M 358K, Spring 06

REVIEW (AND MORE) OF RANDOM VARIABLES

In calculus you learned things like:

> *If an object is tossed straight up from initial height h_0 with initial velocity v_0, then its height at time t seconds after being released is*
>
> $h(t) = h_0 + v_0t - gt^2$.

If you read the textbook carefully and/or listened carefully in lecture, you might have heard a qualification like, "if you ignore air resistance." In reality, air resistance can be important -- in most cases, it results in a "terminal velocity" which the object cannot exceed. However, finding an equation of motion taking air resistance into account presents real problems, since the air resistance depends on the mass and also the shape of the object. (And if the object is a parachutist, its shape may be changing as it falls!) In fact, can you really know the initial velocity exactly? The initial height? And what if there's wind?

As this example shows, in real life we often don't have *deterministic* (that is, exact) formulas like $h(t) = h_0 + v_0t - gt^2$. Instead, we have to deal with *approximations* and *uncertainty*. So we need *stochastic* (that is, *probabilistic*) methods.

A key idea in probabilistic methods is the idea of *random variable*. The height of our *real* falling object can be considered as a random variable -- we may be able to find a formula taking into account air resistance that will give an *approximate* description of the object's motion, but there are still uncertain factors such as wind and our ability to measure the initial velocity or determine the effect of air resistance exactly.

Your probability textbook may have defined a random variable as “A real-valued function defined on a sample space.” This is technically correct, but what is often more helpful in statistics is to think of a random variable as *a variable that depends on a random process*. Here are some examples:

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 a UT student and measure their height. *Picking the student is the random process; their height is the random variable.*

5. Randomly pick 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 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.

c. The formula we obtain, taking everything we can into account, for the height (t minutes after its release) of an object tossed straight up from initial height h_0 with initial velocity v_0.

How are these three random variables different? What is the random process involved in each? [Hints: They're complex.]

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.

These 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 to breakdown of a computer.
- 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.
- How much time is required to read a disk sector into main memory.

As you have seen in Probability and in Section 1.1 of the M 358K textbook, one way to give information about (that is, to say something about the long-term behavior of) a random variable is to show its *distribution*. (In the vocabulary of probability, this is the graph of the *probability density function.)* In Probability, you usually started with a formula for the probability density function and got the distribution from that. In this course, we will often use *empirical distributions*. That means that they are based on actual data plotted in a *histogram* or other graph such as a stem-plot

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

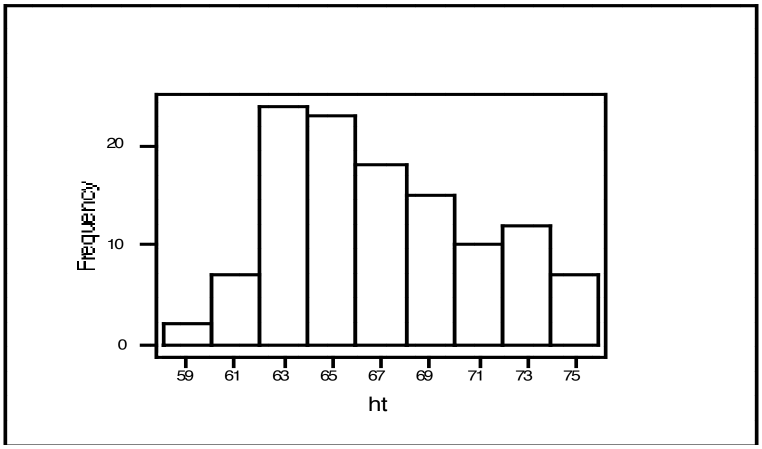

Figure 1

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 bar above each interval shows the *number* of values in the interval.
- In a *density* histogram, 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:

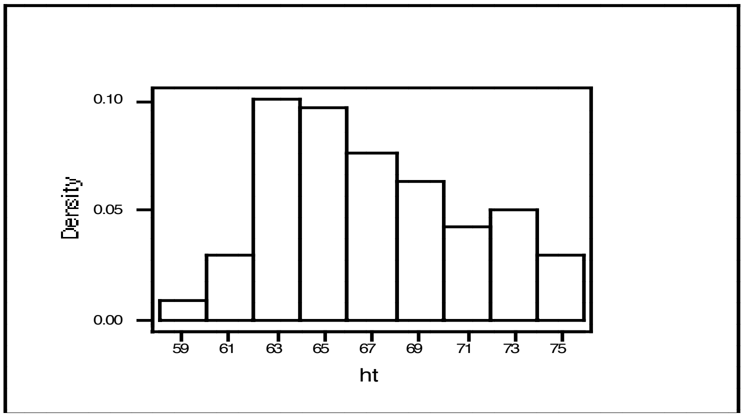

Figure 2

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.)

In the above example, we are only interested in the students in the particular class, so the density histogram is a picture of the probability density function of the random variable in question. If instead we were interested in the heights of all UT students, we would have a related but different (because the random process is different) random variable. Another perspective that is important in statistics is to consider the students in the class to be a *sample* of all UT students. The set of all UT students would then be called the *population*. Using this language, we would talk about the distribution of heights of all UT students and the distribution of the heights of the students in the sample. These usually will not be the same.

Exercises:

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.