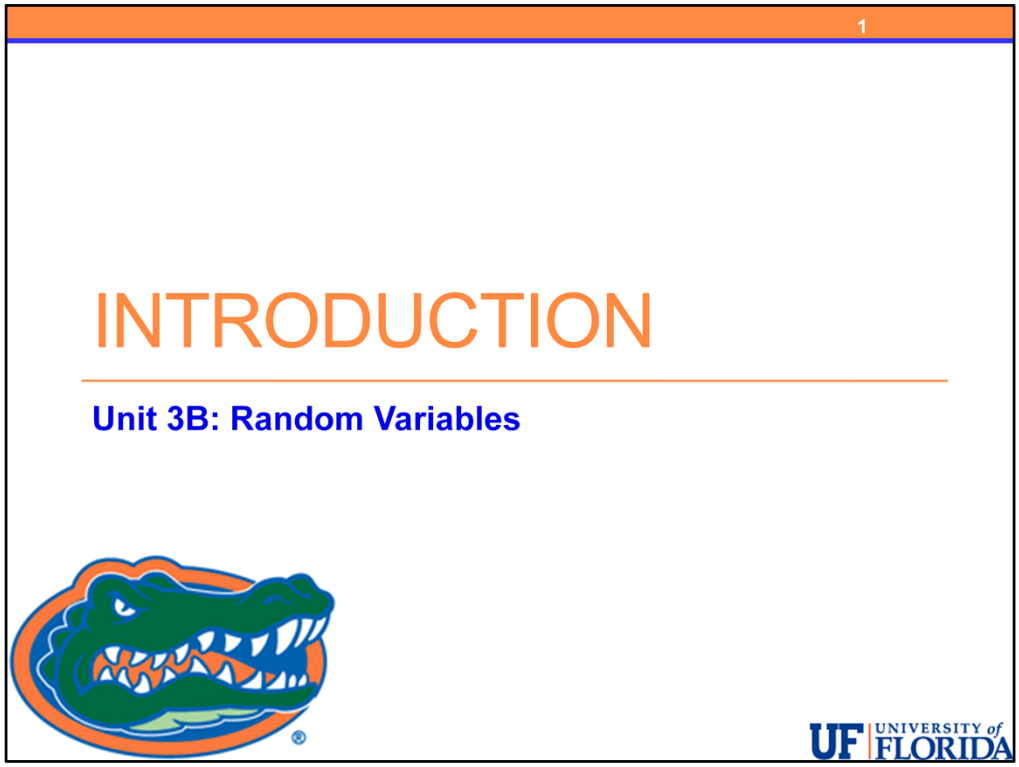


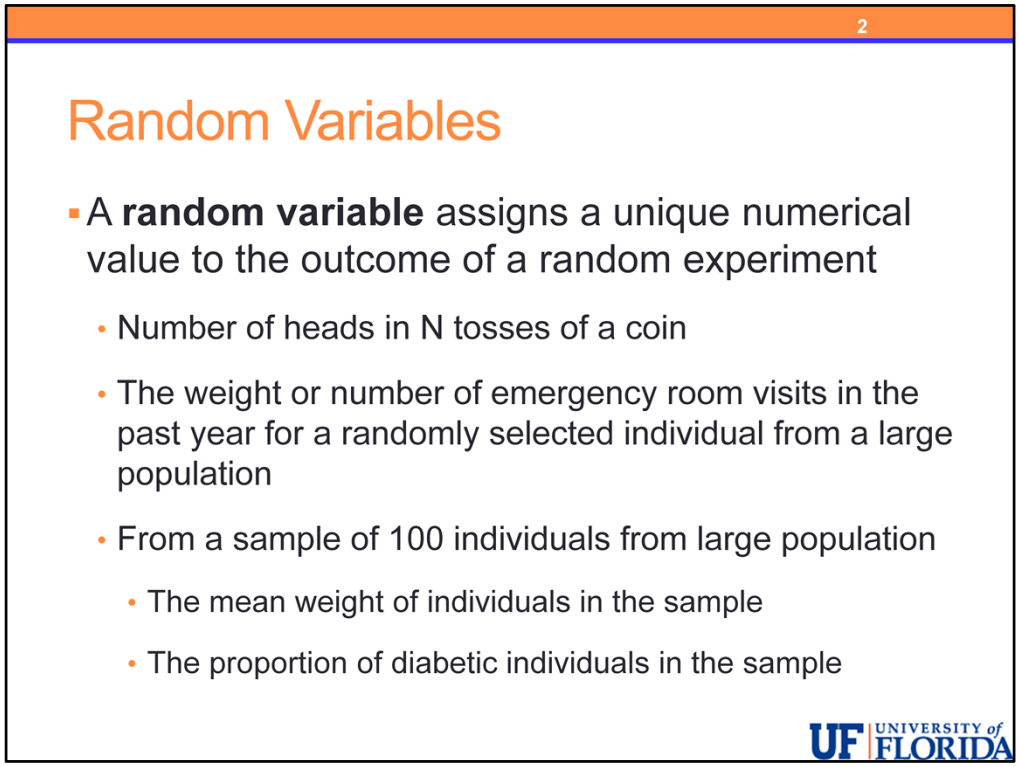
Now we will define, discuss, and apply random variables. This will utilize and expand upon what we have already learned about probability and will be the foundation of the bridge between probability and inferential statistics.

We will develop the theoretical background for some simple situations and then we won't focus too much on the theory once the mathematics becomes more difficult.

Our goal is to give you a good feel for why probability and random variables are useful on their own and then use a few simple examples to help you understand how statistics really works.

Finally, at the end of the semester, we will learn to apply the process of statistical inference using numerous common statistical tests for hypotheses involving one and two variables, such t-tests, analysis of variance, and chi-squared tests.

The foundations we have been building will be important to the development of p-values and confidence intervals which, you may already know, are the basis of our ability to draw conclusions about our population from data.



We need to define what we mean by a random variable. In statistics, a random variable assigns a unique numeric value to the outcome of a random experiment.

The term "random experiment" is very broad, anything from tossing a coin to picking an individual from a large population or even picking a sample of size 100 individuals from a large population. Each of these could be considered one "trial" of a random experiment.

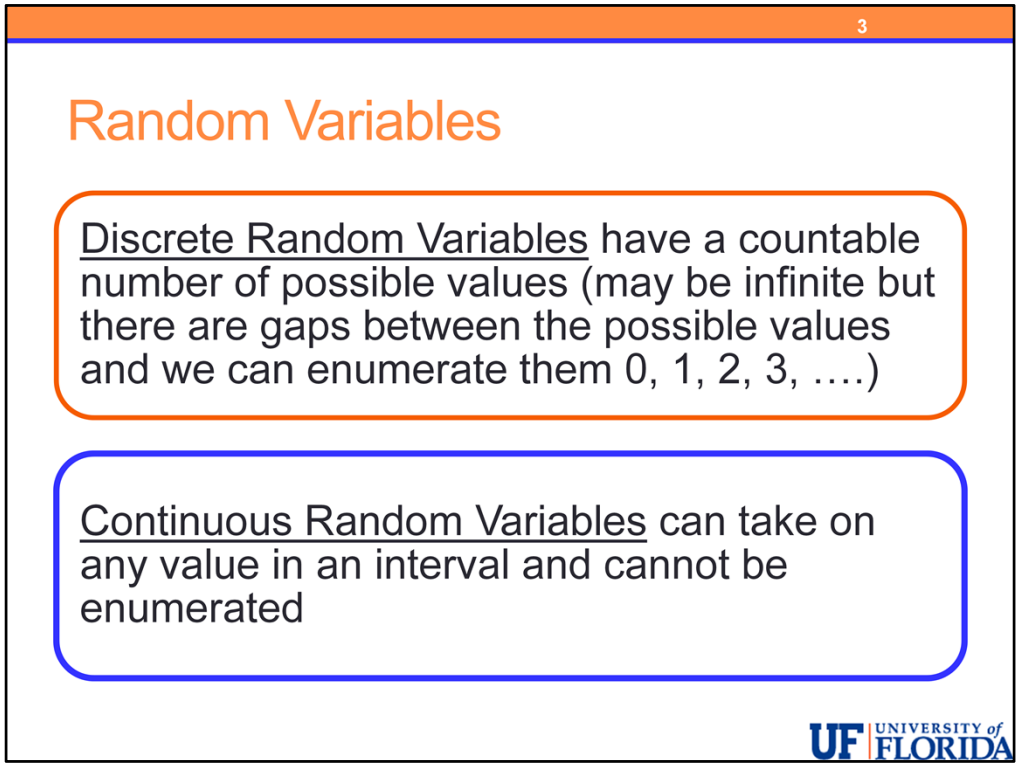
The random variable must be a numeric measure resulting from the outcome of a random experiment.

If we toss a coin, the random variable might be X = the number of heads.

If we pick one person from the population, the random variable might be X = the weight of the person in pounds or Y = the number of emergency room visits in the past year.

If we pick 100 individuals from a large population, the random variable might be X = the number of diabetics in our sample, p-hat = the PROPORTION of diabetics in our sample, x-bar = the average weight of individuals in our sample, or y-bar = the average number of emergency room visits in the past year in our sample.

All of these measures are what statisticians would call random variables. Their values are not known but, under certain assumptions, their processes can be studied. Understanding these processes will be important to understanding the way statistical inference works.



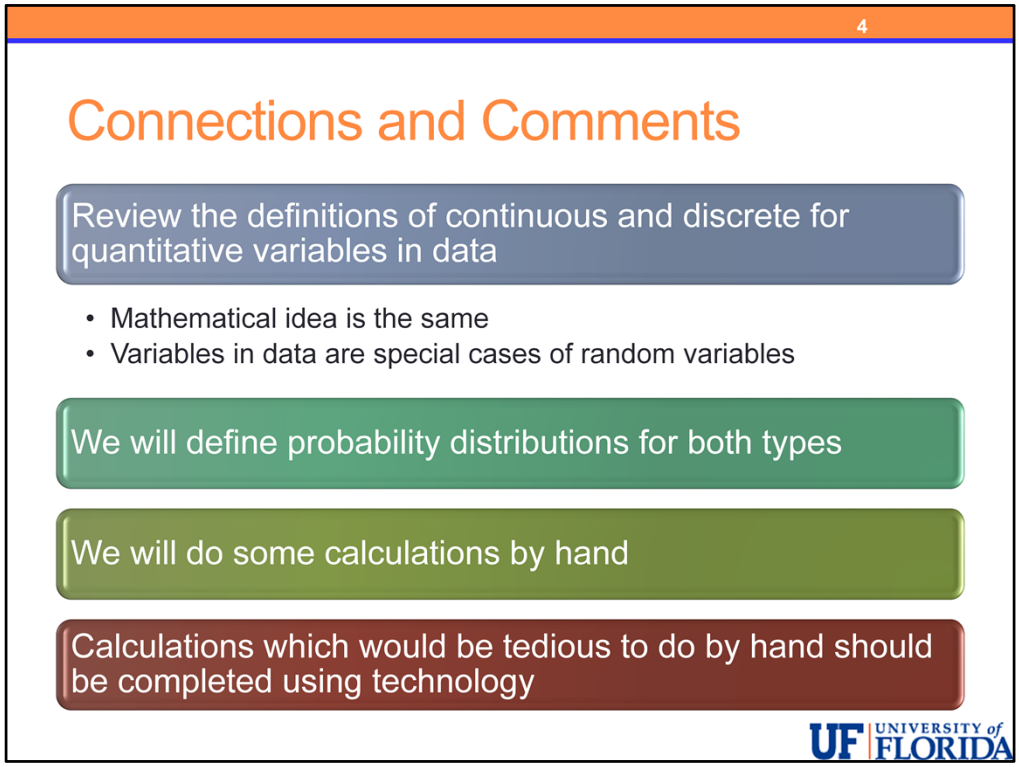
We will discuss two types of random variables.

Discrete random variables have a countable number of possible values. There may be an infinite number of possible values (in the sense that we do not know the potential maximum) but there are gaps between the possible values.

Usually discrete random variables represent a count of some kind. The phrase "the number of ..." usually indicates a discrete random variable. The number of falls in a hospital per week. The number of eggs in an alligator nest. These are counts and are discrete random variables.

It is possible to have a discrete random variable that is not specifically a count, however, the possible values still need to be countable, we must be able to list them. They also must have gaps between the possible values by their nature, not based upon our rounding preferences.

Continuous random variables can take on any value in an interval and can take on an infinite number of possible values. We cannot list them all.



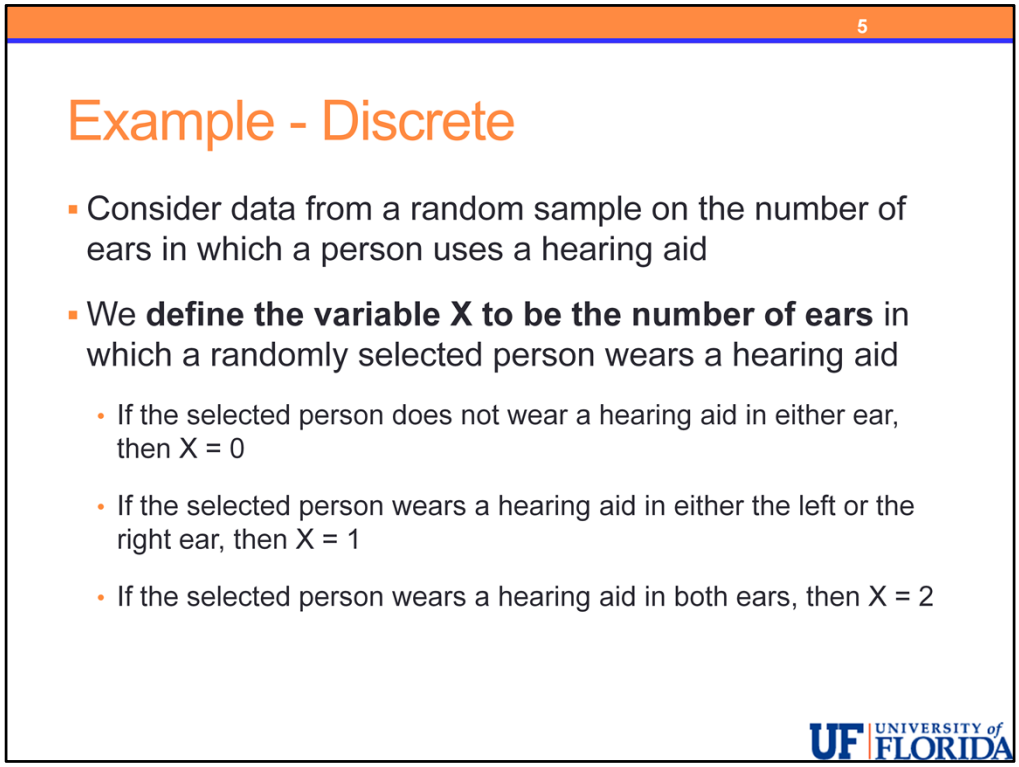
We have discussed discrete and continuous before as it related to quantitative variables. The definitions are the same and in fact a continuous quantitative variable is a special case of a continuous random variable and a discrete quantitative variable is a special case of a discrete random variable but we will see that the definition of a random variable is a much broader concept, especially to statisticians.

In this section we will be learning about how to mathematically define the theoretical PROBABILITY DISTRIBUTION for discrete and continuous random variables and to use this knowledge to calculate probabilities and determine what values are common or rare.

We will begin with a discussion of discrete probability distributions in general and then talk more about the binomial distribution.

Also, we will discuss how to calculate some values by hand and instruct you in using some tables (via the course material examples). For the most part, calculations can be completed using a number of online calculators or other programs such as EXCEL. We won't expect you to do any ridiculous by hand calculations, so if you feel you are going through such a process, stop and ask us how you should be approaching the solution and we will point you back in the right direction!

I am not a fan of meaningless labor in calculation. I am most interested in your understanding and comprehension of the concepts.



In a study of individuals with some degree of hearing loss, individuals were asked in which ear(s) they wear a hearing aid. Possible answers were none, left, right, both.

As recorded, this is a categorical variable. However, we can convert it to a numeric RANDOM VARIABLE by considering the random variable X to be equal to the number of ears for which a hearing aid is used.

This will give X = 0 if no hearing aid is used, X = 2 if a hearing aid is used in both ears, and X = 1 if a hearing aid is used in only one ear (either left or right).

So this is an example of a discrete random variable. It has three possible values with gaps between them. We can list the possible values and they are numeric.

Example - Discrete

- X is a quantitative variable which takes the possible values of 0, 1, or 2.
- We can ask questions like:
  - What is the probability that a randomly selected person will have a hearing aid in both ears?
  - What is the probability that a randomly selected person will not be wearing a hearing aid in either ear?
  - What is the probability that a randomly selected person will have a hearing aid only one ear?

UF UNIVERSITY of FLORIDA

We can ask questions like:

What is the probability that a randomly selected person from our sample will use hearing aids in both ears? Neither ear? Only one ear?

Here we know the maximum number we could observe is 2. We can have discrete random variables where we do not know the upper (or lower) limit. We cannot place a clear maximum on the possible outcome. However, there are still gaps between the values and they can be listed.

Example: Continuous

- Assume we choose a newborn infant at random and record the weight of the infant in grams.
  - Note that we can't list all the possible outcomes here
  - We'll **define X to be the weight of the newborn**
- We can ask questions like:
  - P(X < 2500)
  - P(2800 < X < 3400)

UF UNIVERSITY of FLORIDA

The other type of random variable is a continuous random variable. These definitions are the same as we had for quantitative variables in data where we had discrete and continuous variables. Similarly, a continuous random variable is a random variable that can take on any value in an interval. There are no longer any gaps between the possible values.

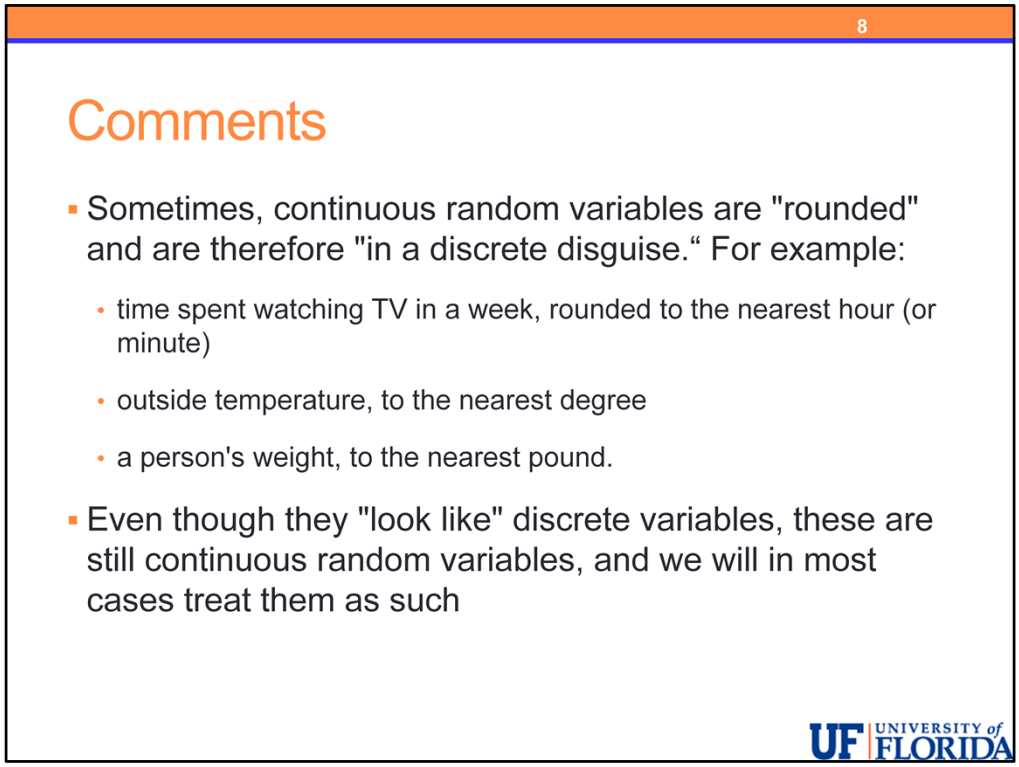
Suppose we consider the weight of newborn infants in grams. We cannot list all of the possible values here. We are only restricted by our ability or interest in measuring the value more precisely. So if we consider X to be the weight of a randomly selected newborn. We can ask questions like

What is the probability that X will be less than 2500 grams? In other words, what is the probability that the newborn will weigh less than 2500 grams?

What is the probability that X will be between than 2800 and 3400 grams? In other words, what is the probability that the newborn will weigh between 2800 and 3400 grams?

We need to be able to work back and forth between the verbal description and the probability notation of the random variable, X.

The difference here is that, for continuous random variables, we aren't going to be listing the values individually 2500, 2501, 2502, etc. In fact we can still have fractions of a gram which makes it impossible to list all values precisely.



Be careful, many continuous random variables may be presented as rounded values which can lead you to conclude they are discrete when in fact they are continuous. Ask yourself if there are TRUE gaps between the possible values you can observe. If there are true gaps then it is discrete. IF there are no gaps between possible values, it is continuous.

One issue with these definitions is that, although the definitions work to classify random variables into two types, there are some situations where we will handle a random variable differently than its type. For example:

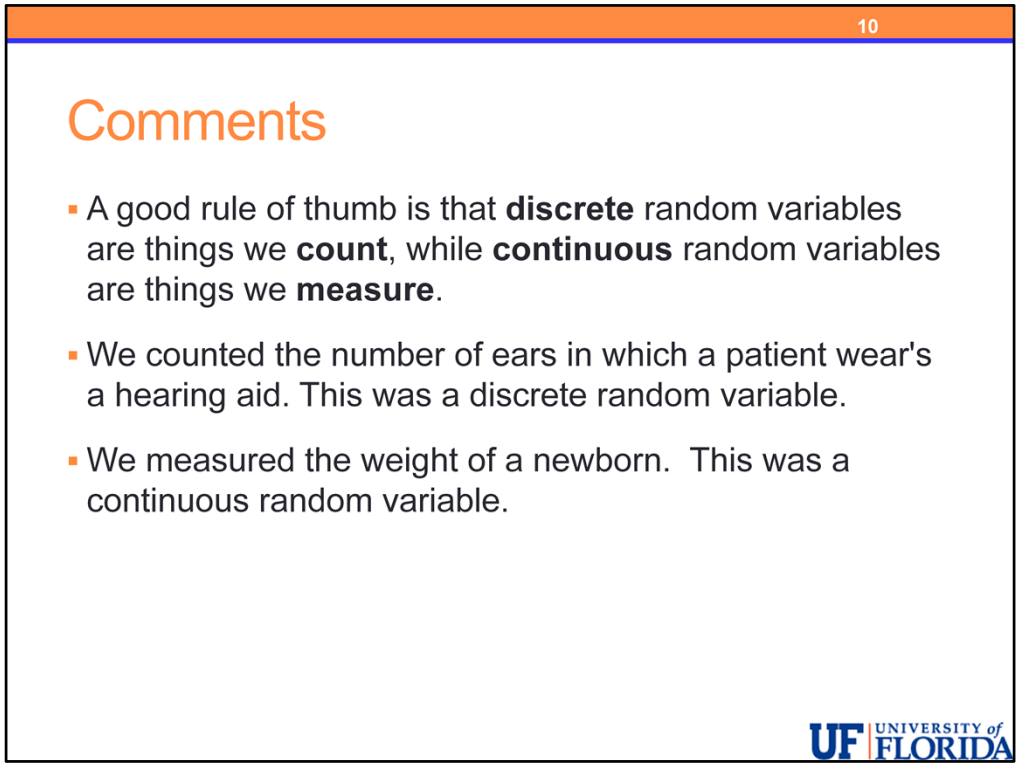
We can have a rounded continuous random variable where we only have a few possible values 130, 131, 132, 133, and 134. If those are my only values, in some analyses, I may treat this as if it is discrete.

Comments

- On the other hand, there are some variables which are discrete in nature, but take so many distinct possible values that it will be much easier to treat them as continuous rather than discrete.
  - the IQ of a randomly chosen person
  - the SAT score of a randomly chosen student
  - the annual salary of a randomly chosen CEO, whether rounded to the nearest dollar or the nearest cent

UF UNIVERSITY of FLORIDA

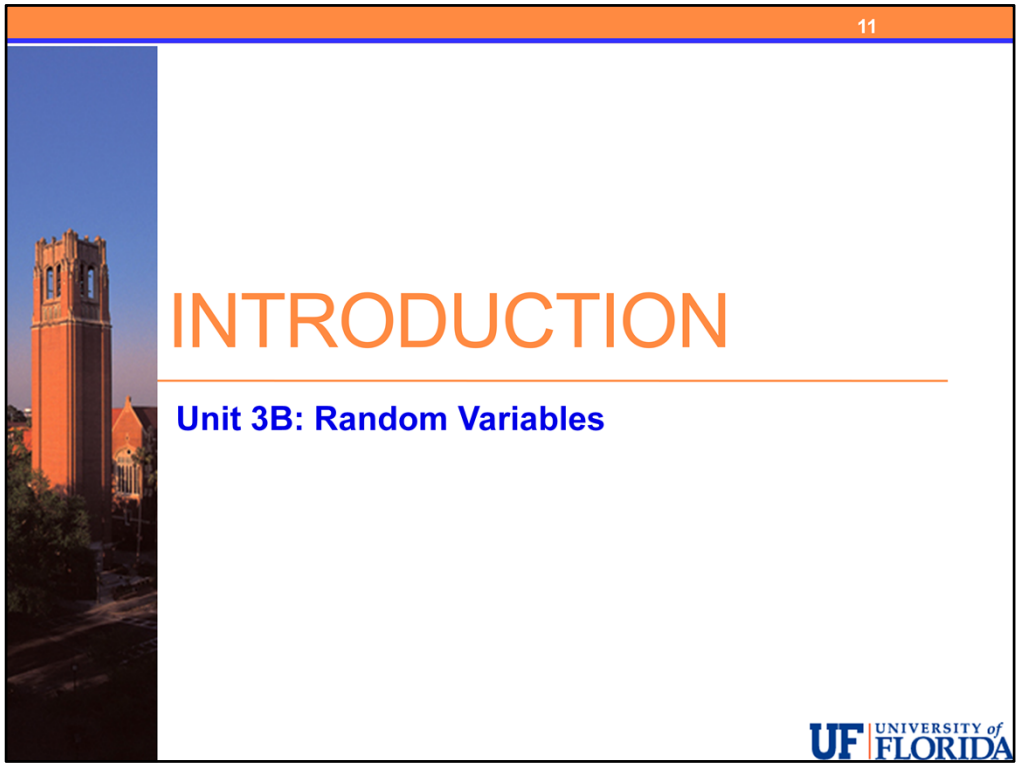
We may have discrete random variables that have so many possible values that we will treat them as if they are continuous. For example, the number of patients treated in a hospital in a particular year. If we had this random variable recorded for 1000 different hospitals, we would have a difficult time listing all possible values and making any sense of the results. Better to analyze these counts as if they were continuous.



A good rule of thumb is that discrete random variables are things we count or list while continuous random variables are things we measure.

We counted the number of ears in which a patient wear's a hearing aid. This was a discrete random variable.

We measured the weight of a newborn. This was a continuous random variable.



The skills we will learn in this section on random variables will be important on our journey toward understanding statistical inference.