1.4 Frequency Distributions

Dr. Lauren Perry

Goals

1. Organize and visualize data using techniques for exploratory data analysis.
   - Create and interpret frequency distributions.
   - Create histograms
2. Identify the shape of a data set.
   - Describe skew and modality
3. Understand and interpret graphical displays.
   - Histograms and bar plots

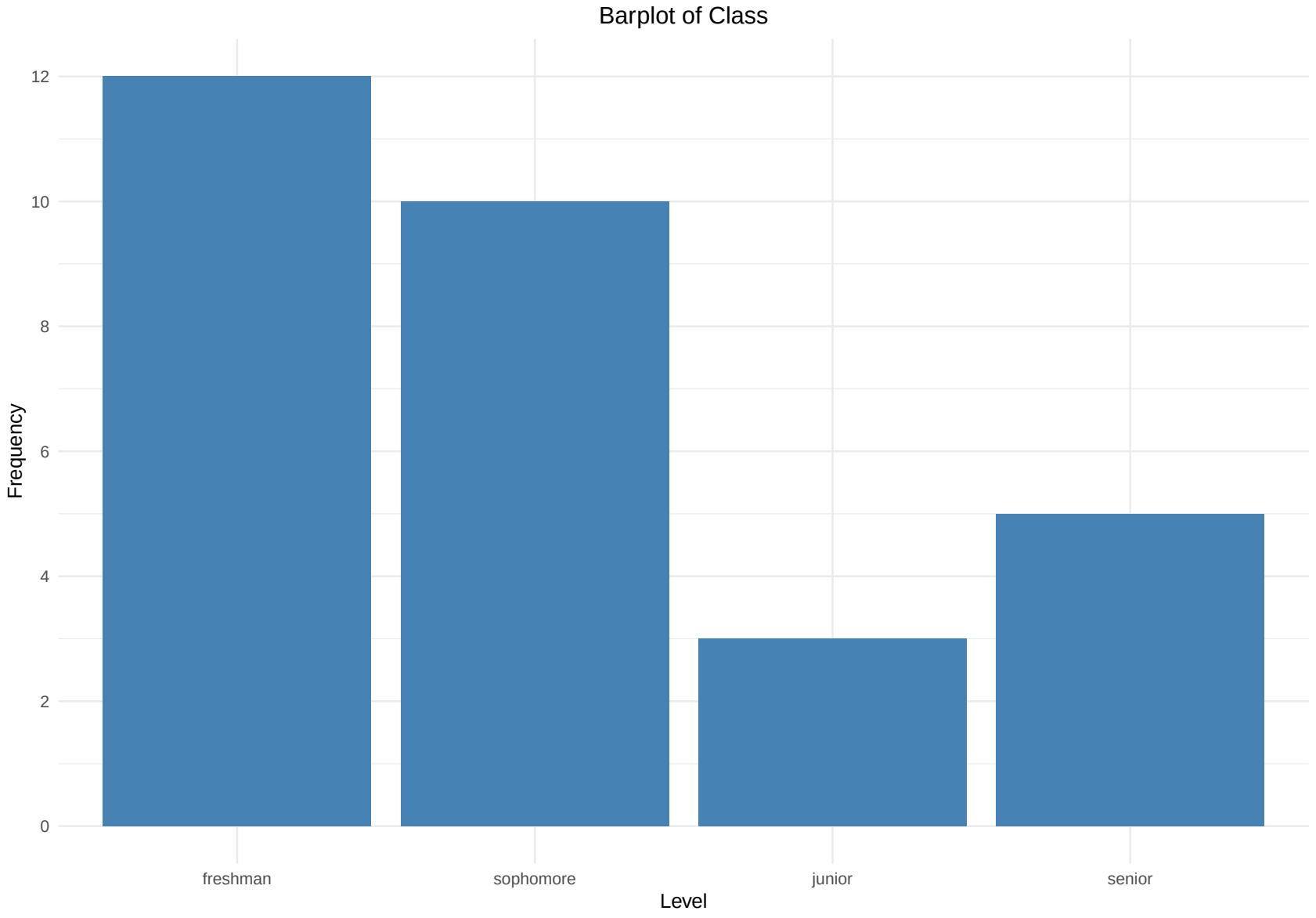
Categorical Variables

Frequency (**count**): the number of times a particular value occurs.

A **frequency distribution** lists each distinct value with its frequency.

Class	Frequency
freshman	12
sophomore	10
junior	3
senior	5

A **bar plot** is a graphical representation of a frequency distribution. Each bar's height is based on the frequency of the corresponding category.



Relative Frequencies

Relative frequency is the ratio of the frequency to the total number of observations.

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations}}$$

This is also called the **proportion**.

The **percentage** can be obtained by multiplying the proportion by 100.

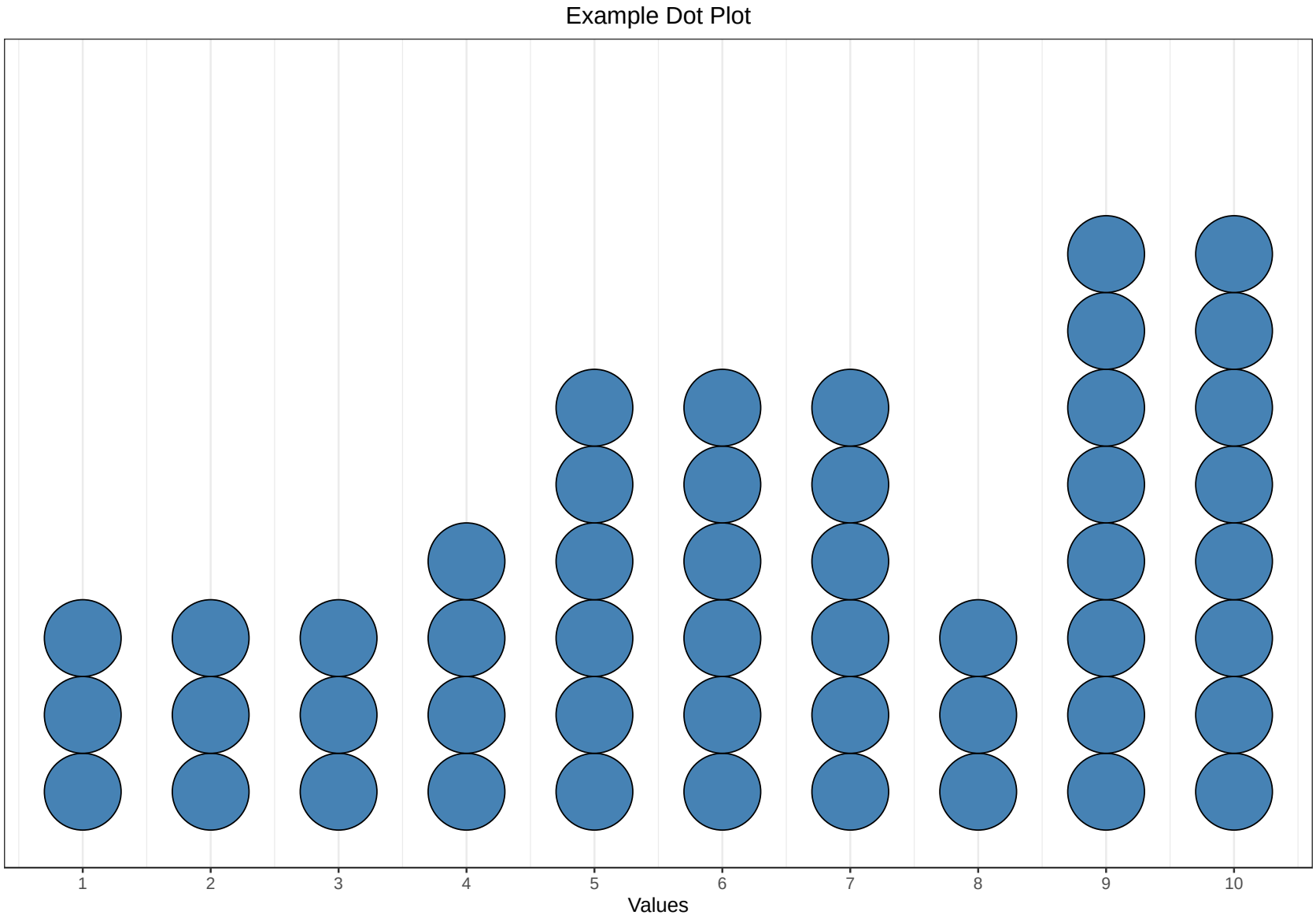
Relative Frequency Distribution

A **relative frequency distribution** lists each distinct value with its relative frequency.

Class	Frequency	Relative Frequency
freshman	12	$12/30 = 0.4$
sophomore	10	0.3333
junior	3	0.1
senior	5	0.1667

Numeric Variables

A **dot plot** shows a number line with dots drawn above the line. Each dot represents a single observation.



Binning Complex Data

- We would also like to be able to visualize larger, more complex data sets.
- This is hard to do using a dot plot!
- Instead, we can do this using **bins**, which group numeric data into equal-width consecutive intervals.

Example

A random sample of weights (in lbs) from 12 cats:

6.2	11.6	7.2	17.1	15.1	8.4
7.7	13.9	21.0	5.5	9.1	7.3

- **Minimum** is 5.5
- **Maximum** is 21

Lots of ways to break these into "bins", but what about...

- 5 - 10
- 10 - 15
- 15 - 20
- 20 - 25

Example

We've suggested bins

- 5 - 10
- 10 - 15
- 15 - 20
- 20 - 25

Each has an equal width of 5 (that's good), but if we had a cat with a weight of exactly 15 lbs, would we use the second or third bin??

Example

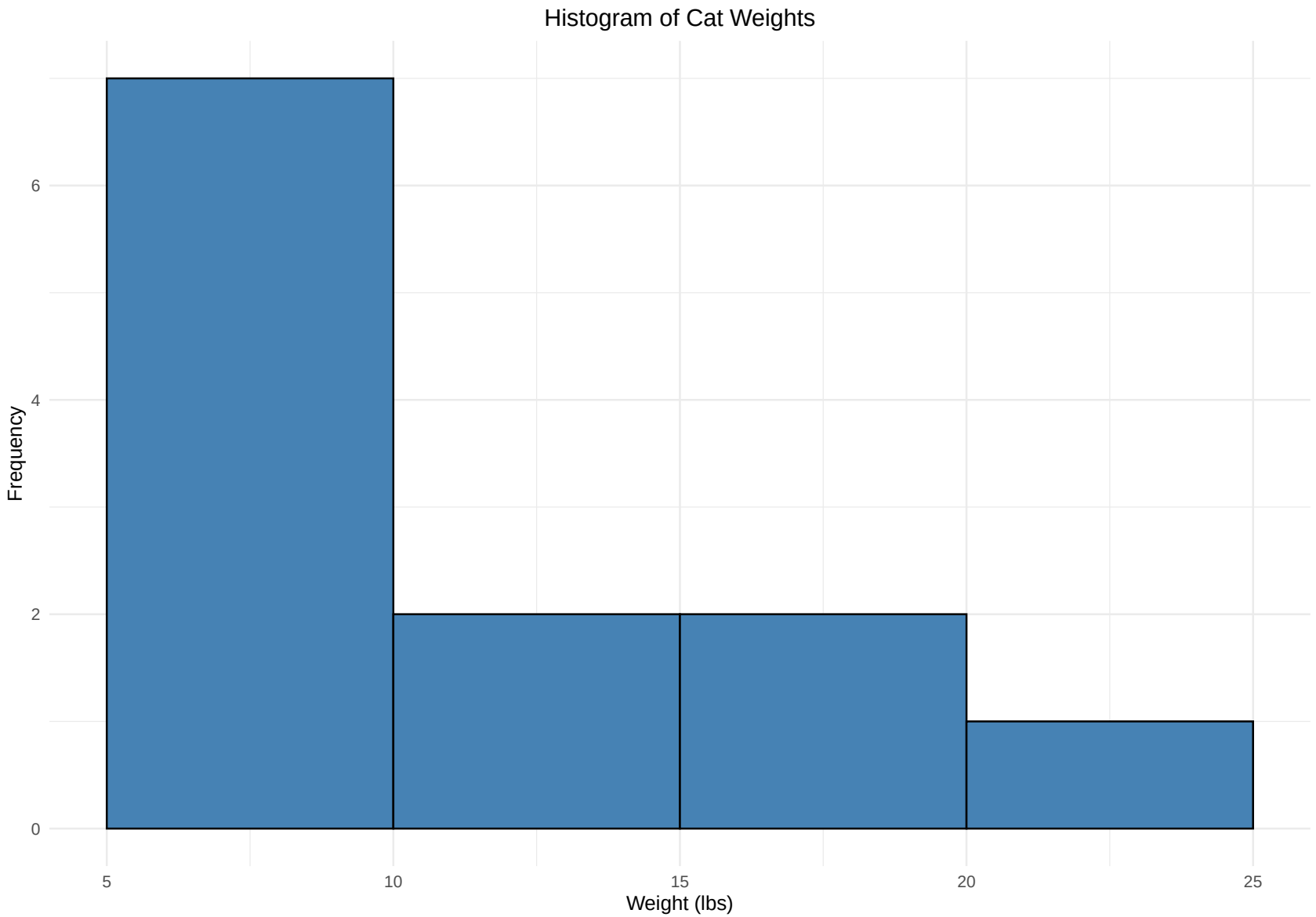
To make this clear, we need there to be no overlap. Instead, we could use:

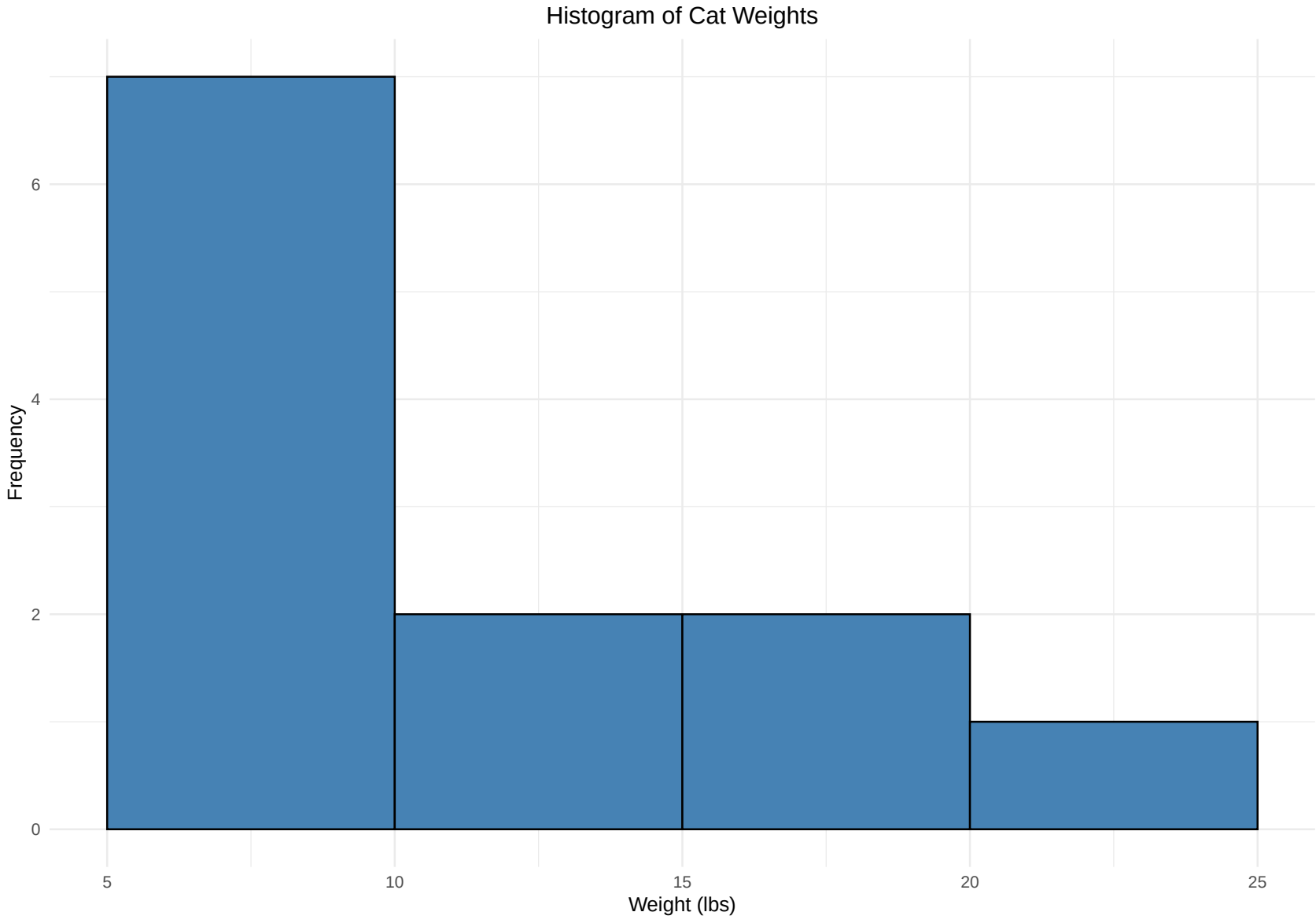
Weight	Count
5 - <10	7
10 - <15	2
15 - <20	2
20 - <25	1

Now, a cat with a weight of 15.0 lbs would be placed in the third bin (but not the second).

Example

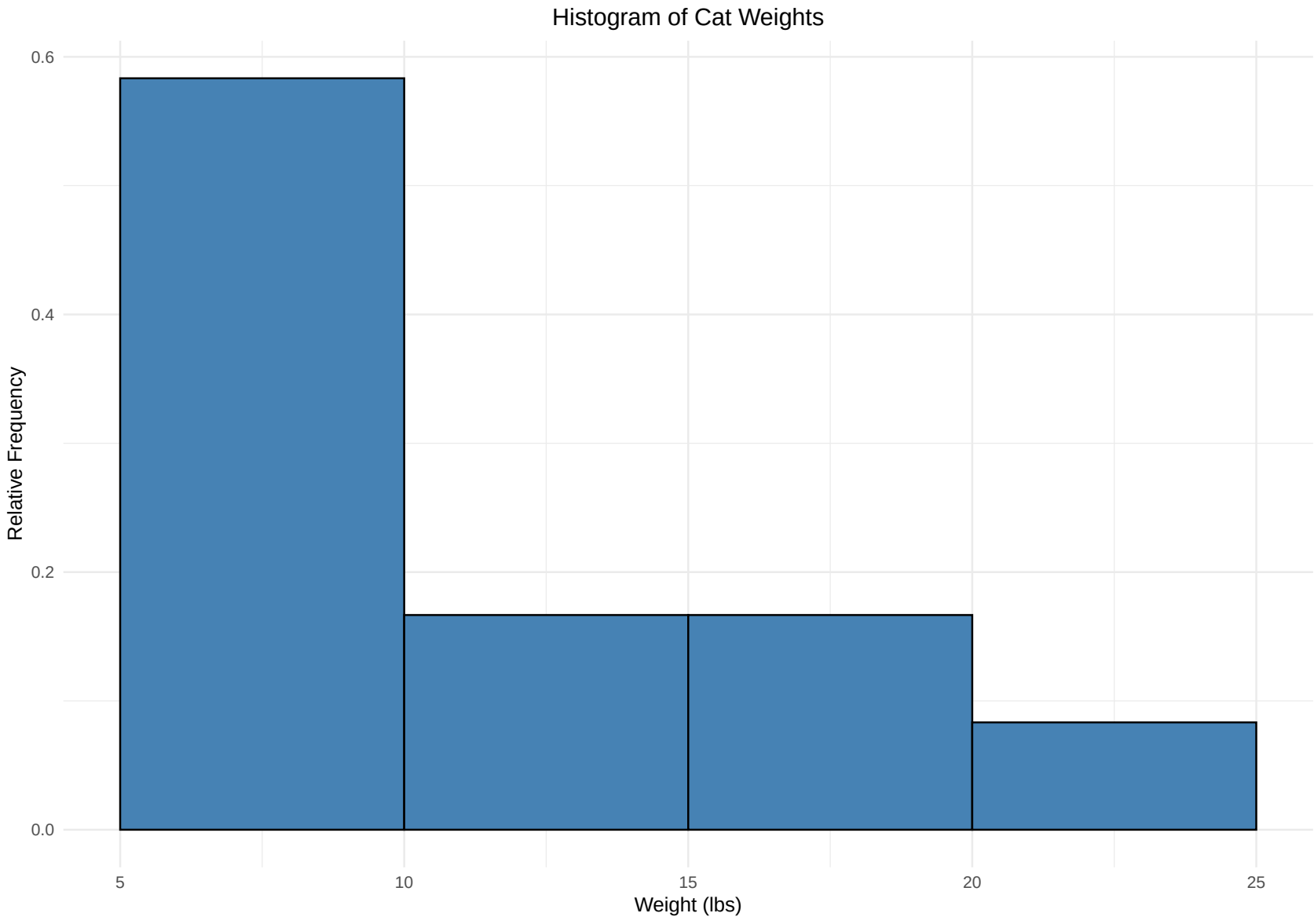
We will visualize this using a **histogram**, which is a lot like a bar plot but for numeric data:





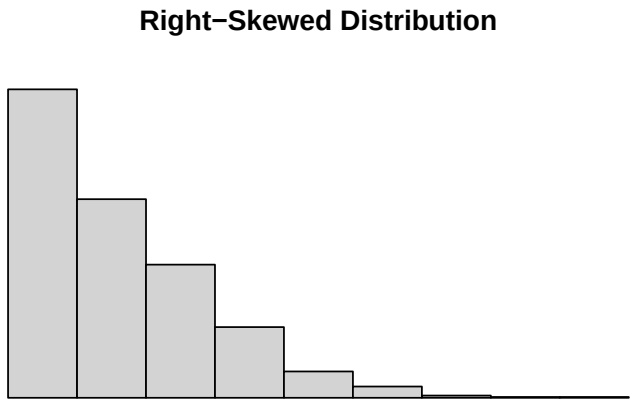
This is a **frequency histogram** because each bar height reflects the frequency of that bin.

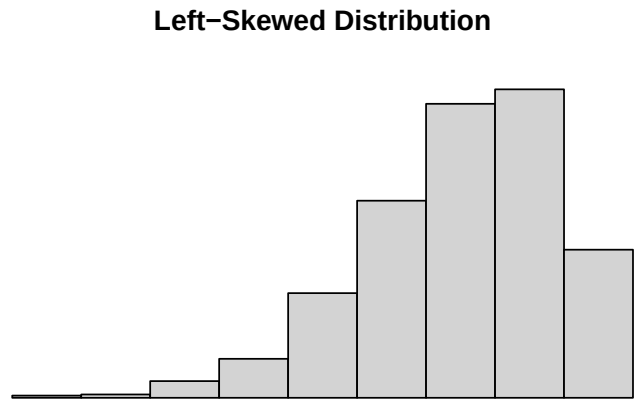
We can also create a **relative frequency histogram** which displays the relative frequency instead of the frequency:

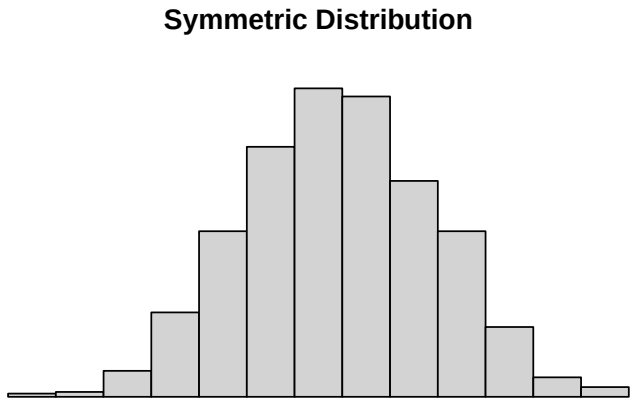


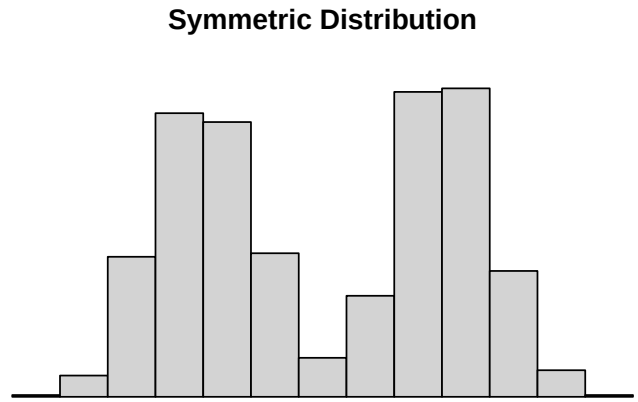
Notice that those last two histograms look the same *except for the numbers on the vertical axis*!

- This gives us insight into the shape of the data **distribution**, literally how the values are distributed across the bins.
- The part of the distribution that "trails off" to one or both sides is called a **tail** of the distribution.
- When a histogram trails off to one side, we say it is **skewed**.
  - right-skewed if it trails off to the right
  - left-skewed if it trails off to the left
- Data sets with roughly equal tails are **symmetric**.

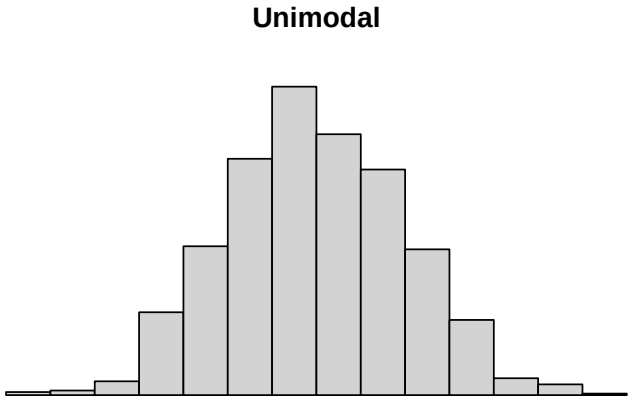

Right−Skewed Distribution

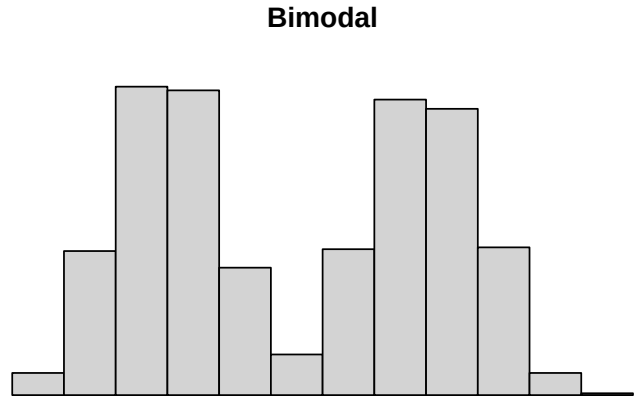

Left−Skewed Distribution

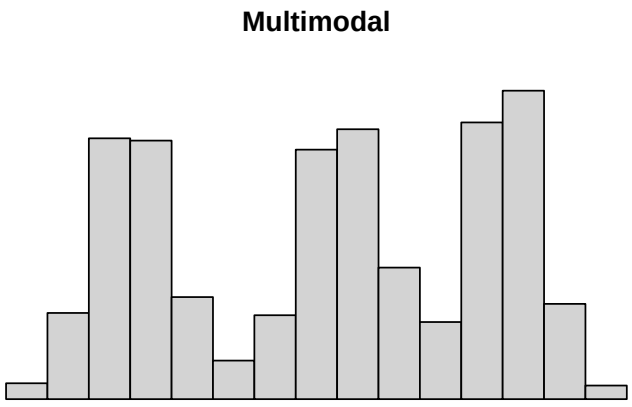

Symmetric Distribution

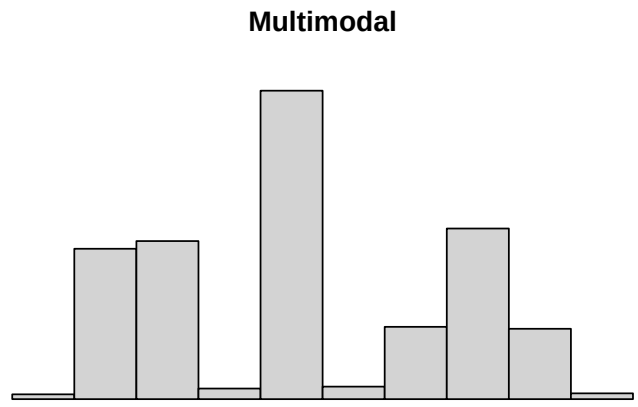

Symmetric Distribution

We can also use a histogram to identify **modes**.





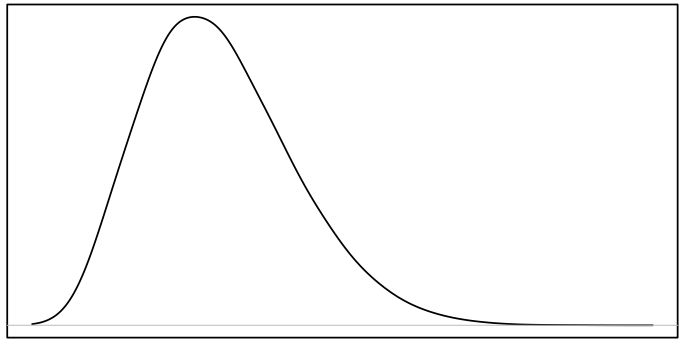




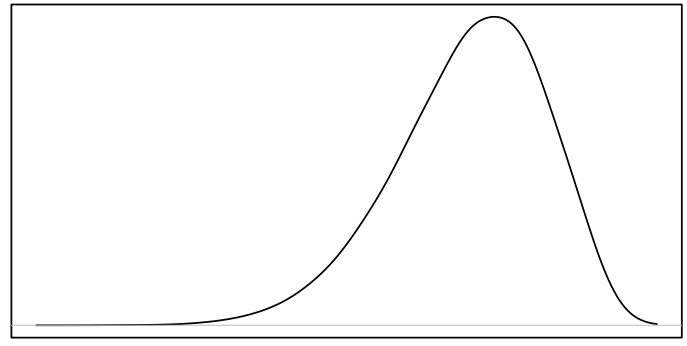
For numeric data, especially continuous variables, we think of modes as *prominent peaks*.

Finally, we can also "smooth out" these histograms and use a smooth curve to examine the shape of the distribution.

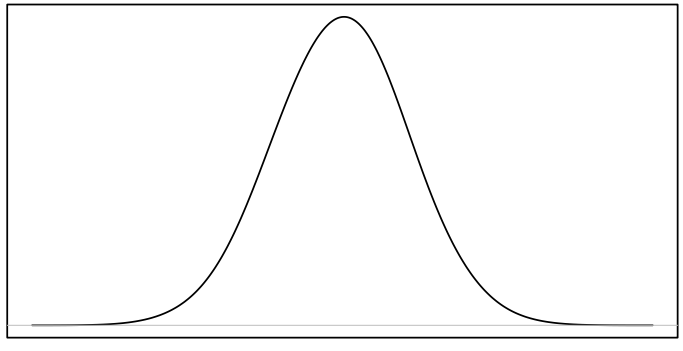
Right−Skewed Distribution



Left−Skewed Distribution



Symmetric Distribution



Symmetric Distribution