Boxplot (or Box-and-Whisker Plot)

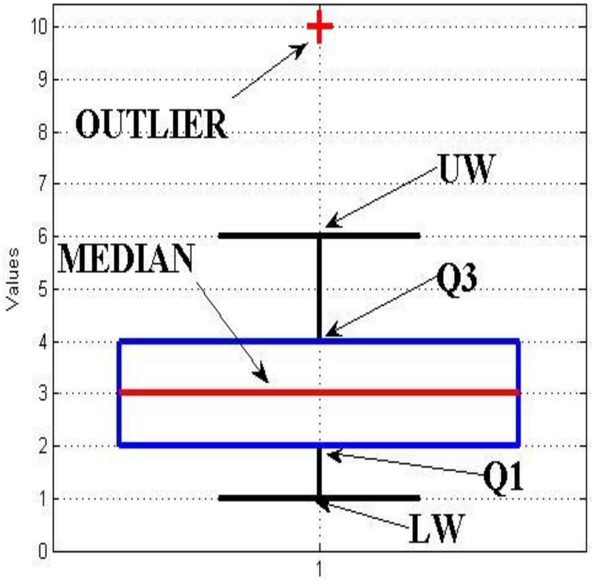
- Summarizes data into a *"5-number"* summary: **median, the first and the third quartiles (Q1 and Q3), minimum, and maximum**.

- Detects extreme observations (**outliers).**

- The centerline of the box marks the **median.**

Boxplot

- **Step 1:** Sort the data.
- **Step 2:** Compute median.
- **Step 3:** Compute quartiles Q1 and Q3
- **Step 4:** Compute IQR and identify whiskers.
- **Step 5:** Draw the boxplot.



Example

Data ID	1	2	3	4	5	6	7	8	9	10
Data(X)	10	3	1	6	2	3	4	2	3	4
Sorted(X)	1	2	2	3	3	3	4	4	6	10

Step 1: Sort the data.
Step 2: Compute median: n=10 is even

$$\tilde{X} = \frac{X_{(n/2)} + X_{(n/2)+1}}{2} = \frac{3+3}{2} = 3$$

Example

Data ID	1	2	3	4	5	6	7	8	9	10
Sorted(X)	1	2	2	3	3	3	4	4	6	10

Step 3: Compute quartiles Q1 and Q3.

Recall that the ***positions of the quartiles*** are determined by the following formula (D'Agostino, p.37):

$$pos = \begin{cases} \left[\frac{n+3}{4}\right], when\ n\ is\ odd \\ \left[\frac{n+2}{4}\right], when\ n\ is\ even \end{cases}.$$

The quartiles are in the **position = *pos*** from the top (**Q3**) and bottom (**Q1**) of the ordered data set, hence Q1=2 and Q2 = 4.

Example

Data ID	1	2	3	4	5	6	7	8	9	10
Sorted(X)	1	2	2	3	3	3	4	4	6	10

Step 4: Compute IQR and identify whiskers.

IQR = Q3- Q1 = 4-2 =2

Lower Bound = Q1 - 1.5*IQR = 2-1.5*2 = -1

Lower Whisker (LW) equals to <u>minimum</u> data observation value that is <u>greater</u> than or equal to **Lower Bound**. LW = 1

Example

Data ID	1	2	3	4	5	6	7	8	9	10
Sorted(X)	1	2	2	3	3	3	4	4	6	10

Step 4: Compute IQR and identify whiskers.
IQR = Q3- Q1 = 4-2 =2

Upper Bound = Q3 + 1.5*IQR = 4+1.5*2 = 7
Upper Whisker (UW) equals to maximum data observation value that is less than or equal to **Upper Bound.** UW = 6

Values greater than Upper Bound or less than Lower Bound are considered to be **outliers.**

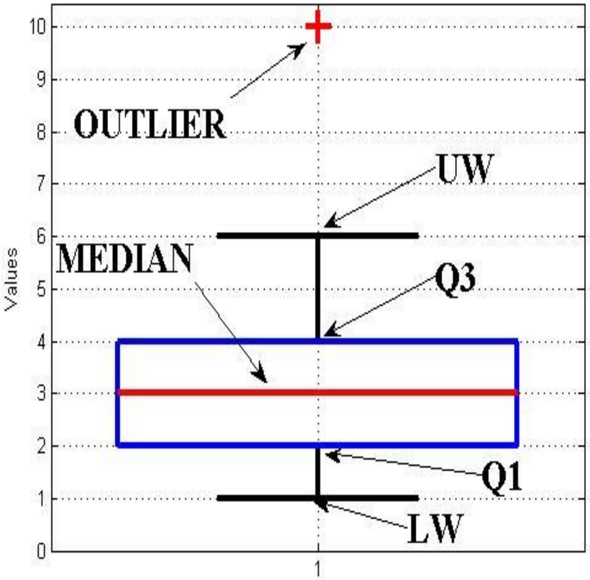
Example

Data ID	1	2	3	4	5	6	7	8	9	10
Sorted(X)	1	2	2	3	3	3	4	4	6	10

Step 5: Draw the boxplot.

Median = 3

Q1 = 2
Q3 = 4
Lower Whisker = 1
Upper Whisker = 6
Outlier = 10



Histogram

- Displays the distribution of a quantitative variable by showing the frequencies (counts) the values that fall in various *classes*.
  - For **continuous** variables, the classes are typically intervals of numbers that cover the full range of the variable.
- Determines **the shape of distribution** and helps to assess the **symmetry, modality, center, and spread.**

Example

Data ID	1	2	3	4	5	6	7	8	9	10	11
Data(X)	12	40	27	15	31	21	34	40	35	37	45
Sorted(X)	12	15	21	27	31	34	35	37	40	40	45

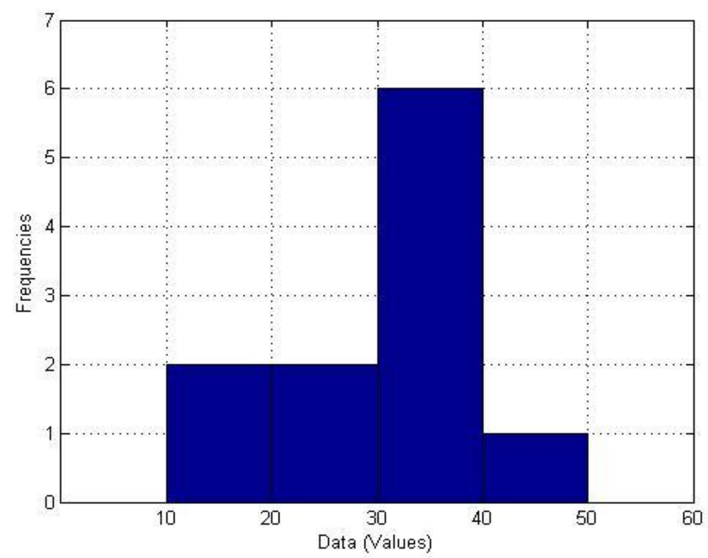
Frequency Class	Frequency
10 - 19	2
20 - 29	2
30 - 39	4
40 - 49	3

- **Step 1:** Sort the data.
- **Step 2:** Convert your data into Frequency Table.
- **Step 3:** Draw the histogram.

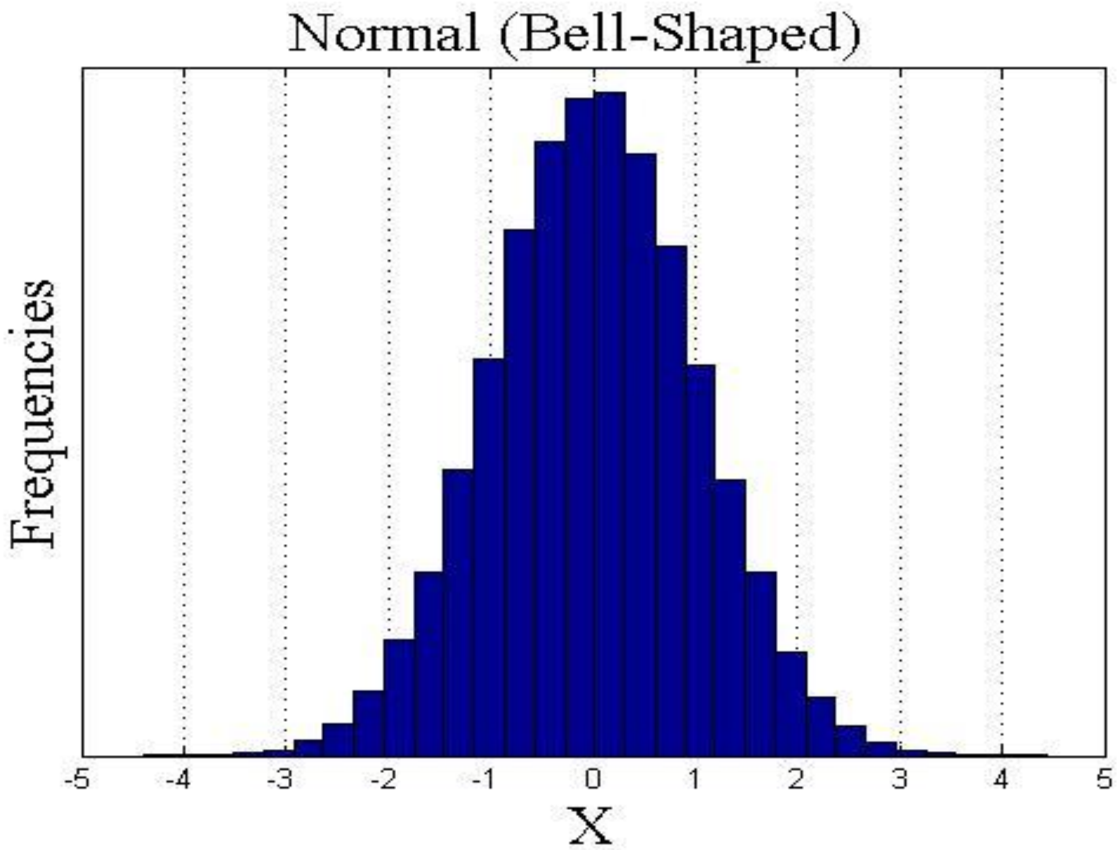
Example

Data ID	1	2	3	4	5	6	7	8	9	10	11
Data(X)	12	40	27	15	31	21	34	40	35	37	45
Sorted(X)	12	15	21	27	31	34	35	37	40	40	45

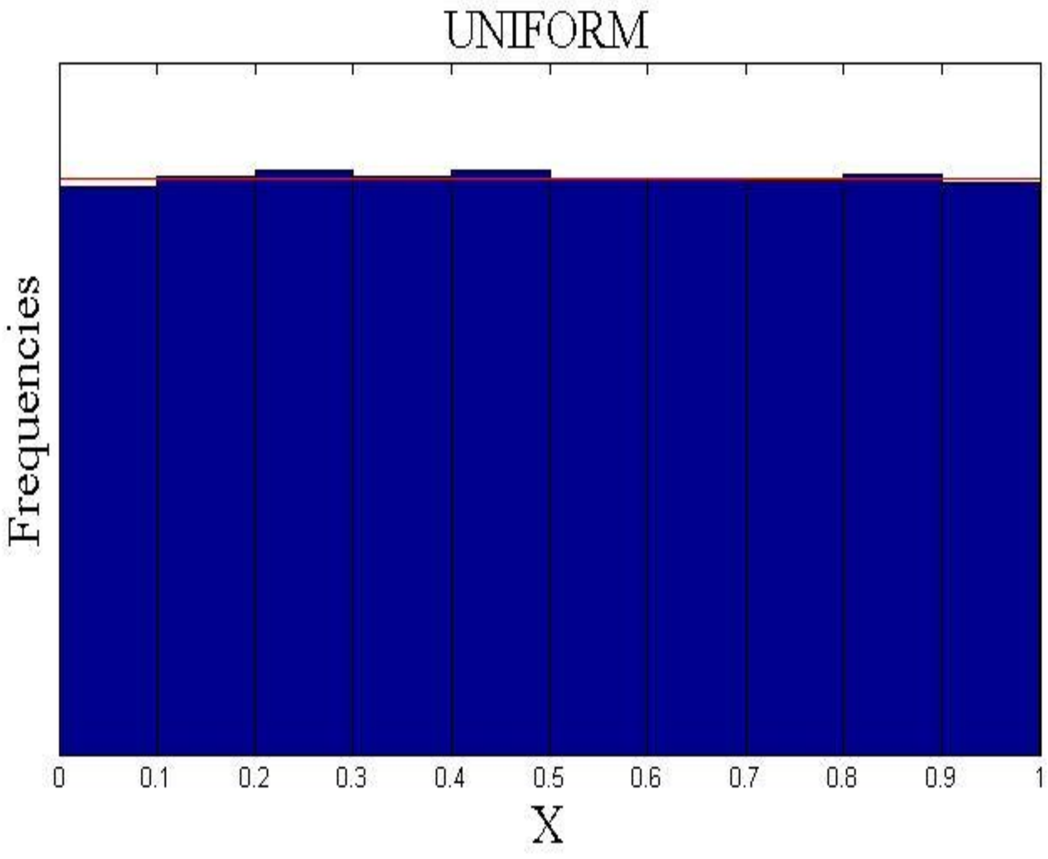
Frequency Class	**Frequency**
10 - 19	2
20 - 29	2
30 - 39	4
40 - 49	3



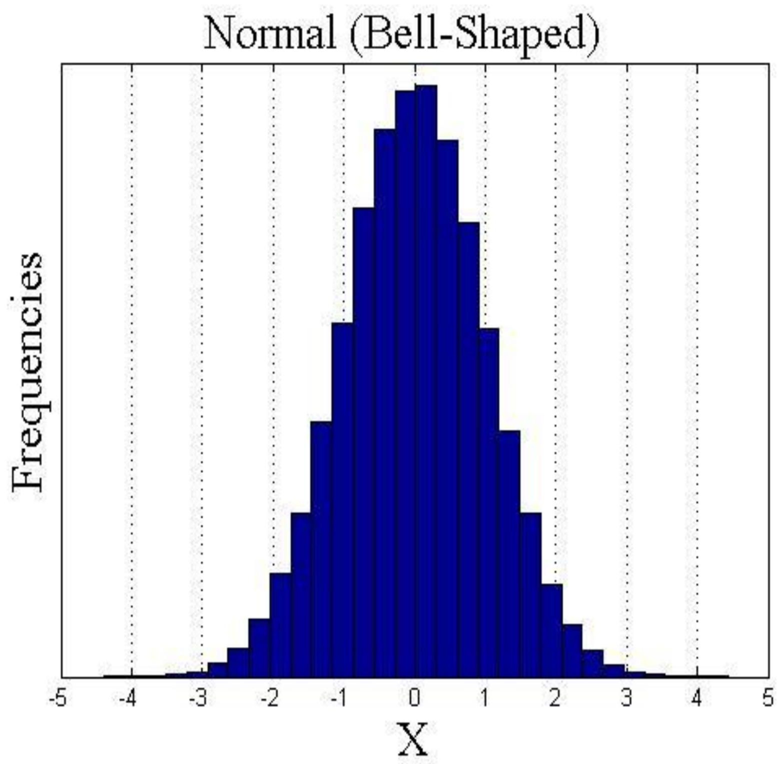
The Shapes of the Distribution

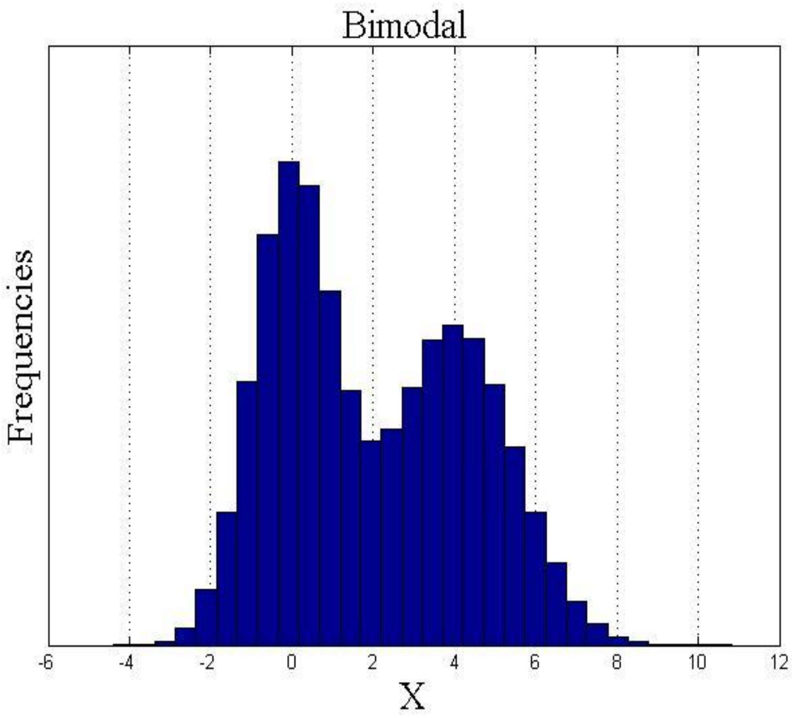


The Shapes of the Distribution

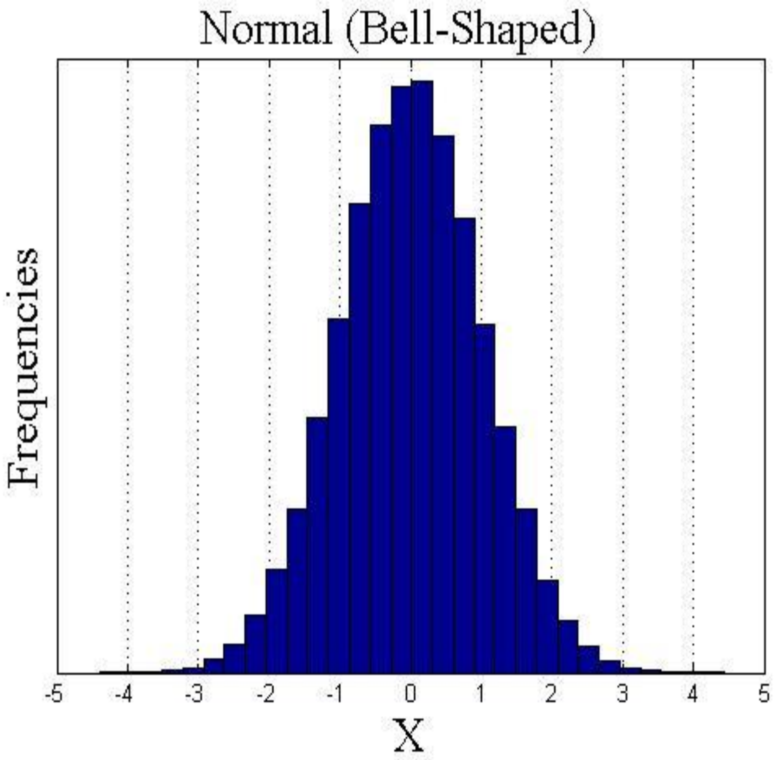


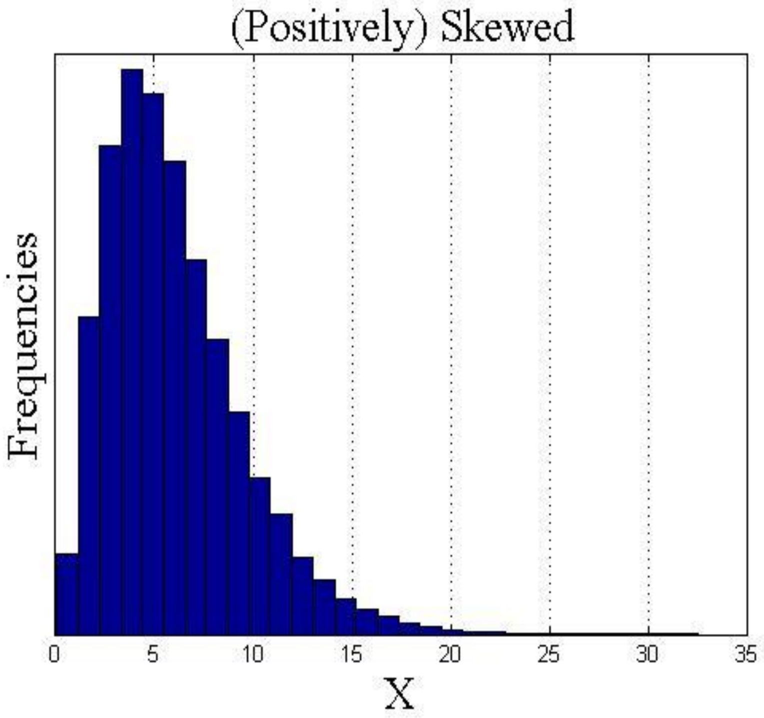
Unimodal vs. Bimodal





Symmetrical vs. Skewed





Symmetrical vs. Skewed

- **The relationship between mean, median and the shape of the distribution:**

  http://onlinestatbook.com/stat_sim/descriptive/index.html

Symmetrical vs. Skewed

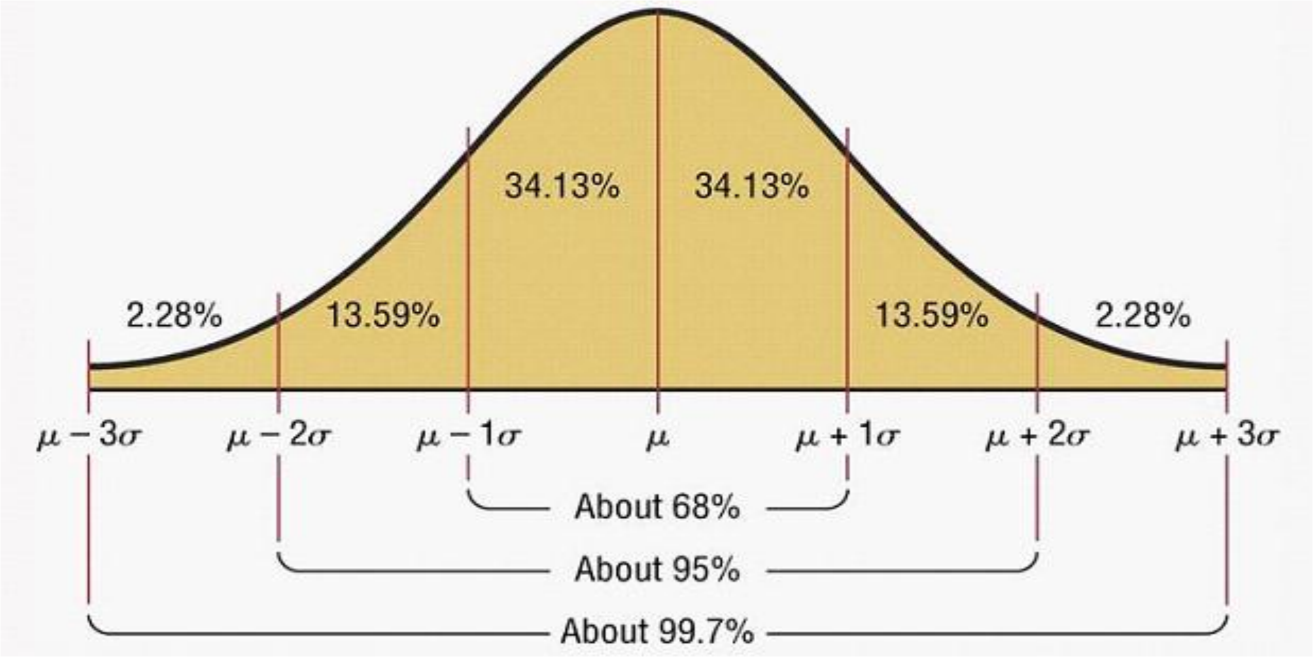
- In a **<u>symmetric</u>** distribution, **the mean = the median**.

- In a **positively (right) skewed** distributions (with longer tails to the right), **the mean $\geq$ the median**.

- In a **negatively (left) skewed** distributions (with longer tails to the left), **the mean $\leq$ the median**.

Empirical Rule for Normal Distribution

Empirical Rule states that for a **normal (bell-shaped) distribution**, nearly all values lie *within* **3 standard deviations of the mean**.



Experiment: Random vs. Deterministic

- An **experiment** is defined as a process, by which *observations are made*, or as a procedure that *generates* specific type of *outcome (data).*

  - In **deterministic experiment**, the same outcome is observed each time the experiment is performed.

  - In **random experiment**, one of several (random) outcomes is observed each time the experiment is observed.

Deterministic Experiment

- In **deterministic experiment**, the result is predictable with certainty and is known prior to its conduct.
- Examples:
  - An Experiment conducted to verify the Newton's Laws of Motion.
  - An Experiment conducted to verify the Economic Law of Demand.
  - (More Examples)______________________________

Random Experiment

- In **random experiments,** the result is unpredictable , unknown prior to its conduct, and can be one of several choices.
- Examples:
  - The Experiment of tossing a coin (head, tail)
  - The Experiment of rolling a die (1,2,3,4,5,6)
  - (More Examples) ______________________________

Sample Space

- The enumeration of all possible outcomes of an experiment is called the **sample space**, denoted **S.**

E.g.: S={head, tail}

- Collection of some outcomes is called an **event** and usually denoted with capital letters (e.g., A, B, C).

- Individual events are called **simple events.**

E.g.:{head}, {tail}