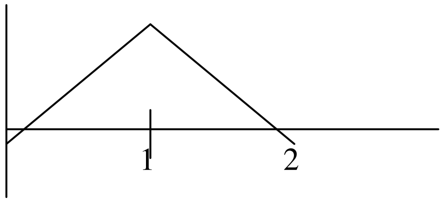


(The [illegible] ew Moore, http://www.cs.cmu.edu/~awm/tu

Exercises:

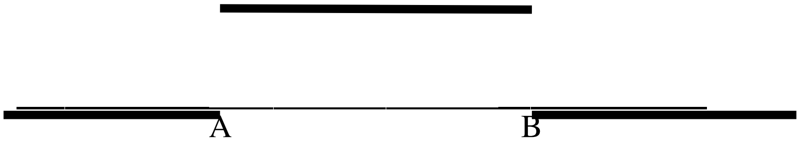
-> Be sure to give reasons in all problems and exercises.

3. If the graph of the probability density function (*pdf* for short) is



a. What is the area of the triangle?
b. What is the height of the triangle?
c. What is P(0 < X < 1)? P(0 < X < ½)? P(X > ½)? P(|X – 1| > ½)?
d. Find a formula for the pdf

4. The graph of the pdf of the *uniform* distribution on the interval [A, B] looks like this:



a. What is the y-coordinate of the upper line segment? (The lower line is the x-axis)
b. Find a formula for the pdf of this distribution.

Empirical Distributions:

In practice, we may not always be able to know the pdf of a random variable, but sometimes we can get an idea of what the distribution looks like from an *empirical distribution*. This means a *histogram* (or other graph such as a stem-plot) using data collected from a suitable sample of data.

Example: Here is a histogram that gives the empirical distribution of the heights of students in a particular beginning statistics class:

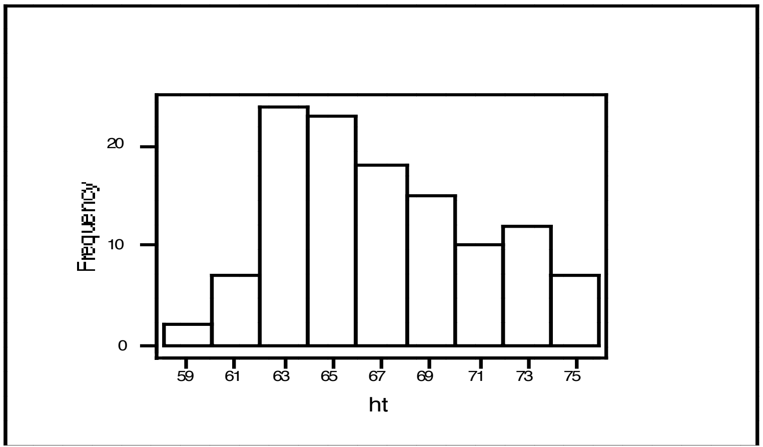


Figure 1

The first bar shows the number of people whose height is between 58 and 59.5 inches (inclusive); the second bar, the number of people whose height is between 60 and 61.5 inches, and so on. (*Please note*: There are lots of different conventions for drawing histograms, so please use caution in interpreting them.)

Based on this histogram and what you know about peoples' heights, what would you guess the proportion of males and females in this class to be? How would you expect the histogram to differ if that proportion were different?

The histogram above is an example of a *frequency histogram*. Sometimes we may use a *density histogram*.

- In a *frequency* histogram, the *height* of the bar above each interval shows the *number* of values in the interval.
- In a *density* histogram, the *area* of the bar above each interval shows the *proportion* of values in the interval.

Here is a *density histogram* for the heights of students in the same class:

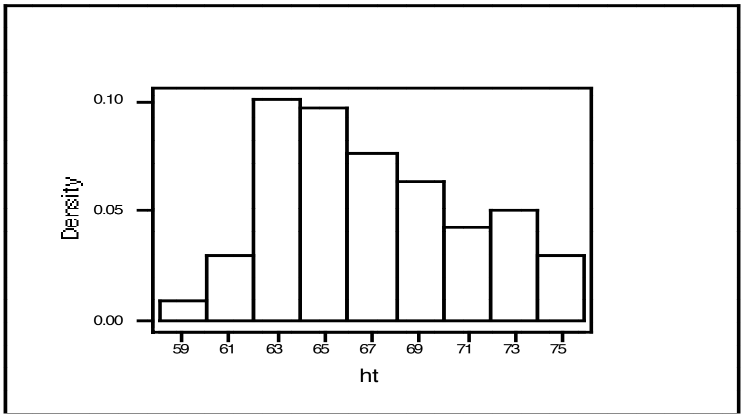


Figure 2

What is the same about the two histograms? Why?
What is different?
How can you get the height of a bar in the density histogram from the height of the corresponding bar in the frequency histogram? (Hint: There were 118 students in the class.)

Problems: *->Be sure to give your reasons in all problems.*

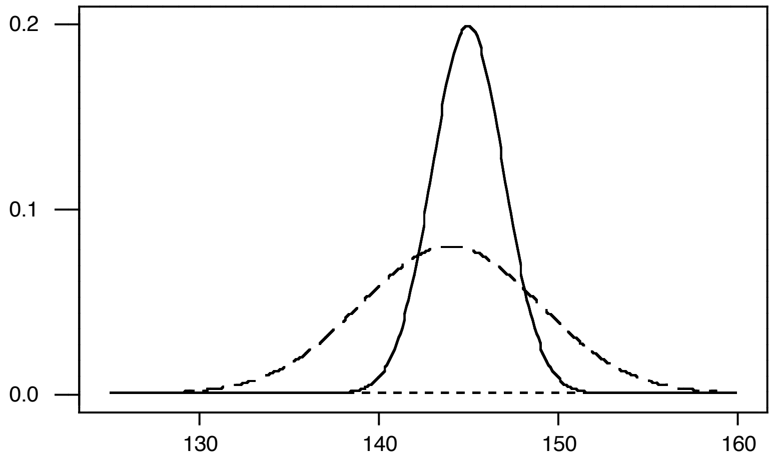
5. Sketch a possible distribution for each of the random variables in Examples 4 and 8. Explain why your sketches have the features you have sketched.

(*Next problem on next page)*

6. A physician friend has asked for your advice on whether or not to prescribe a new medication for lowering high blood pressure. She has obtained the following diagram based on data collected in a large clinical trial that compared the new drug with an older drug for the same purpose. It shows the distribution of the systolic blood pressure (the top number in blood pressure reports) of people taking the new drug (dashed lines) and the distribution of the systolic blood pressure of people taking the old drug (solid lines). She also tells you:

- Patients taking the drugs all started out (before taking their respective drugs) with systolic blood pressure between 145 and 155 mmHg (milligrams of mercury), with average (mean) reading 150 mmHg.
- Achieving a drop of 5 - 6 mmHg in systolic blood pressure is considered worthwhile.
- An increase in blood pressure can mean an increased risk for stroke and heart attack.
- The average (mean) systolic blood pressure for patients taking the new drug is 144 mmHg.
- The average (mean) systolic blood pressure for patients taking the old drug is 145 mmHg.

What would you tell her?



old drug = solid, new drug = dashed

7. A *normal* (or *Gaussian*) distribution is one whose probability density function (pdf) has the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)},$$

which is often more convenient to write as

$$\text{f(x)} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Normal distributions occur frequently, in lots of applications. (We'll be discussing this more later.) The numbers μ ("mu") and σ ("sigma") are called *parameters*; each choice of parameters gives a different normal distribution. (Analogy: In talking about lines in the plane, we often talk about the form y = mx + b. The graph of any such equation is a line; m and b are parameters; each choice of parameters gives a different line. Caution: This is different from, although similar to, the use of the word "parameter" in talking about parametric equations. Can you see both the difference and the similarity?)

Note: σ must be positive, since a probability density function is always non-negative and a power of e is always positive.

Suppose the random variable U has pdf $f_U(u) = \frac{1}{\sqrt{\pi}}\exp\left(-(u+2)^2\right)$. Explain why U is normal. (Hint: Find μ and σ to express $f_U(u)$ in the form of the pdf of a normal random variable.)

8. You have probably heard of the normal distributions described as "bell curves." The purpose of this problem is to help you understand why.

a. Recall that the graph of a function is said to be *symmetric about the line x = a* if $f(x_1) = f(x_2)$ whenever x_1 and x_2 are equal distance from a and both in the domain of f. In other words, f(a-x) = f(a+x) for every x with both a-x and a+x in the domain of f. Find a number a with the property that the normal distribution curve with parameters μ and σ is symmetric about the line x = a, and use the definition to show that the curve indeed is symmetric about this line. (Remember: Reasons required.)

b. Hand in. *Use calculus* to find

i. All local maxima and minima of the normal distribution pdf with parameters μ and σ. (Be sure to show whether you really do have a maximum or a minimum!)

ii. The points of inflection of the normal distribution pdf with parameters μ and σ.

c. Explain how parts a and b tell you that a normal distribution is bell shaped.

d. The *standard normal* distribution is the normal distribution with parameters μ = 0 and σ = 1. Let g(x) be the pdf of the standard normal distribution.

i. Verify that, if f(x) is the pdf of the the normal distribution with parameters μ and σ, then $\text{f(x)} = \frac{1}{\sigma}g\left(\frac{x-\mu}{\sigma}\right)$. In other words, you can get f(x) by first subtracting μ from x, then dividing the result by σ, then taking g of that, then dividing the result by σ.

ii. Use the above to explain, step-by-step, how to get the graph of f from the graph of g. That is, subtracting μ from x does what to the graph? Dividing x - μ by σ does what to the graph? Dividing $g\left(\frac{x-\mu}{\sigma}\right)$ by σ does what to the graph?