

## CHAPTER OVERVIEW

### Unit 3B: Random Variables

**CO-6:** Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

#### Video

**Video:** [Unit 3B Random Variables](#) (10:00)

#### Introduction

In the remaining sections in Unit 3 we will begin to make the **connection** between **probability** and **statistics** so that we can apply these concepts in the final Unit on statistical inference.

These concepts **bridge the gap** between the **mathematics of descriptive statistics** and **probability** and **true “Inferential Statistics”** where we will formalize **statistical hypothesis tests**.

In other words, the topics in Unit 3B provide the **mathematical background** and **concepts** that will be needed for our **study of inferential statistics**.

In the previous sections we learned principles and tools that help us find probabilities of events in general.

Now that we’ve become proficient at doing that, we’ll talk about **random variables**.

Just like any other variable, random variables can take on multiple values.

What differentiates **random variables** from other variables is that the **values** for these variables are **determined by a random trial, random sample, or simulation**.

The probabilities for the values can be determined by theoretical or observational means.

Such probabilities play a vital role in the theory behind statistical inference, our ultimate goal in this course.

#### Random Variables

##### Learning Objectives

**LO 6.11:** Distinguish between discrete and continuous random variables

We first discussed variables in the Exploratory Data Analysis portion of the course. A variable is a characteristic of an individual.

We also made an important distinction between **categorical variables**, whose values are groups or categories (and an individual can be placed into one of them), and **quantitative variables**, which have numerical values for which arithmetic operations make sense.

In the previous sections, we focused mostly on events which arise when there is a categorical variable in the background: blood type, pierced ears (yes/no), gender, on time delivery (yes/no), side effect (yes/no), etc.

Now we will begin to consider quantitative variables that arise when a random experiment is performed. We will need to define this new type of variable.

A **random variable** assigns a unique numerical value to the outcome of a random experiment.

A random variable can be thought of as a function that associates exactly one of the possible numerical outcomes to each trial of a random experiment. However, that number can be the same for many of the trials.

Before we go any further, here are some simple examples:

#### ✓ EXAMPLE: Theoretical

Consider the random experiment of flipping a coin twice.

- The sample space of possible outcomes is  $S = \{ HH, HT, TH, TT \}$ .

Now, let's **define the variable X to be the number of tails** that the random experiment will produce.

- If the outcome is HH, we have no tails, so the value for **X is 0**.
- If the outcome is HT, we got one tail, so the value for **X is 1**.
- If the outcome is TH, we again got one tail, so the value for **X is 1**.
- Lastly, if the outcome is TT, we got two tails, so the value for **X is 2**.

As the definition suggests, **X is a quantitative variable that takes the possible values of 0, 1, or 2**.

It is **random** because we do not know which of the three values the variable will eventually take.

**We can ask questions like:**

- What is the probability that X will be 2? In other words, what is the probability of getting 2 tails?
- What is the probability that X will be at least 1? In other words, what is the probability of getting at least 1 tail?

As you can see, random variables are not really a new thing, but just a different way to look at the same problem.

Note that if we had tossed a coin three times, the possible values for the number of tails would be 0, 1, 2, or 3. In general, if we toss a coin “n” times, the possible number of tails would be 0, 1, 2, 3, ... , or n.

#### ✓ EXAMPLE: Observational

Consider getting data from a random sample on the number of ears in which a person wears one or more earrings.

We **define the variable X to be the number of ears** in which a randomly selected person wears an earring.

- If the selected person does not wear any earrings, then **X = 0**.
- If the selected person wears earrings in either the left or the right ear, then **X = 1**.
- If the selected person wears earrings in both ears, then **X = 2**.

As the definition suggests, **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 earrings in both ears?
- What is the probability that a randomly selected person will not be wearing any earrings in either ear?

NOTE... We identified the first example as **theoretical** and the second as **observational**.

Let's discuss the distinction.

- To answer probability questions about a theoretical situation, we only need the principles of probability.
- However, if we have an observational situation, the only way to answer probability questions is to use the relative frequency we obtain from a random sample.

Here is a different type of example:

#### ✓ EXAMPLE: Lightweight Boxer

Assume we choose a lightweight male boxer at random and record his exact weight.

According to the boxing rules, a lightweight male boxer must weigh between 130 and 135 pounds, so the sample space here is

- $S = \{ \text{All the numbers in the interval } 130\text{-}135 \}$ .

Note that we can't list all the possible outcomes here!

We'll **define X to be the weight of the boxer** again, as the definition suggests, **X is a quantitative variable whose value is the result of our random experiment.**

Here **X can take any value between 130 and 135.**

We can ask questions like:

- What is the probability that X will be more than 132? In other words, what is the probability that the boxer will weigh more than 132 pounds?
- What is the probability that X will be between 131 and 133? In other words, what is the probability that the boxer weighs between 131 and 133 pounds?

What is the difference between the random variables in these examples? Let's see:

- They all arise from a random experiment (tossing a coin twice, choosing a person at random, choosing a lightweight boxer at random).
- They are all quantitative (number of tails, number of ears, weight).

Where they differ is in the type of possible values they can take:

- In the first two examples, X has three distinct possible values: 0, 1, and 2. You can list them.
- In contrast, in the third example, X takes any value in the interval 130-135, and thus the possible values of X cover an infinite range of possibilities, and cannot be listed.

## Types of Random Variables

A random variable like the one in the first two examples, whose possible values are a list of distinct values, is called a **discrete random variable**.

A random variable like the one in the third example, that can take any value in an interval, is called a **continuous random variable**.

The main distinction between these two types of random variables is that,

- although they can both take on a potentially infinite number of values,
- for **discrete** random variables there is always a **GAP** between any two possible values
- whereas for **continuous** random variables there are no gaps in the range of possible values – it can take on any value in an interval; our precision in measurement is only limited by our level of technology in taking that measurement.

Just as the distinction between categorical and quantitative variables was important in Exploratory Data Analysis, the distinction between discrete and continuous random variables is important here, as each one gets a different treatment when it comes to calculating probabilities and other quantities of interest.

Before we go any further, a few observations about the nature of discrete and continuous random variables should be mentioned.

### Comments:

- Sometimes, continuous random variables are “rounded” and are therefore “in a discrete disguise.” For example:
  - time spent watching TV in a week, rounded to the nearest hour (or minute)
  - outside temperature, to the nearest degree
  - a person's weight, to the nearest pound.

Even though they “look like” discrete variables, these are still continuous random variables, and we will in most cases treat them as such.

- 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

- Sometimes we have a discrete random variable but do not know the extent of its possible values.
  - For example: How many accidents will occur in a particular intersection this month?
  - We may know from previously collected data that this number is from 0-5. But, 6, 7, or more accidents could be possible.
- A good rule of thumb is that **discrete** random variables are things we **count**, while **continuous** random variables are things we **measure**.
  - We counted the number of tails and the number of ears with earrings. These were discrete random variables.
  - We measured the weight of the lightweight boxer. This was a continuous random variable.

Often we can have a subject matter for which we can collect data that could involve a discrete or a continuous random variable, depending on the information we wish to know.

#### ✓ EXAMPLE: Soft Drinks

Suppose we want to know **how many days per week you drink a soft drink**.

- The sample space would be  $S = \{ 0, 1, 2, 3, 4, 5, 6, 7 \}$ .
- There are a finite number of values for this variable.
- This would be a **discrete** random variable.

Instead, suppose we want to know **how many ounces of soft drinks you consume per week**.

- Even if we round to the nearest ounce, the answer is a measurement.
- Thus, this would be a **continuous** random variable.

#### ✓ EXAMPLE: x-bar

Suppose we are interested in the weights of all males.

- We take a random sample and get the mean for that sample, namely x-bar.
- We then take another random sample (with the same sample size) and get another x-bar.
- We would expect the values of the x-bars from these two samples to be different, but pretty close in value.
- Each time we take a sample we'll get a different x-bar.
- We will take lots of samples and thus get many x-bar values.

The value of x-bar from these repeated samples is a **random variable**.

Since it can take on any value within an interval of possible male weights it is a **continuous** random variable.

#### Did I Get This?: [Random Variables](#)

We devote a great deal of attention to random variables, since random variables and the probabilities that are associated with them play a vital role in the theory behind statistical inference, our ultimate goal in this course.

We'll start with discrete random variables, including a discussion of binomial random variables and then move on to continuous random variables where we will formalize our understanding of the normal distribution.

[Binomial Random Variables](#)

[Continuous Random Variables](#)

[Discrete Random Variables](#)

[Normal Random Variables](#)

[Summary \(Unit 3B - Random Variables\)](#)