
# STATISTICS FOR DATA SCIENCE

## Data Visualization and Interpretation

**D. Uma**
Department of Computer Science and Engineering
**umaprabha@pes.edu**

# STATISTICS FOR DATA SCIENCE

## Data Visualization and Interpretation – Histogram

**D. Uma**
Department of Computer Science and Engineering

# STATISTICS FOR DATA SCIENCE

## Data Visualization

Data visualization is a **BIG buzz word** these days, but what does it actually mean ????

At a basic level, data is just information — **facts, figures, words, percentages, measurements, and observations**, but it's just computerized information.

In order for you to make it useful, you need to **find creative ways** to **make it user friendly** for your audience.

This is where the art of **data visualization** comes in!

Source: nutsandboltspeedtraining.com

# STATISTICS FOR DATA SCIENCE

## Data Visualization

**Data visualization** is the **secret art** of **turning data** into **visual graphics** that people can understand (**graphs, charts, info graphics**, etc.).

Words may be mightier than the sword, but in a battle for our brains, **visual images win every time**. - Colin Ware.

Source: nutsandboltspeedtraining.com

# STATISTICS FOR DATA SCIENCE

## Data Visualization

- Histogram
- Box plot
- Scatter plot
- Bar chart
- Heat map

Source: https://www.martinsights.com/?p=1256

# STATISTICS FOR DATA SCIENCE

## Data Distribution

To understand some of the fundamental concepts of statistical analysis, it is important to **appreciate** the **importance** of the **distribution of data points** in the sample.

**Data type** and the **distribution pattern** of their values **influence** the **choice of appropriate statistical tests**.

Emphasis will be placed on the **normal, or Gaussian, distribution.**

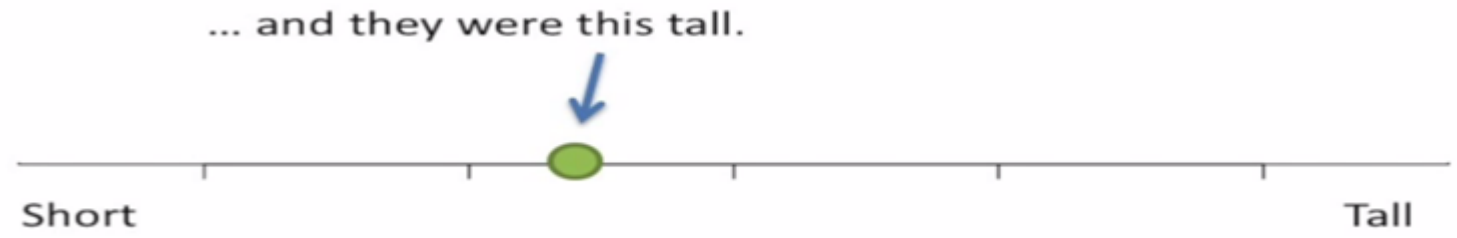
This is an important distribution to understand because the **assumption of this distribution** underlies the **use of many** common **statistical tests**.

Source: Sciencedirect.com

# STATISTICS FOR DATA SCIENCE

## Let's get an idea

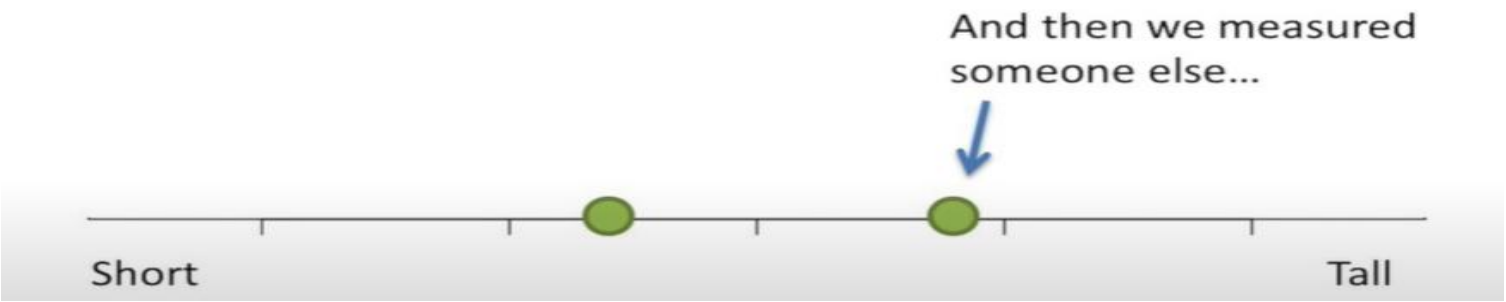
Imagine we went out and measured someone…

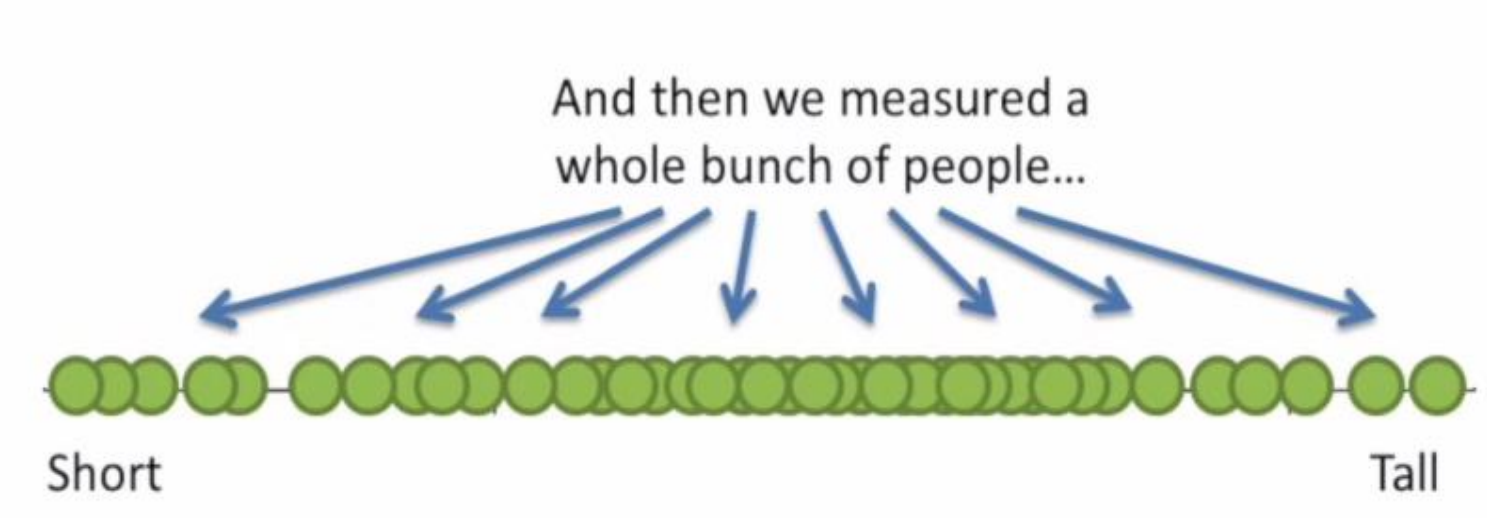




Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Let's get an idea

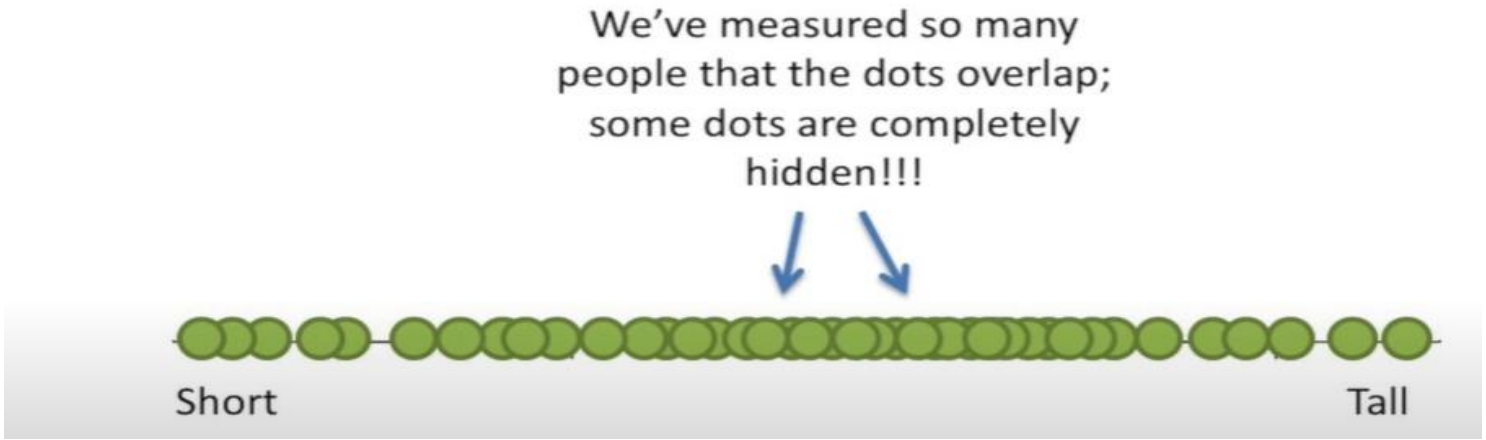


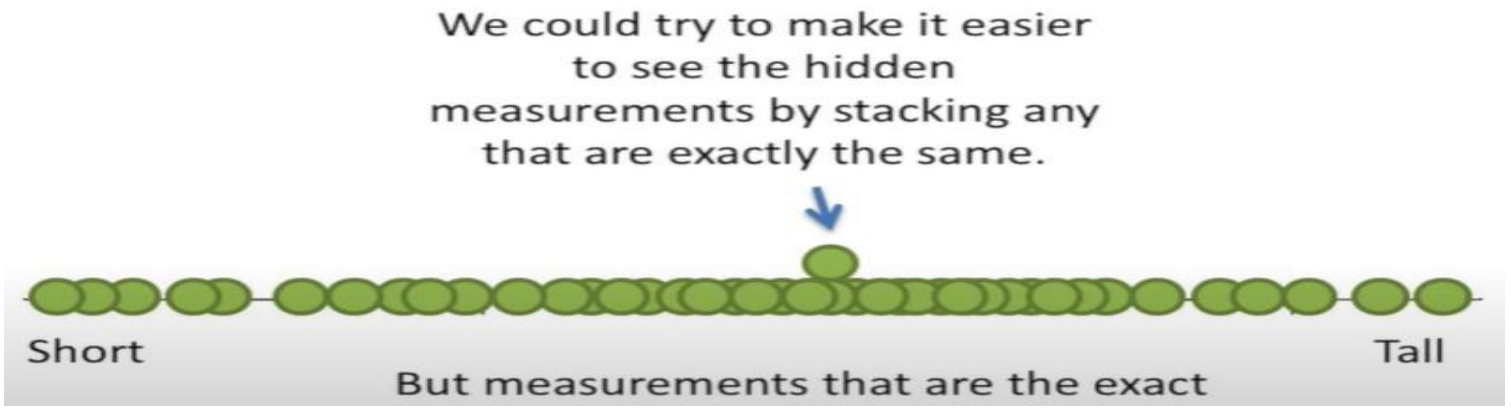




Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Let's get an idea

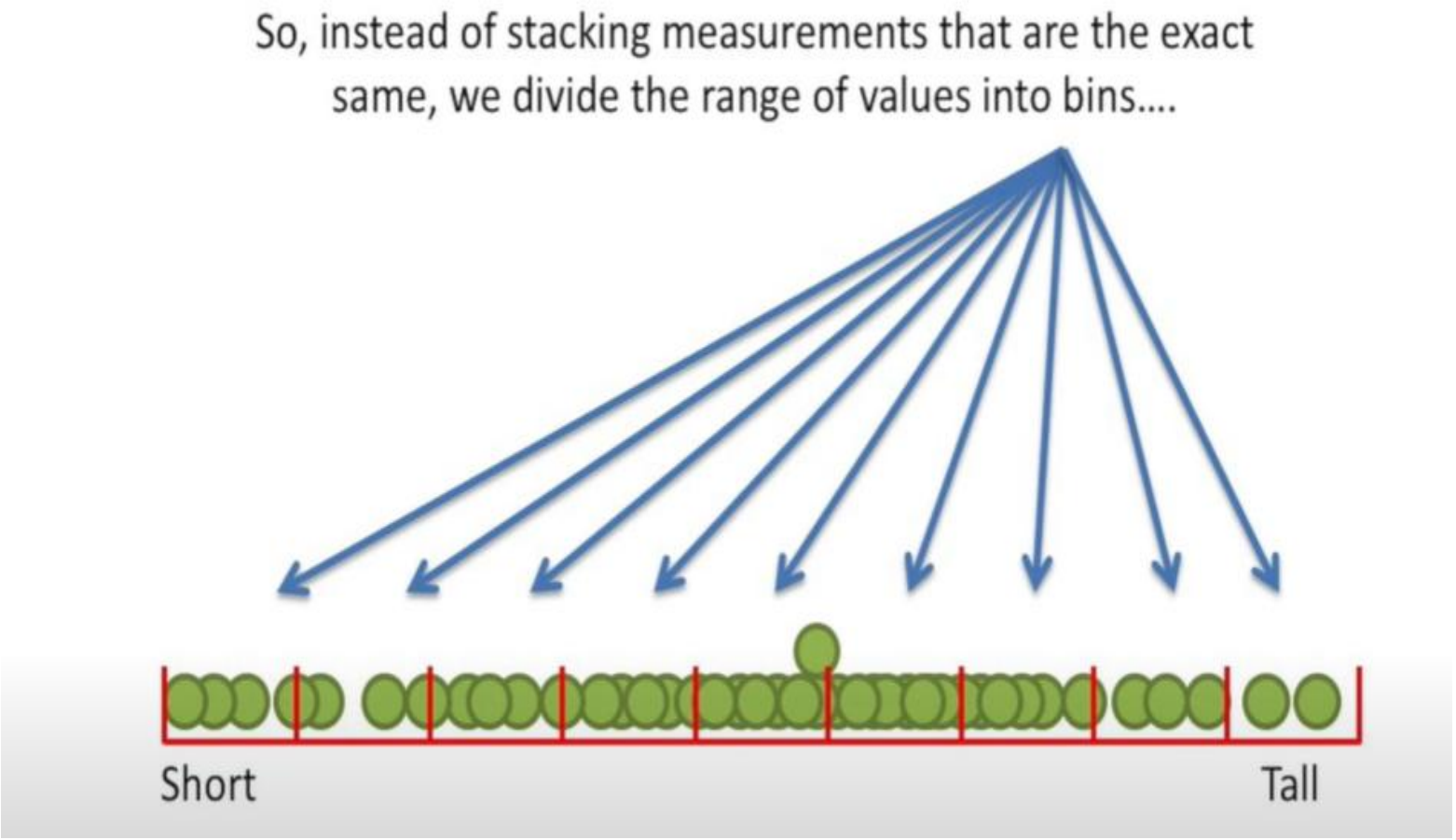



Source: StatQuest

# STATISTICS FOR DATA SCIENCE

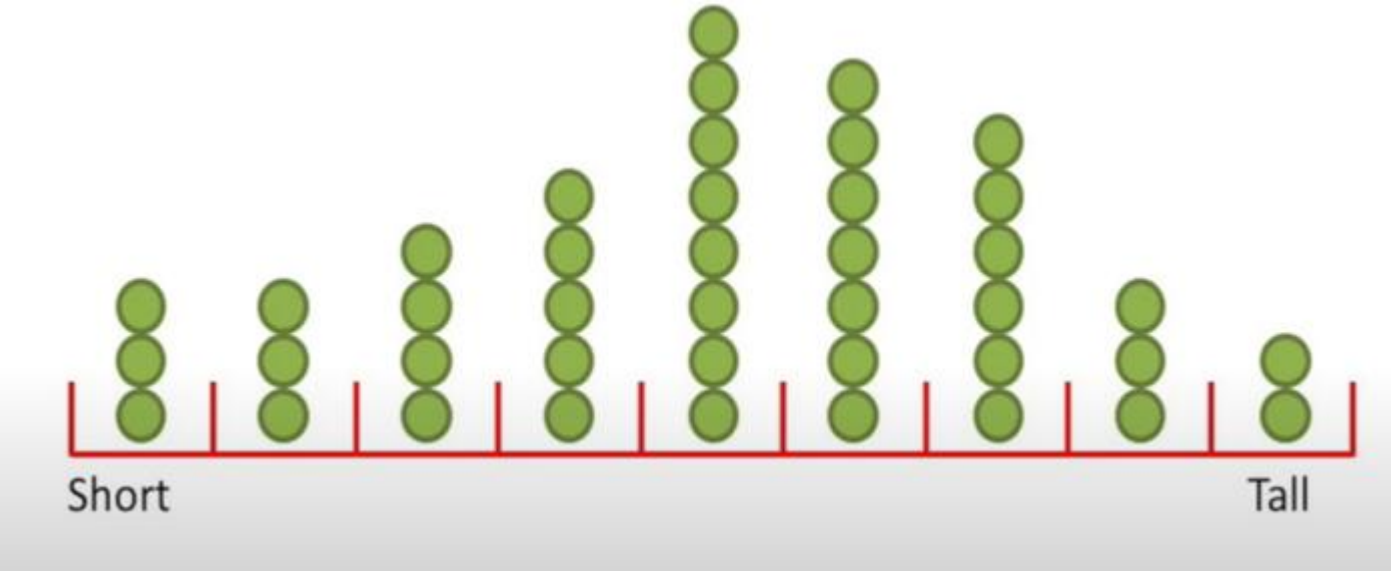
## Let’s get an idea




Source: StatQuest

# STATISTICS FOR DATA SCIENCE

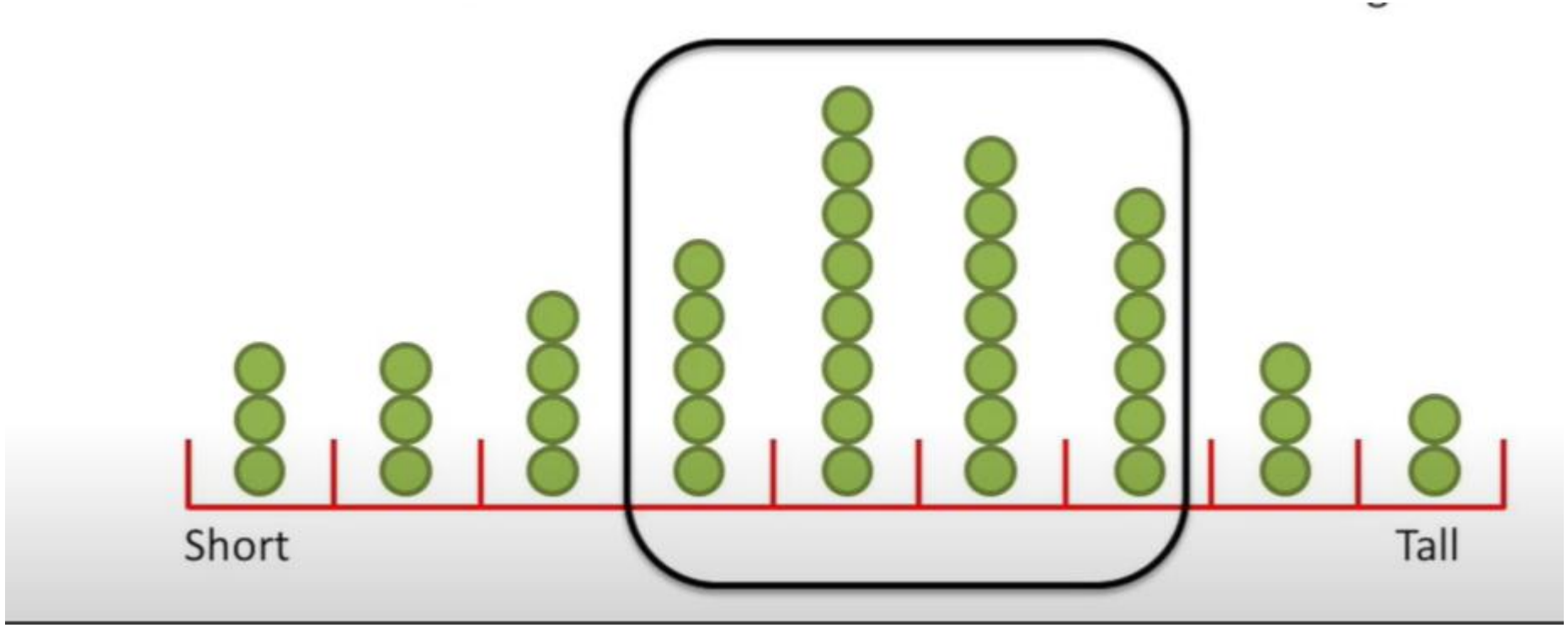
## Histogram




Source: StatQuest

# STATISTICS FOR DATA SCIENCE

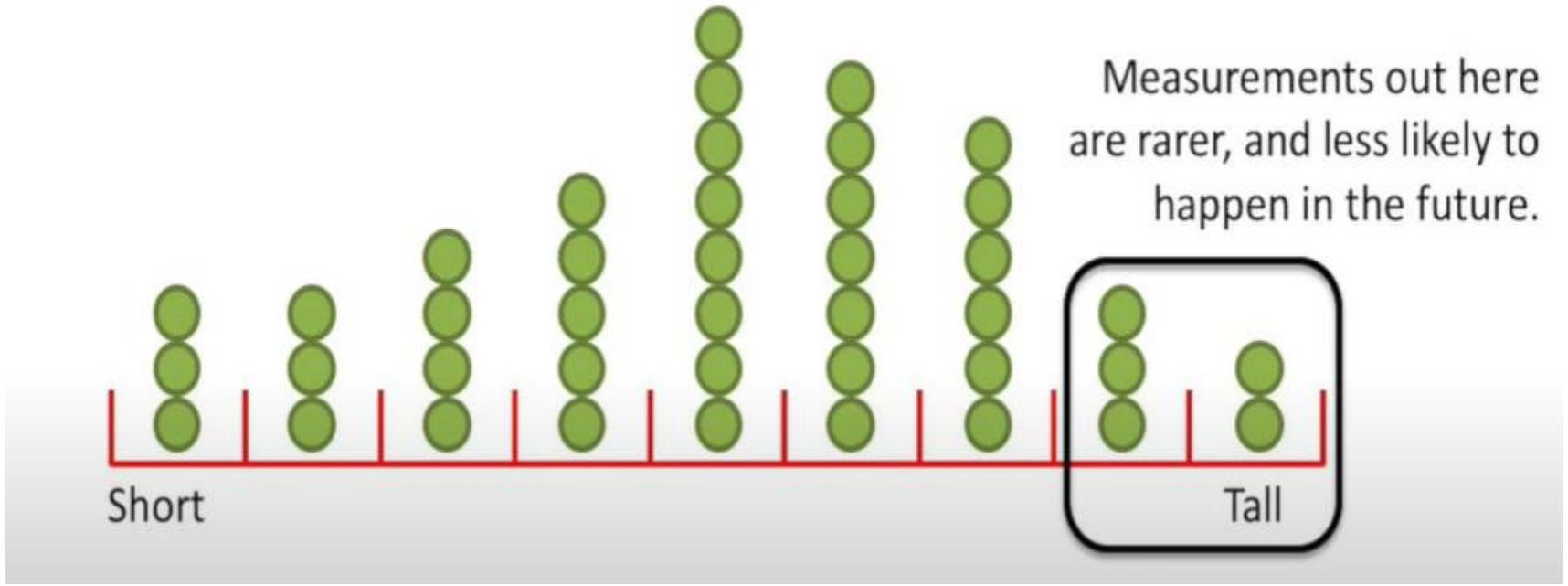
## Histogram




Source: StatQuest

# STATISTICS FOR DATA SCIENCE

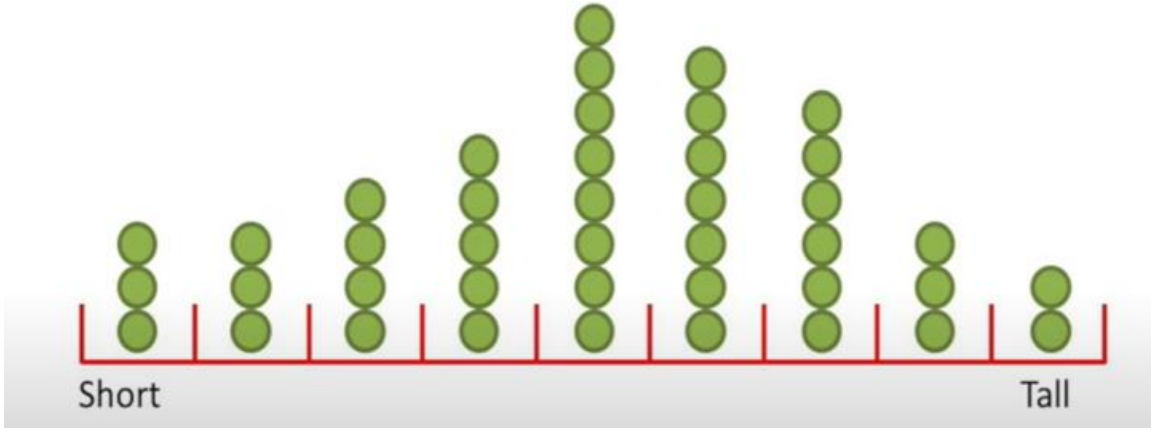
## Histogram

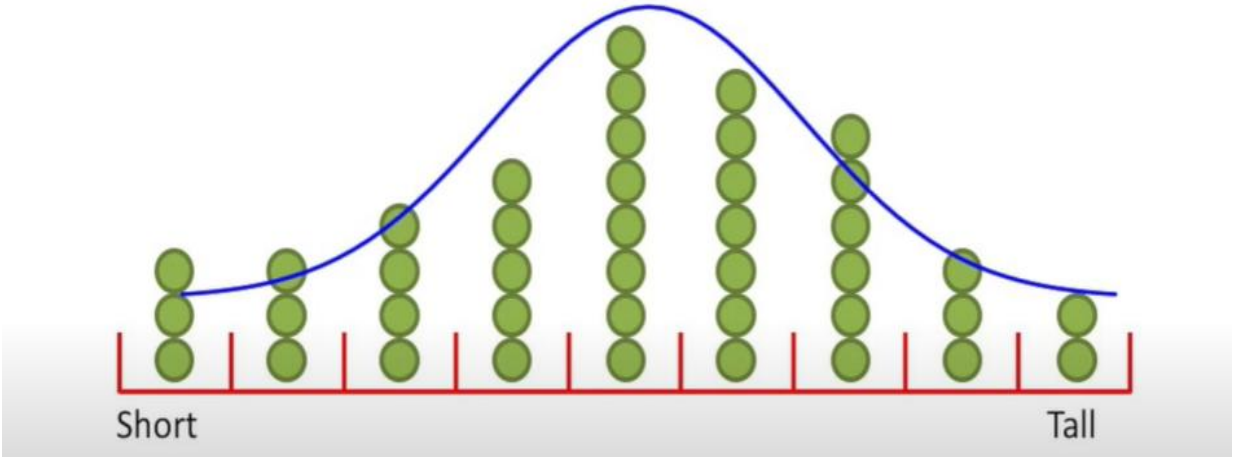



Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Data Distribution

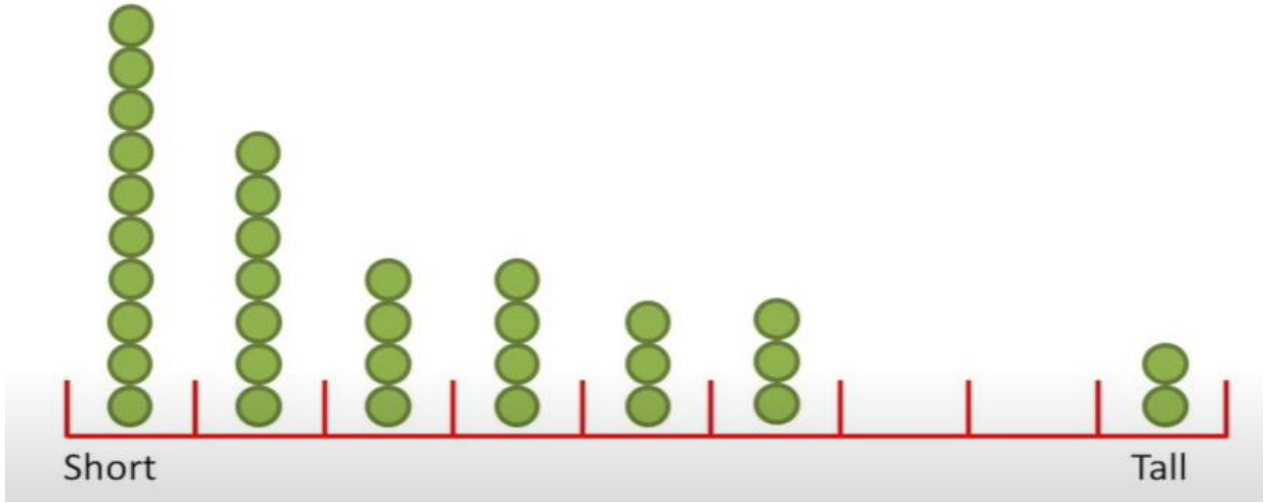


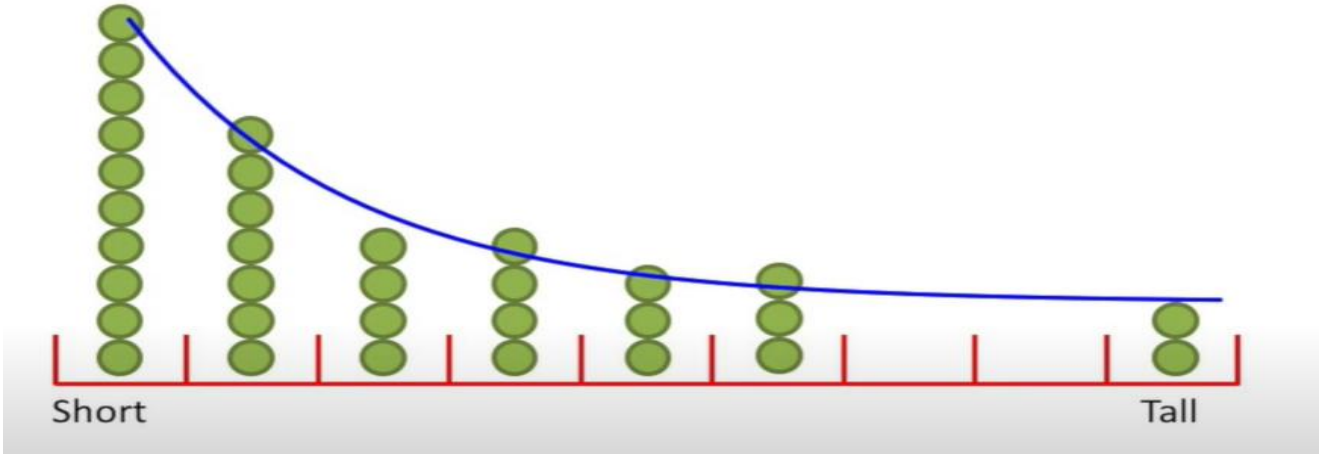




Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Data Distribution

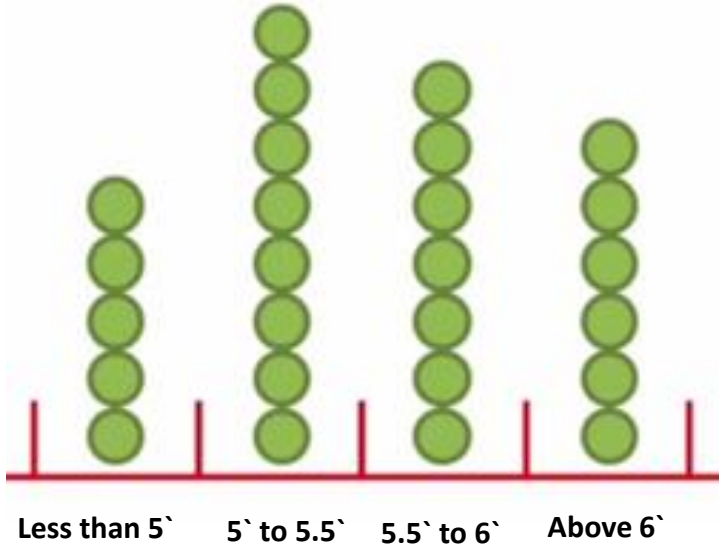



Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Histogram !!!!!!!!




**Note:** Figuring out how wide to make the bins is tricky!!!!

Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Bin Width

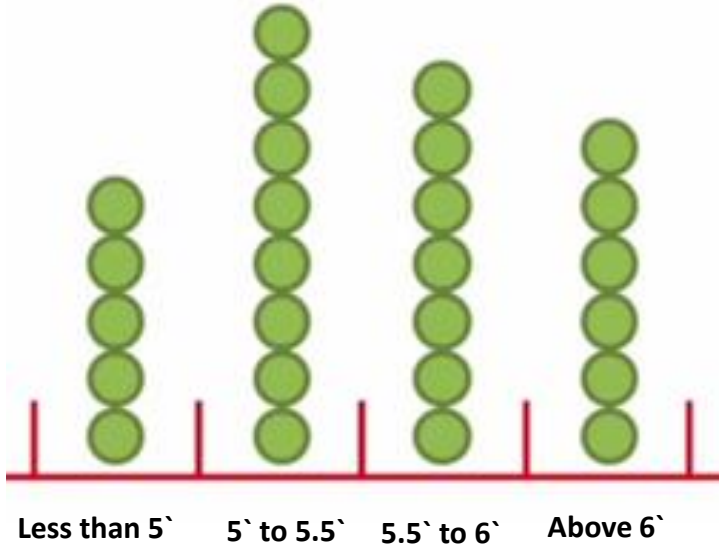

**If the bins are too narrow, then they are not much help**


**If the bins are too wide, then they are not much help**

Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Histogram !!!!!!



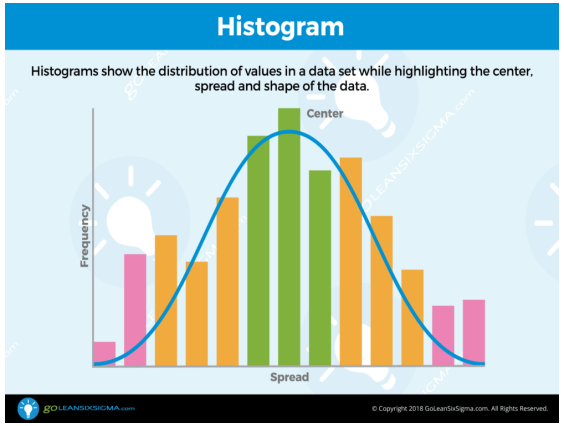

**Note:** Figuring out how wide to make the bins is tricky!!!!

Source: StatQuest

# STATISTICS FOR DATA SCIENCE

## Histogram

**Histograms** show the **distribution of data values** in a data set while highlighting the **center, spread and shape** of the data.




Histograms, also known as **Frequency Plots**, are a visual displays of **how much variation exists** in a process.

They highlight the **center of the data** measured as the **mean, median and mode.**

They highlight the **distribution of the data** measured as the **range and standard deviation.**

The **shape of a Histogram** indicates whether the distribution is **normal, bi-modal, or skewed.**

Source: goLEANSIXSIGMA.com

# STATISTICS FOR DATA SCIENCE

## Histogram

A histogram is used to **summarize discrete or continuous data.**

In other words, it provides a visual interpretation of numerical data by showing the **number of data points** that fall within a specified range of values (called "**bins**").

It is similar to a vertical bar graph.

However, a histogram, unlike a vertical bar graph, shows **no gaps between the bars.**

Source: corporatefinanceinstitute.com

## Histogram

A histogram is a **graphical display of data** using bars of different heights.

In a histogram, each bar groups **numbers into ranges**.

**Taller bars** show that **more data falls** in that **range.**

A histogram displays the **shape and spread** of **continuous** sample data.

A histogram is used to **summarize discrete or continuous data.**

## Importance of Histogram

Creating a histogram provides a visual representation of data distribution.

Histograms can display a large amount of data and the frequency of the data values.

The median and distribution of the data can be determined by a histogram.

In addition, it can show any outliers or gaps in the data.

Source: corporatefinanceinstitute.com

# STATISTICS FOR DATA SCIENCE

## Distributions of a Histogram

**A normal distribution:**

In a normal distribution, points on one side of the average are as likely to occur as on the other side of the average.

Source: corporatefinanceinstitute.com

## Distributions of a Histogram

**A bimodal distribution:** In a bimodal distribution, there are two peaks.

In a bimodal distribution, the data should be separated and analyzed as separate normal distributions.

Source: corporatefinanceinstitute.com

## Distributions of a Histogram

**A right-skewed distribution:** A right-skewed distribution is also called a positively skewed distribution.

In a right-skewed distribution, a large number of data values occur on the left side with a fewer number of data values on the right side.

A right-skewed distribution usually occurs when the data has a range boundary on the left-hand side of the histogram. For example, a boundary of 0.

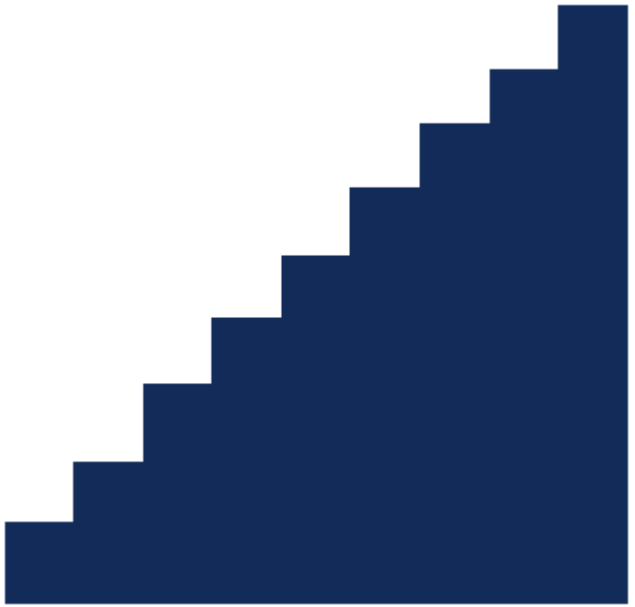
Source: corporatefinanceinstitute.com

## Distributions of a Histogram

**A left-skewed distribution:** A left-skewed distribution is also called a negatively skewed distribution.

In a left-skewed distribution, a large number of data values occur on the right side with a fewer number of data values on the left side.

A right-skewed distribution usually occurs when the data has a range boundary on the right-hand side of the histogram. For example, a boundary such as 100.



Source: corporatefinanceinstitute.com

# STATISTICS FOR DATA SCIENCE

## Distributions of a Histogram

**A random distribution:**

A random distribution lacks an apparent pattern and has several peaks.

In a random distribution histogram, it can be the case that different data properties were combined.

Therefore, the data should be separated and analyzed separately.

Source: corporatefinanceinstitute.com

## Distributions of a Histogram

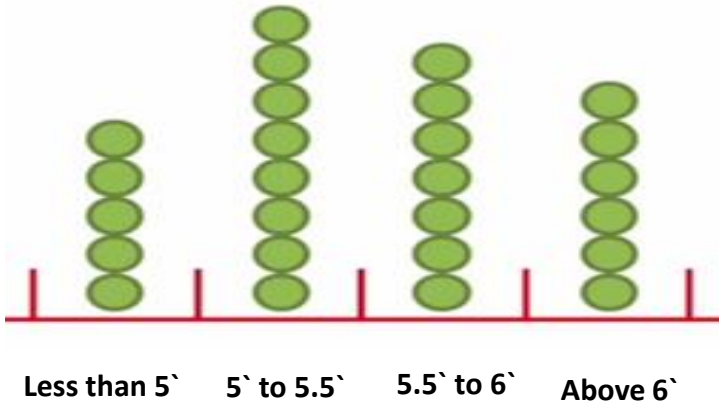
**A bimodal distribution:** In a bimodal distribution, there are two peaks.

In a bimodal distribution, the data should be separated and analyzed as separate normal distributions.

Source: corporatefinanceinstitute.com

# STATISTICS FOR DATA SCIENCE

## Histogram !!!!!




**Note: Figuring out how wide to make the bins is tricky!!!!**

Good to have more intervals rather than fewer.

Good to have large numbers of sample points in the intervals.

Source: StatQuest

## Bins in Histogram

Histograms are based on area, not height of bars.

In a histogram, it is the area of the bar that indicates the frequency of occurrences for each bin.

In statistics, the Freedman – Diaconis rule can be used to select the size of the bins to be used in a histogram

$$\text{Bin size} = 2\,\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

# STATISTICS FOR DATA SCIENCE

## Equal Bins / Class Widths

Suppose we are looking at the history grades of students in 10th grade and have the classes corresponding to letter grades: A, B, C, D, F. The number of each of these grades gives us a frequency for each class:

- 7 students with an F
- 9 students with a D
- 18 students with a C
- 12 students with a B
- 4 students with an A

**Frequency**

To determine the relative frequency for each class we first add the total number of data points: $7 + 9 + 18 + 12 + 4 = 50$. Next we, divide each frequency by this sum 50.

- 0.14 = 14% students with an F
- 0.18 = 18% students with a D
- 0.36 = 36% students with a C
- 0.24 = 24% students with a B
- 0.08 = 8% students with an A

**Relative Frequency**

Source: www.thoughtco.com

## Unequal Bin / Class Widths

Histograms are drawn with class intervals of differing widths rarely.

When the **class intervals** are of **unequal widths**, the **heights** of the **rectangles or bars** must be set equal to the densities.

**Compute** the **density** for each class, according to the formula

$$\text{Density} = \frac{\text{Relative Frequency}}{\text{Class Width}}$$

The **areas of the rectangles** will then be the **relative frequencies**.

# STATISTICS FOR DATA SCIENCE

## Parts of Histogram

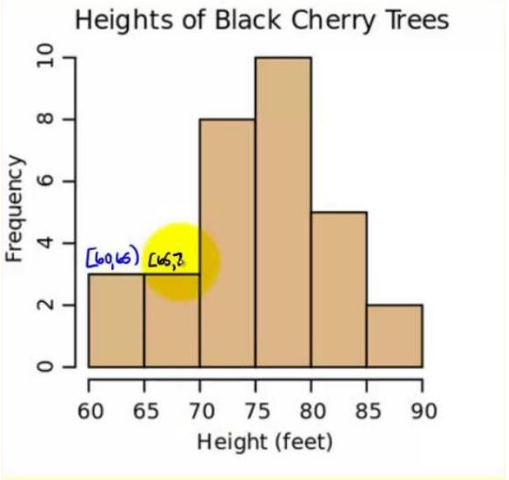
**The title:** The title **describes the information** included in the histogram.

**X-axis:** The X-axis are intervals that show the **scale of values** which the measurements fall under.

**Y-axis:** The Y-axis shows the **number of times** that the values occurred within the intervals set by the X-axis.

**The bars:** The **height of the bar** shows the **number of times** that the values occurred within the interval, while the **width** of the bar shows the **interval that is covered**.

For a histogram with **equal bins**, the **width** should be the **same** across all bars.




Source: Mathispower4u

## Creating a Histogram

Divide the range of values into series of intervals .

Check how many values falls into each intervals.

Bins are consecutive and non-overlapping intervals of a variable.

They must be adjacent and are often of equal size.

Width of each bin may or may not be equal.

If they’re equal then, the height of bins represents the frequency of data points in that  range.

## Example – Construct a Histogram

The weather in Los Angeles is dry most of the time, but it can be quite rainy in the winter. The rainiest month of the year is February. The following table presents the annual rainfall in Los Angeles, in inches, for each February from 1965 to 2006.

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.2 | 3.7 | 1.2 | 13.7 | 1.5 | 0.2 | 1.7 |
| 0.6 | 0.1 | 8.9 | 1.9 | 5.5 | 0.5 | 3.1 |
| 3.1 | 8.9 | 8.0 | 12.7 | 4.1 | 0.3 | 2.6 |
| 1.5 | 8.0 | 4.6 | 0.7 | 0.7 | 6.6 | 4.9 |
| 0.1 | 4.4 | 3.2 | 11.0 | 7.9 | 0.0 | 1.3 |
| 2.4 | 0.1 | 2.8 | 4.9 | 3.5 | 6.1 | 0.1 |

## Step:1 – Prepare the Data

Arrange the values in ascending order (number of data points (n) = 42)

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 |
| 0.3 | 0.5 | 0.6 | 0.7 | 0.7 | 1.2 | 1.3 |
| 1.5 | 1.5 | 1.7 | 1.9 | 2.4 | 2.6 | 2.8 |
| 3.1 | 3.1 | 3.2 | 3.5 | 3.7 | 4.1 | 4.4 |
| 4.6 | 4.9 | 4.9 | 5.5 | 6.1 | 6.6 | 7.9 |
| 8.0 | 8.0 | 8.9 | 8.9 | 11.0 | 12.7 | 13.7 |

## Step: 2 Identify the Bin Widths

By using the Freedman – Diaconis , the bin width / class intervals can be found.

$$\text{Bin Width} = 2\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

**Find the IQR (InterQuartile Range)**

$$\text{IQR} = Q_3 - Q_1$$

Quartile 1, $Q_1 = 0.25\,(n+1) = 0.25\,(43) = 10.75$

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 |
| 0.3 | 0.5 | 0.6 | 0.7 | 0.7 | 1.2 | 1.3 |

$$\frac{0.6 + 0.7}{2} = 0.65$$

## Step: 2 Conti..

Quartile 3, $Q_3 = 0.75\ (n+1) = 0.75\ (43) = 32.25$

| | | | | | | |
|---|---|---|---|---|---|---|
| 3.1 | 3.1 | 3.2 | 3.5 | 3.7 | 4.1 | 4.4 |
| 4.6 | 4.9 | 4.9 | 5.5 | 6.1 | 6.6 | 7.9 |

$$\frac{5.5 + 6.1}{2} = 5.8$$

IQR = 5.8 – 0.65 = 5.15

Substitute in the formula, lets find the Bin width

$$\frac{2 * 5.15}{\sqrt[3]{42}} = 2.9 = (\sim 3)$$

## Step: 3 Build the Frequency Distribution Table

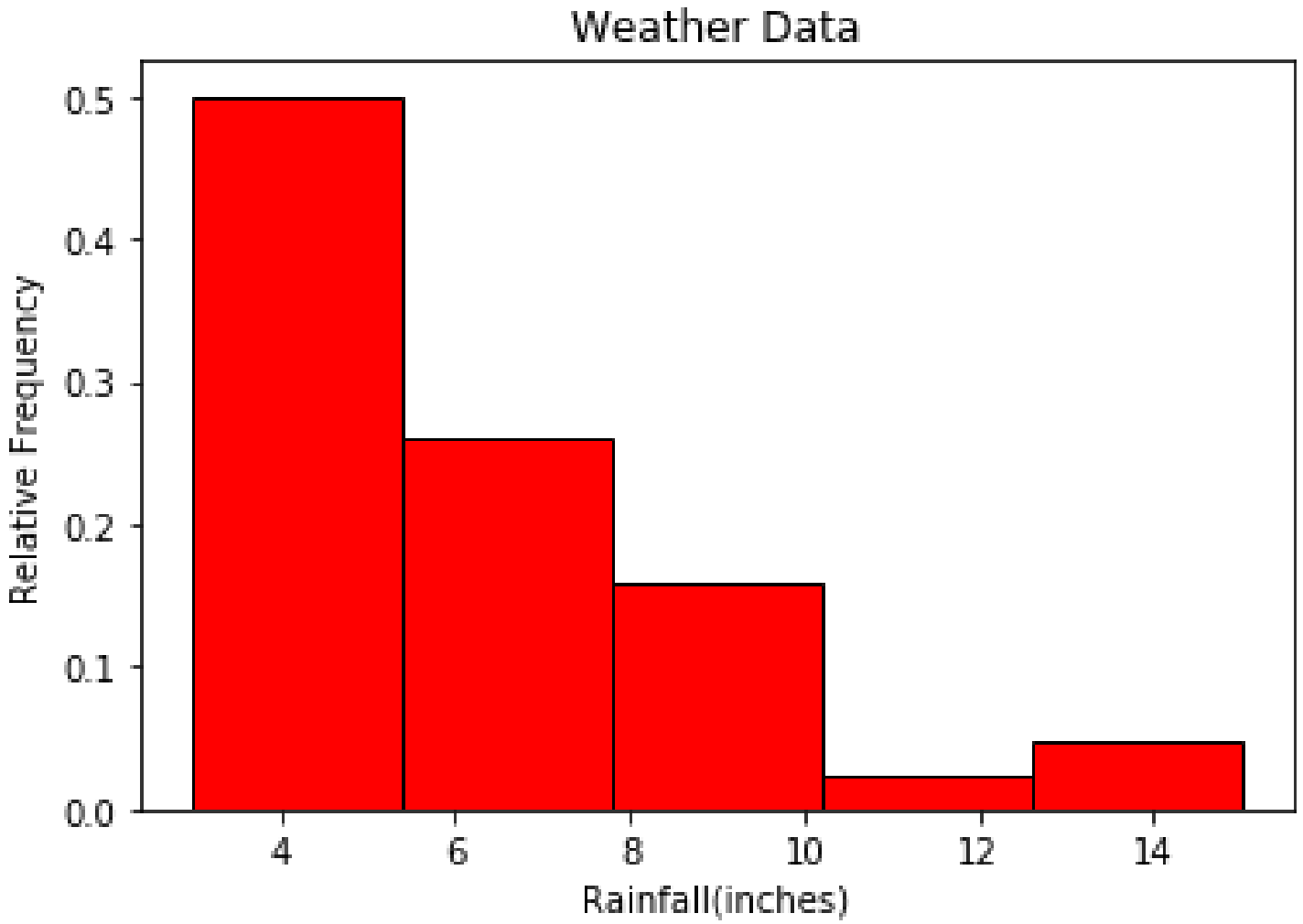
| Class | Frequency | Relative Frequency | Density |
|---|---|---|---|
| 0 – 3 | 21 | 0.5 | 0.1667 |
| 3 – 6 | 11 | 0.2619 | 0.0873 |
| 6 – 9 | 7 | 0.1667 | 0.0555 |
| 9 – 12 | 1 | 0.0238 | 0.0073 |
| 12 - 15 | 2 | 0.0476 | 0.0159 |
| | Sum = 42 | Sum = 1 | |

## Step: 4 Find the number of Bins / Buckets

Number of bins / buckets = $\frac{\text{Max} - \text{Min}}{\text{Bin Width}}$

$$\frac{15 - 0}{3} = 5$$

## Step: 5 Plot the Histogram



## Histogram

Choose boundary points for the class intervals.

Compute the frequency and relative frequency for each class.

Compute the density for each class, according to the formula

$$\text{Density} = \frac{\text{Relative Frequency}}{\text{Class Width}}$$

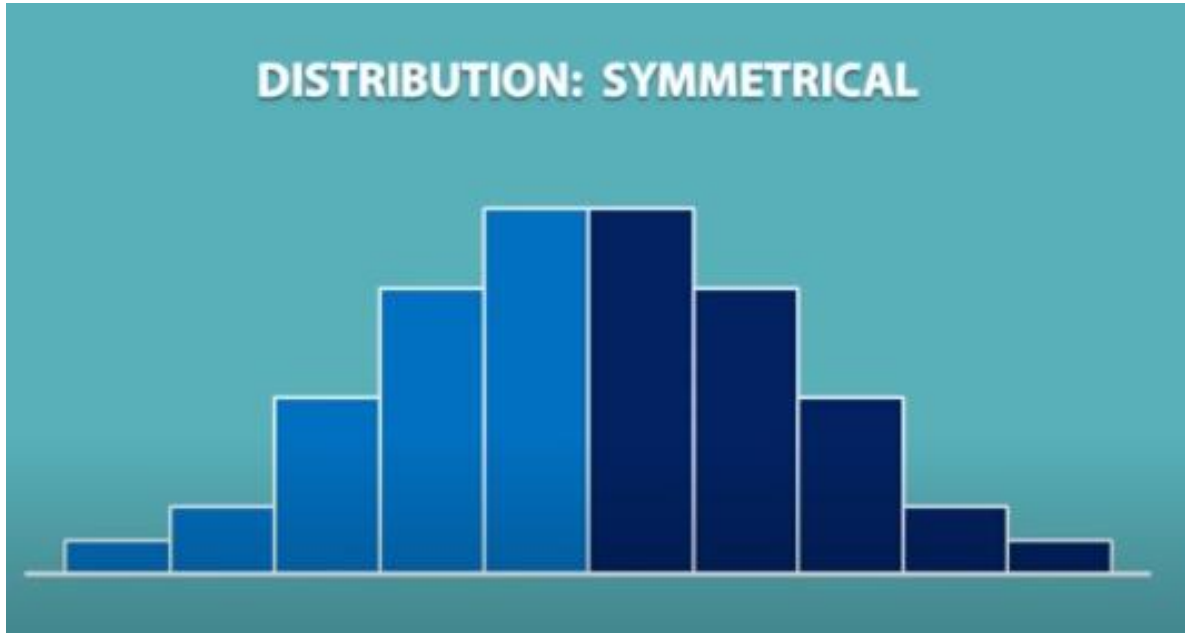
Draw a rectangle for each class. If the classes all have the same width, the heights of the rectangles may be set equal to the frequencies, the relative frequencies, or the densities.

If the classes do not all have the same width, the heights of the rectangles must be set equal to the densities.

# STATISTICS FOR DATA SCIENCE

## Symmetry

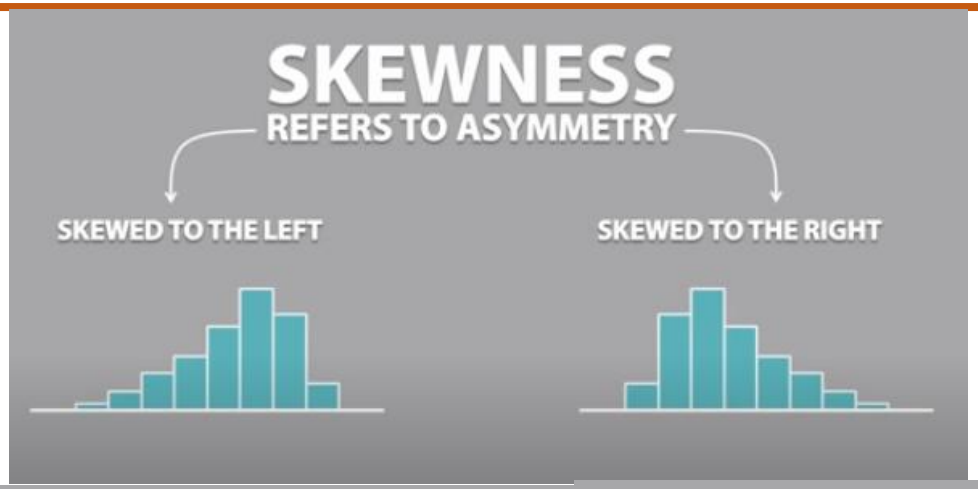
A distribution is said to be symmetrical if it can be divided into two equal sizes of the same shape.




Source: Simple Learning Pro

# STATISTICS FOR DATA SCIENCE

## Skewness



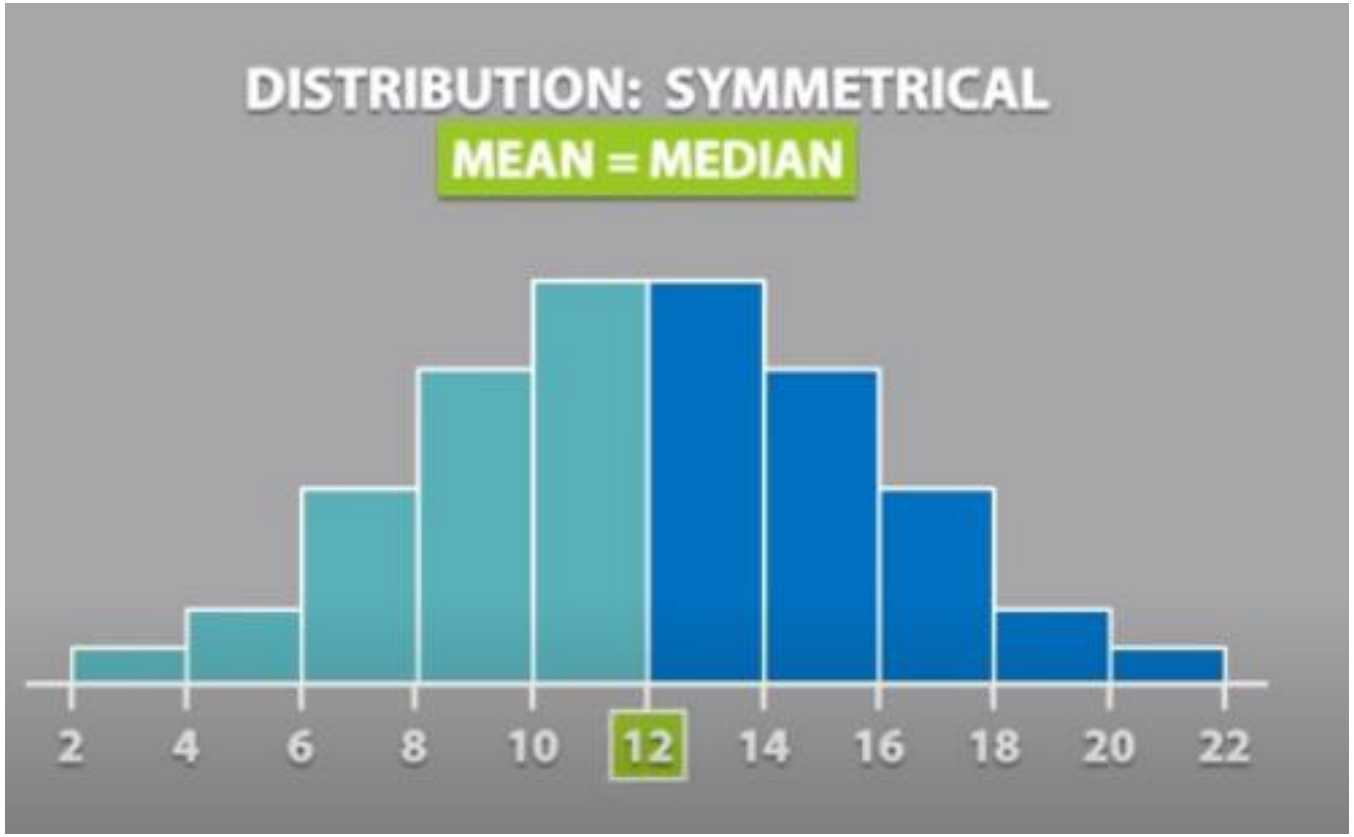


Source: Simple Learning Pro

# STATISTICS FOR DATA SCIENCE

## Distribution - Symmetry

When a histogram is roughly **symmetric**, the mean and the median are approximately equal.

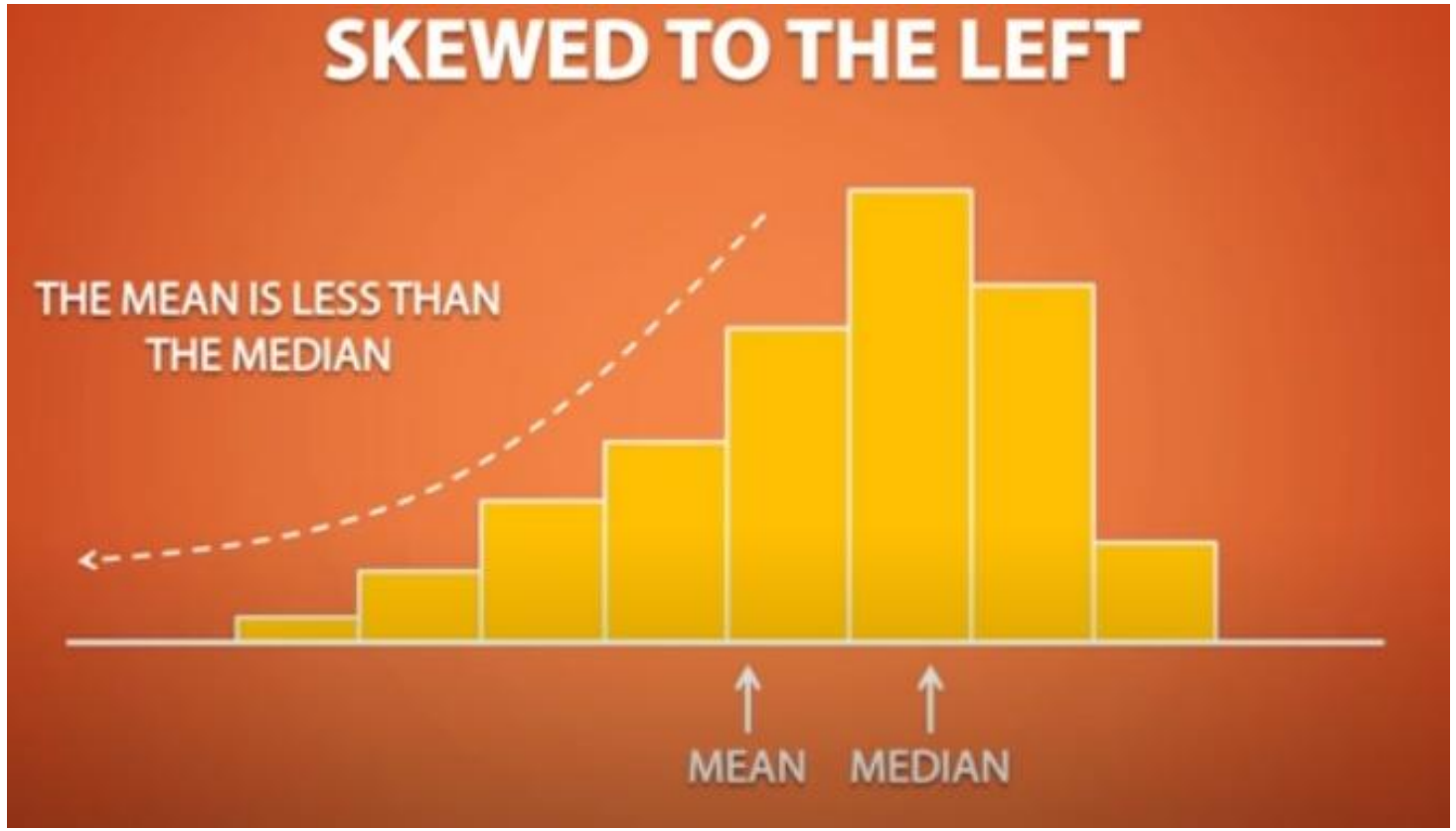



Source: Simple Learning Pro

# STATISTICS FOR DATA SCIENCE

## Distribution – Left Skewed

When a histogram is **left-skewed**, the mean is less than the median.

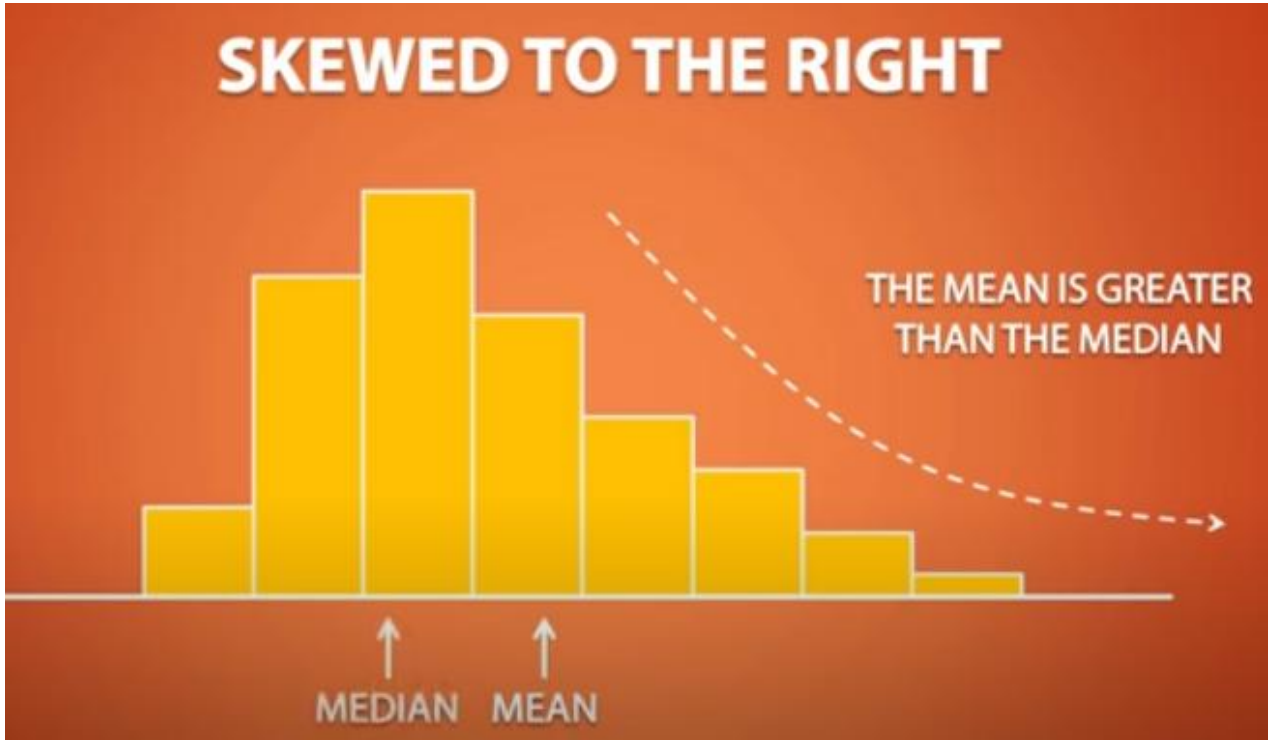



Source: Simple Learning Pro

# STATISTICS FOR DATA SCIENCE

## Distribution – Right Skewed

When a histogram is **right-skewed**, the mean is greater than the median.

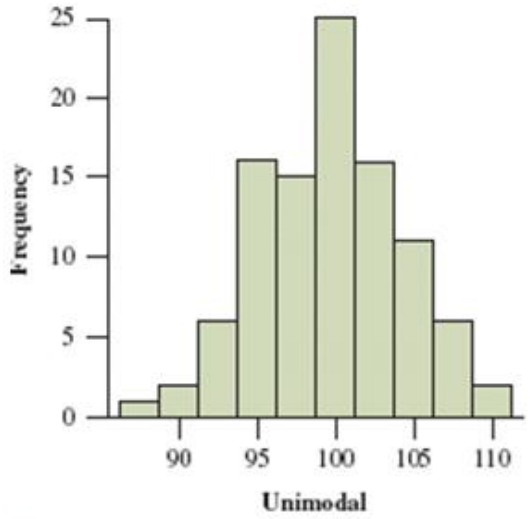



Source: Simple Learning Pro

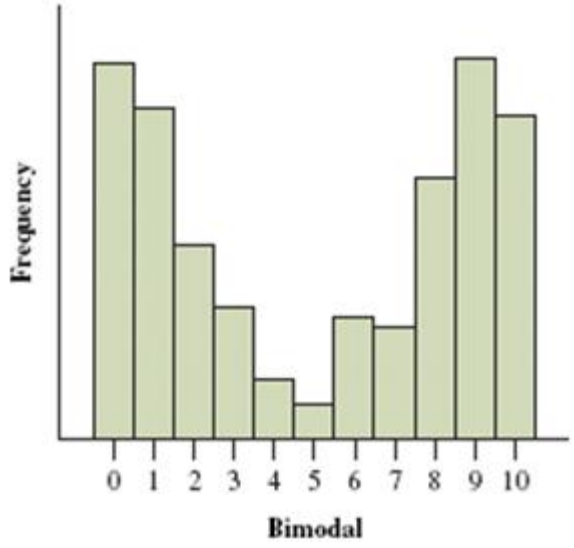
# STATISTICS FOR DATA SCIENCE

## Unimodal and Bimodal Histogram

A histogram with only one peak is what we call **unimodal**.

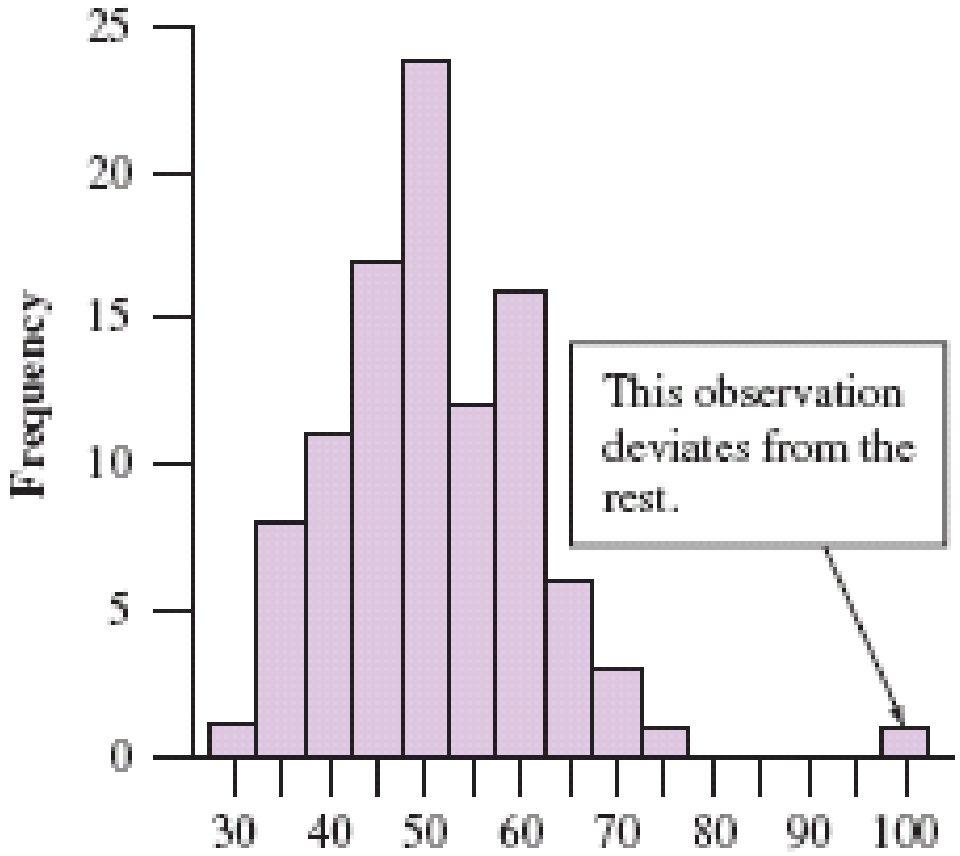



If a histogram has two peaks then we say that it is **bimodal**.



## Outlier

An outlier falls far from the rest of the data.




# THANK YOU

**D. Uma**

Department of Computer Science and Engineering

**umaprabha@pes.edu**

+91 99 7251 5335