Lesson 2: Introduction to Simulations

Nicky Wakim

2025-10-01

Learning Objectives

1. Describe random variables and distinguish between discrete and continuous types
2. Explain the role of simulation in approximating probabilities and distributions
3. Use R to run simulations of discrete random variables
4. Apply and interpret the four-step simulation process

Where are we?

Basics of probability

- Outcomes and events
- Sample space
- Simulations
- Probability axioms
- Probability properties
- Counting
- Rules of probability

Probability for random variables

- Probability functions: pmfs/pdfs
- Cumulative distribution functions (CDFs)
- Important distributions

Two random variables

- Transformations
- Joint distributions
- Independence and conditioning

Expected values

- Means / expected values
- Linear combinations and conditioning
- Variance

Advanced probability

- Central limit theorem
- Functions: moment generating functions

Learning Objectives

1. Describe random variables and distinguish between discrete and continuous types
2. Explain the role of simulation in approximating probabilities and distributions
3. Use R to run simulations of discrete random variables
4. Apply and interpret the four-step simulation process

Recall: Outcomes, events, sample spaces

Definition: Outcome

The possible results in a random phenomenon.

Definition: Sample Space

The **sample space** S is the set of *all* outcomes

Definition: Event

An **event** is a *collection of some* outcomes. An event can include multiple outcomes or no outcomes (a subset of the sample space).

When thinking about events, think about outcomes that you might be asking the probability of. For example, what is the probability that you get a heads or a tails in one flip? (Answer: 1)

We need to understand random variables

Definition: Random Variable

For a given sample space S, a **random variable** (r.v.) is a **function** whose domain is S and whose range is the set of real numbers $\mathbb{R}$. **A random variable assigns a real number to each outcome in the sample space.**

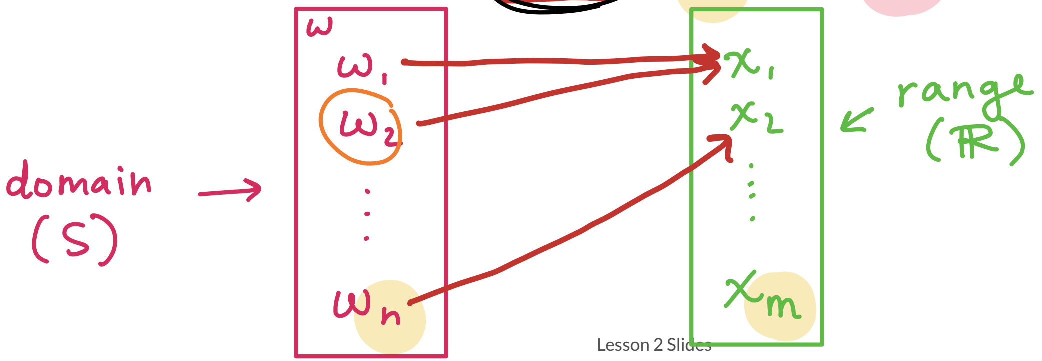
→ 0, 1, 2, 3, 4

- A random variable's value is completely determined by the outcome ω, where $\omega \in S$
  - What is *random* is the outcome ω
- A **random variable** is a function from the sample space (with outcomes ω) to the set of real numbers
  - We typically write $X(\omega)$ (or X for short), where X is our random variable
- Thus, we can take our sample space (all outcomes) and make functional transformations to it

$X(\omega)$ $X(\omega) = x$ → output / realised value of R.V.

2 ω's can point to same x

1 ω cannot point to 2 x's



The cool (and tricky) thing about random variables

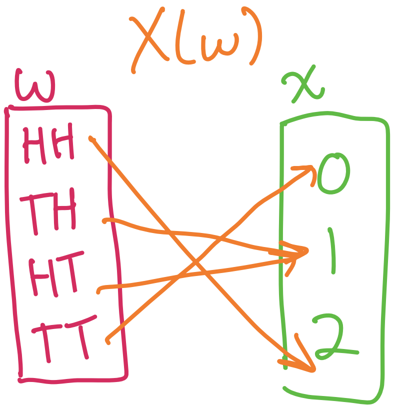
Do you remember our coin example from Lesson 1? We tossed one or two coins.

- For each coin, the **sample space** is heads and tails ($S = \{H, T\}$)
- If we want the **sample space** for both coins in order, then we have combinations ($S = \{(H,H), (H,T), (T,H), (T,T)\}$)

$$X(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}$$

We make the **random variable** a *function* of the **sample space**.

- For one coin toss, we can say random variable X is 1 if we toss a heads ($\omega = \mathrm{H}$) and $X = 0$ if we get a tails
- For the two coins, we can say X is the count of heads, so if $\omega = (\mathrm{H, T})$, then $X = 1$



$$X(\omega) = \begin{cases} 0 & \omega = (T,T) \\ 1 & \omega = (H,T), (T,H) \\ 2 & \omega = (H,H) \end{cases}$$

Types of random variables

There are two types of random variables:

- **Discrete random variables** (RVs): the set of possible values is either finite or can be put into a countably infinite list
  - You could *theoretically* list the specific possible outcomes that the variable can take
  - If you sum the rolls of three dice, you must get a whole number. For example, you can't get any number between 3 and 4.

- **Continuous random variables** (RVs): take on values from continuous *intervals*, or unions of continuous intervals
  - Variable takes on a range of values, but there are infinitely possible values within the range
  - If you keep track of the time you sleep, you can sleep for 8 hours or 7.9 hours or 7.99 hours or 7.999 hours …

- **Discrete random variables** (RVs) are a little easier to simulate right now
  - We will only do discrete RVs today

Learning Objectives

1. Describe random variables and distinguish between discrete and continuous types
2. Explain the role of simulation in approximating probabilities and distributions
3. Use R to run simulations of discrete random variables
4. Apply and interpret the four-step simulation process

What is a simulation?

A **probability model** for a random phenomenon includes a sample space, events, random variables, and a probability measure.

$P(A)$

Simulation

Simulation involves using a probability model to artificially recreate a random phenomenon, many times, usually using a computer.

↳ or random variable

We simulate outcomes and values of random variables according to the model's assumptions.

The Foundation: Relative Frequencies

- Probabilities can be interpreted as **long-run relative frequencies**

- By simulating a random phenomenon a large number of times, we can approximate the probability of an event by calculating the relative frequency of its occurrence
  - Basically, out of all the trials we run, how many times did the event happen?

$\hookrightarrow P(A) = \frac{|A|}{|S|} \xleftarrow{\text{approx}} P(A) = \frac{\#\text{ times get } A}{\#\text{ simulation}}$

- Simulation is a powerful tool to **approximate a few things**:
  - Probabilities
  - Distributions of random variables
  - Long-run averages

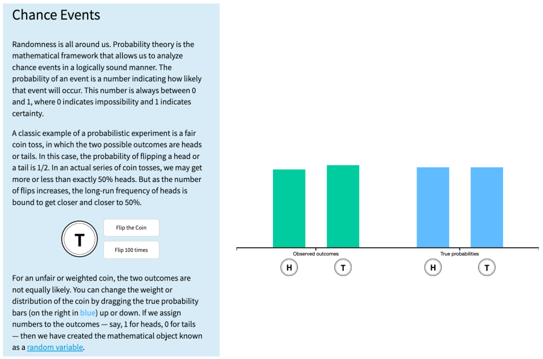
We saw an example of long-run relative frequency in our coin flip

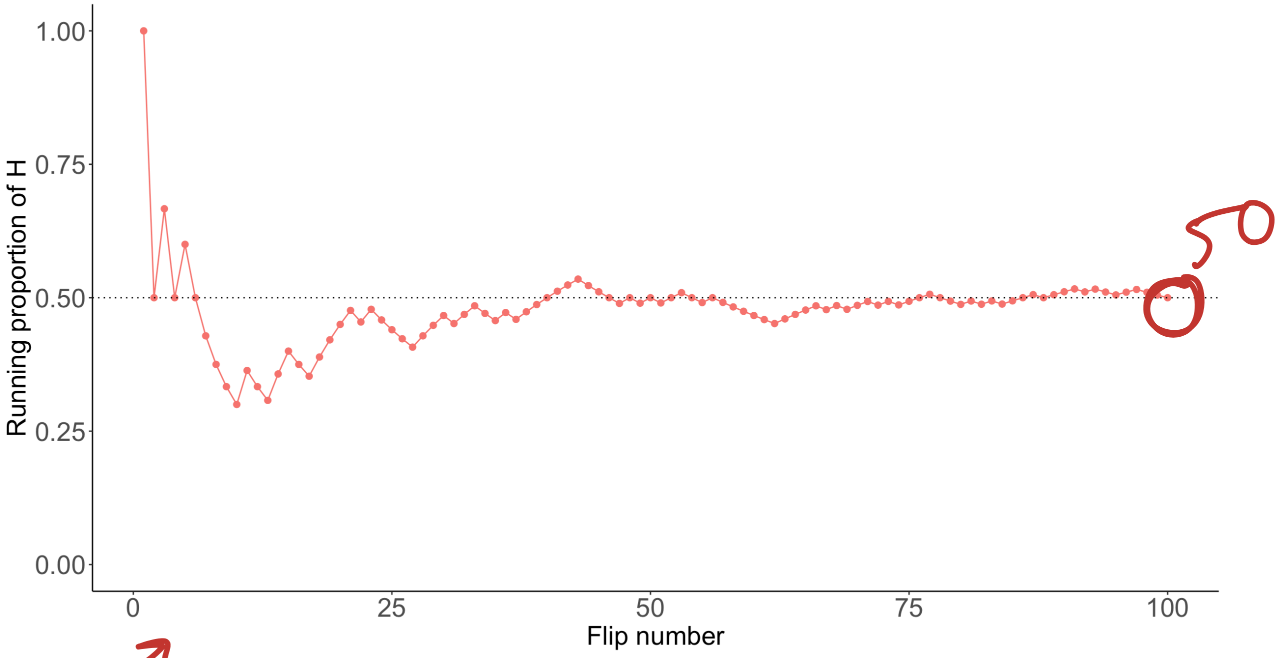
In Lesson 1, we flipped a coin 100 times and recorded the proportion of heads.

$A = H$

$$P(A) = P(H) = \frac{\#\text{ of Heads}}{\#\text{ sims}}$$

- We tossed 50 heads out of the 100 flips
- Our long-run frequency was $50/100 = 0.5$, which **approximated the probability** of getting a head on any one flip





Tactile simulations

- We've already seen coin flips!
- We can also use cards, dice, and other objects to simulate discrete random variables

- Other common method: A **box model** uses a box/hat/bucket of "tickets" with labels to represent possible outcomes
  - Allows us to increase the number of "tickets" with appropriate labels
  - **Coin flip as box model:** A box with two tickets (H and T).
  - **90% free throw shooter:** A box with 10 tickets (9 "make" and 1 "miss").
  - Draws can be **with replacement** (e.g., coin flips) or **without replacement** (e.g., dealing a poker hand).

Learning objectives

1. Describe random variables and distinguish between discrete and continuous types
2. Explain the role of simulation in approximating probabilities and distributions
3. Use R to run simulations of discrete random variables
4. Apply and interpret the four-step simulation process

Example to build our simulation skills

Example: Simulating Two Rolls of a Fair Four-Sided Die

We're going to roll two four-sided die. Let X be the sum of two rolls, and Y be the larger of the two rolls. How would we simulate X and Y separately?

- Note: this example is not asking for a probability!
  - We can simulate a random variable and looks at its distribution without calculating any probabilities.

- We will focus on simulating X first

Let's build up some coding tools to do this!

How do we simulate something like a single dice roll?

- We can also use R to sample from the box or spinner
- The `sample()` function is a powerful tool for simulating draws from a box model.
- For example, we can simulate a coin flip
  - What is `x`? → possible outcomes
  - What is `size`? → how many draws?

sample() is best for equally likely outcomes

```
sample(x = c("H", "T"), size = 1)
```

```
[1] "T"
```

~~wrong~~ x = "H", "T"

sample(x = c("make", "make", "make", ..., "miss"), size = 1)

- Or a dice roll

```
sample(x = c(1, 2, 3, 4), size = 1)
```

```
[1] 1
```

1:4

What if we have multiple rolls at once?

* default is replace = F

- We can set `size` to be larger than 1 to simulate multiple draws at once

```
sample(x = c("H", "T"), size = 5, replace = TRUE)
```

```
[1] "T" "T" "T" "H" "T"
```

- We can simulate our example of the two four-sided dice

```
sample(x = c(1, 2, 3, 4), size = 2, replace = TRUE)
```

```
[1] 3 3
```

- What happens if we set `replace = FALSE`? sample w/out replacement

```
sample(x = c(1, 2, 3, 4), size = 2, replace = FALSE)
```

```
[1] 4 2
```

```
sample(x = c(1, 2, 3, 4), size = 4, replace = FALSE)
```

```
[1] 2 1 3 4
```

size ≤ length of x

Can we start to simulate many rolls of two dice?

Example: Simulating Two Rolls of a Fair Four-Sided Die

We're going to roll two four-sided dice. Let X be the sum of two rolls, and Y be the larger of the two rolls. How would we simulate X and Y separately?

- We've seen how to simulate a single pair of rolls

```
rolls <- sample(x = c(1, 2, 3, 4), size = 2, replace = TRUE)
```

- We can use the `replicate()` function to repeat this process many times (we'll do 10)

```
reps <- 10
replicate(reps, sample(x = 1:4, size = 2, replace = TRUE))
```

```
        [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
D1 [1,]    2    3    3    3    3    3    3    1    3     3
D2 [2,]    1    1    2    2    2    1    4    1    4     2
           1    2    3    4    5    6    7    8    9    10
```

We need more reps for long-run relative frequencies

```
1 reps <- 10000
2 simulations <- replicate(reps, sample(x = 1:4, size = 2, replace = TRUE))
```

10000

- Let's show the first 14 simulations

```
1 simulations[, 1:14]
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]    4    1    2    2    4    2    1    2    4     1     4     2     2     3
[2,]    4    4    1    3    1    1    2    3    2     4     1     2     2     2
```

1st 14 pairs of rolls

- X is the sum of the two rolls: we could calculate that for each column

```
1 X_simulated <- apply(simulations, 2, sum)
2 X_simulated[1:14]
```

```
[1] 8 5 3 5 5 3 3 5 6 5 5 4 4 5
```

1: across rows

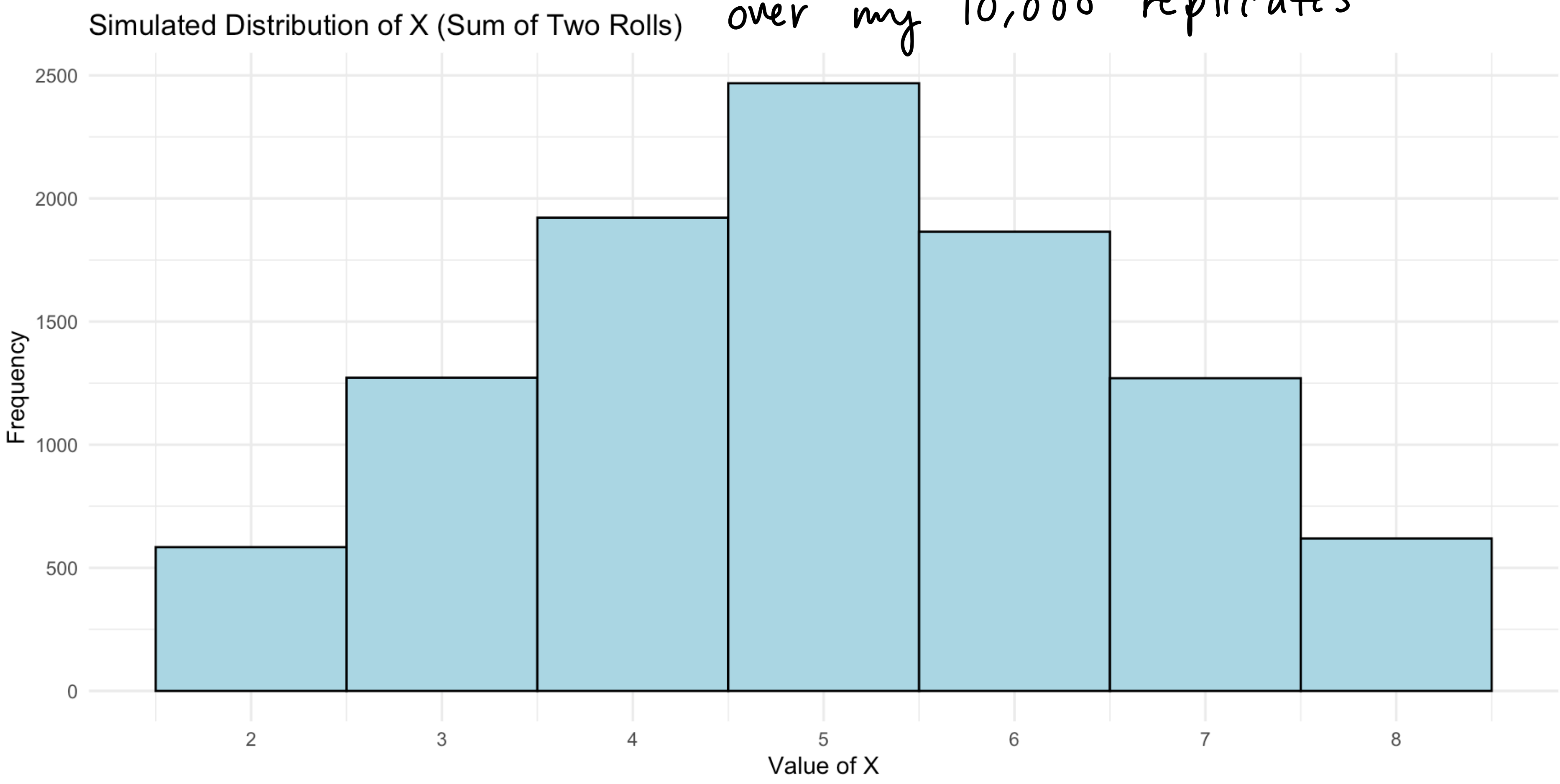
2: across columns

F T F T T F

X_simulated == 5

We can look at the plot of random variable X

over my 10,000 replicates



If we want to calculate something else, we can!

- Average:

```r
mean(X_simulated)
```

```
[1] 5.0044
```

- Standard deviation:

```r
sd(X_simulated)
```

```
[1] 1.574303
```

- Probability that $X = 5$:

```r
sum(X_simulated == 5) / reps
```

```
[1] 0.2468
```

- Probability that $X < 3$:

```r
sum(X_simulated < 3) / reps
```

```
[1] 0.0584
```

Probabilities (relative frequencies) are calculated by:

- summing the number of times an event occurs and
- dividing by the total number of simulations (reps)

Learning objectives

1. Describe random variables and distinguish between discrete and continuous types
2. Explain the role of simulation in approximating probabilities and distributions
3. Use R to run simulations of discrete random variables
4. Apply and interpret the four-step simulation process

4 (S)teps of a Simulation

1. Set up

Define the probability space and related random variables and events, including assumptions.

2. Simulate

Run the simulation to generate outcomes according to the assumptions.

3. Summarize

Analyze the output using plots and summary statistics like relative frequencies and averages.

4. Sensitivity analysis

Investigate how results change when assumptions or parameters of the model are altered.

Example: Dice Rolls

Example: Simulating Two Rolls of a Fair Four-Sided Die

We're going to roll two four-sided die. Let X be the sum of two rolls, and Y be the larger of the two rolls. How would we simulate X and Y separately?

$Y = \max$

- Use the steps to run simulation for for Y now

1. Set up

Define the probability space and related random variables and events, including assumptions.

- **Random variable:** $Y(\omega)$ is the larger of the two rolls in outcome ω
- **Goal:** simulate Y, the larger of two rolls of a fair four-sided die
- **Sample space:** all possible outcomes of rolling two four-sided die
  - Not necessary, but helpful to define the sample space

Y can be
1, 2, 3, 4

$$S = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(2,4),\\(3,1),(3,2),(3,3),(3,4),(4,1),(4,2),(4,3),(4,4)\}$$

$|S| = 16$

- **Assumptions:** each die is fair and rolls are independent
  - Each outcome in S is equally likely with probability $1/16$

2. Simulate

Run the simulation to generate outcomes according to the assumptions.

```r
reps <- 10000
simulations <- replicate(reps, sample(x = 1:4, size = 2, replace = TRUE))
Y_simulated <- apply(simulations, 2, max)
```

- We can look at the first 30 simulations

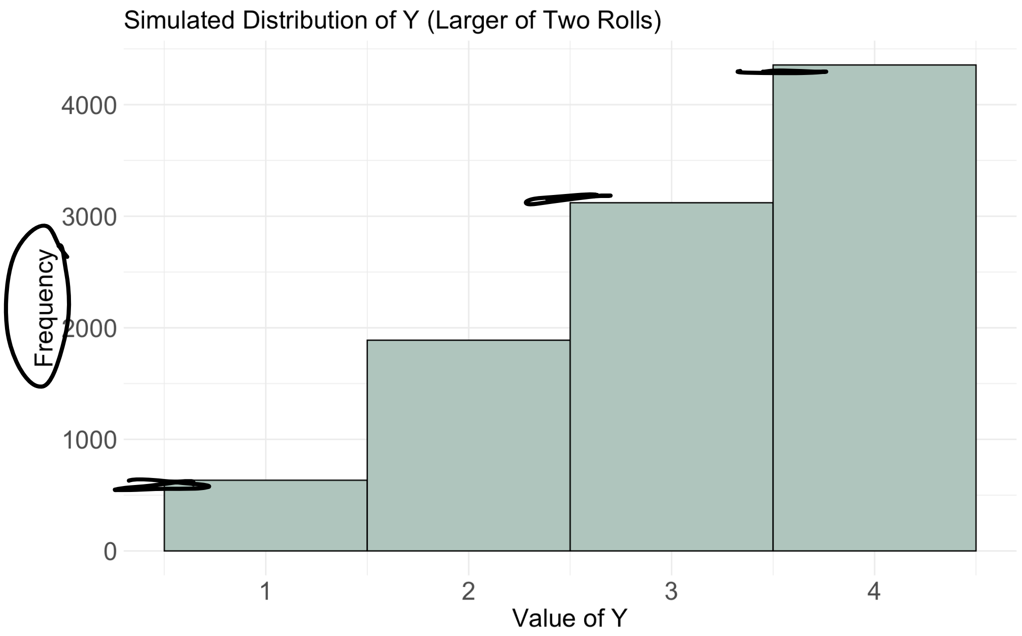
```r
Y_simulated[1:30]
```

```
[1] 3 4 4 4 2 4 4 4 2 3 4 4 1 4 4 3 4 4 2 4 4 4 1 4 2 2 4 4 3 3
```

3. Summarize

Analyze the output using plots and summary statistics like relative frequencies and averages.

▶ Show/Hide Code for plotting Y



If the problem asked us for something else, we could compute it:

- Average:

```
1 mean(Y_simulated)
```

```
[1] 3.1199
```

- Probability that $Y = 1$:

```
1 sum(Y_simulated == 1) / reps
```

```
[1] 0.0634
```

- Probability that $Y > 3$: $Y = 4$

```
1 sum(Y_simulated > 3) / reps
```

```
[1] 0.4356
```

4. Sensitivity analysis

Investigate how results change when assumptions or parameters of the model are altered.

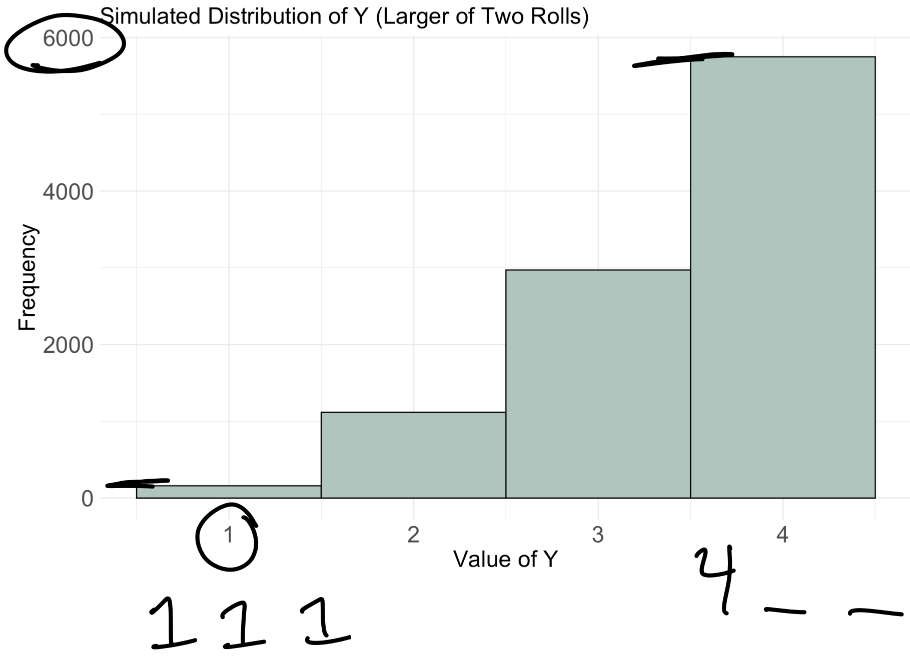
What if we rolled three die instead of two?

```
reps <- 10000
die <- 3
simulations <- replicate(
  reps,
  sample(x = 1:4,
         size = die,
         replace = TRUE)
)
simulations[, 1:6]
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
D1 [1,]	1	3	1	4	2	3
D2 [2,]	4	1	2	1	2	3
D3 [3,]	4	4	4	4	3	3

```
Y_simulated <- apply(simulations, 2, max)
```

▶ Show/Hide Code for plotting Y



Learning Objectives

1. Describe random variables and distinguish between discrete and continuous types
2. Explain the role of simulation in approximating probabilities and distributions
3. Use R to run simulations of discrete random variables
4. Apply and interpret the four-step simulation process