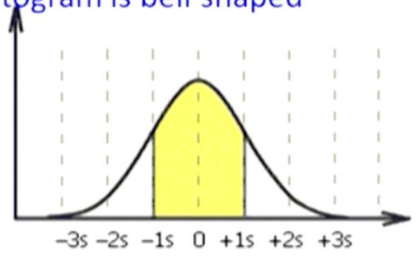
DATA ANALYTICS

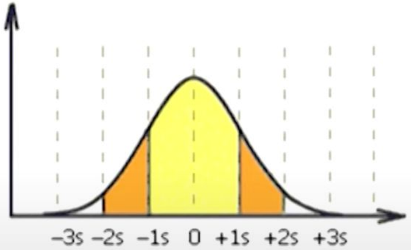
Lecture No: 04

Central Tendency and Dispersion

The Empirical Rule… If the histogram is bell shaped

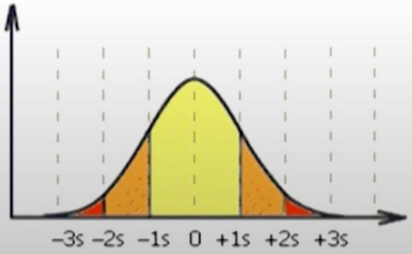
- Approximately 68% of all observations fall within **one** standard deviation of the mean.





- Approximately 95% of all observations fall within **two** standard deviations of the mean.

- Approximately 99.7% of all observations fall within **three** standard deviations of the mean.



Empirical Rule

- Data are normally distributed (or approximately normal)

Distance from the Mean	Percentage of Values Falling Within Distance
$\mu \pm 1\sigma$	68
$\mu \pm 2\sigma$	95
$\mu \pm 3\sigma$	99.7

Chebychev's Theorem— Not often used because interval is very wide

- A more general interpretation of the standard deviation is derived from **Chebychev's Theorem**, which applies to all shapes of histograms (not just the bell shaped).
- The proportion of observations in any sample that lie within **k** standard deviations of the mean is at least:

$$1-\frac{1}{k^2}\ for\ k > 1$$

For k=2 (for instance), the theorem states that atleast ¾ of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%)

Coefficient of variation

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion.

$$Coefficient\ of\ \ Variation = \frac{\sigma}{\mu}(100)$$

Coefficient of Variation

$$\mu_1 = 29$$
$$\sigma_1 = 4.6$$
$$C.V._{1} = \frac{\sigma_1}{\mu_1}(100)$$
$$= \frac{4.6}{29}(100)$$
$$= 15.86$$

$$\mu_2 = 84$$
$$\sigma_2 = 10$$
$$C.V._{2} = \frac{\sigma_2}{\mu_2}(100)$$
$$= \frac{10}{84}(100)$$
$$= 11.90$$

Variation and Standard deviation of Grouped Data

Population

$$\sigma^2 = \frac{\sum f\left(M - \mu\right)^2}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

Sample

$$S^2 = \frac{\sum f\left(M - \overline{X}\right)^2}{n-1}$$

$$S = \sqrt{S^2}$$

Population Variation and standard deviation of Grouped data (μ= 43)

Class Interval	f	M	fM	$M-\mu$	$(M-\mu)^2$	$f(M-\mu)^2$
20-under 30	6	25	150	-18	324	1944
30-under 40	18	35	630	-8	64	1152
40-under 50	11	45	495	2	4	44
50-under 60	11	55	605	12	144	1584
60-under 70	3	65	195	22	484	1452
70-under 80	1	75	75	32	1024	1024
	50		2150			7200

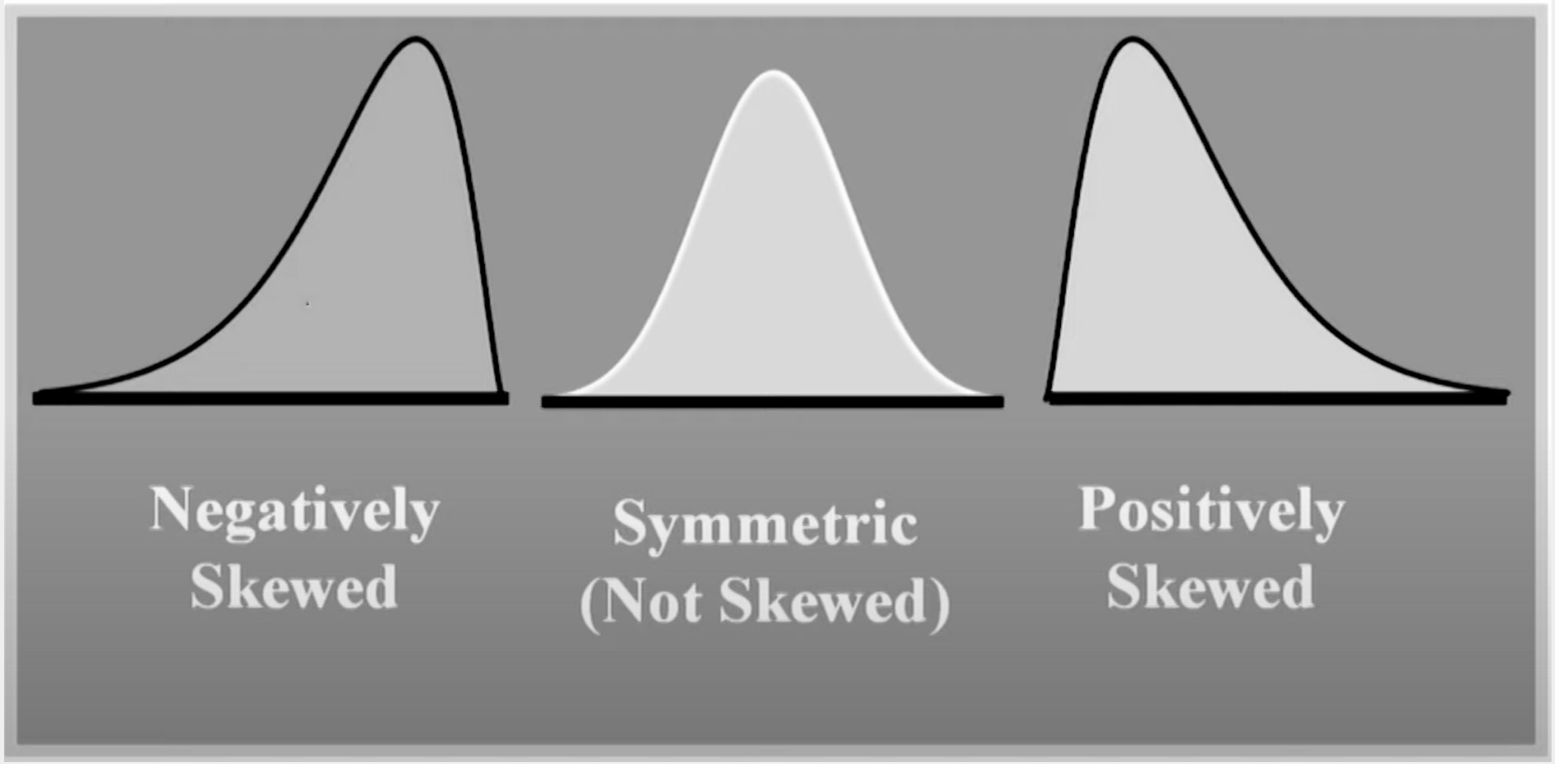
$$\sigma^2 = \frac{\sum f(M-\mu)^2}{N} = \frac{7200}{50} = 144$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{144} = 12$$

Measures of Shape

- **Skewness**
  - Absence of Symmetry
  - Extreme values in one side of a distribution
- **Kurtosis**
  - Peakness of the distribution
    - **Leptokurtic:** high and thin.
    - **Mesokurtic**: normal shape
    - **Platykurtic**: flat and spread out.
- **Box and whisker plots:**
  - Graphic display of a distribution.
  - Reveals skewness.

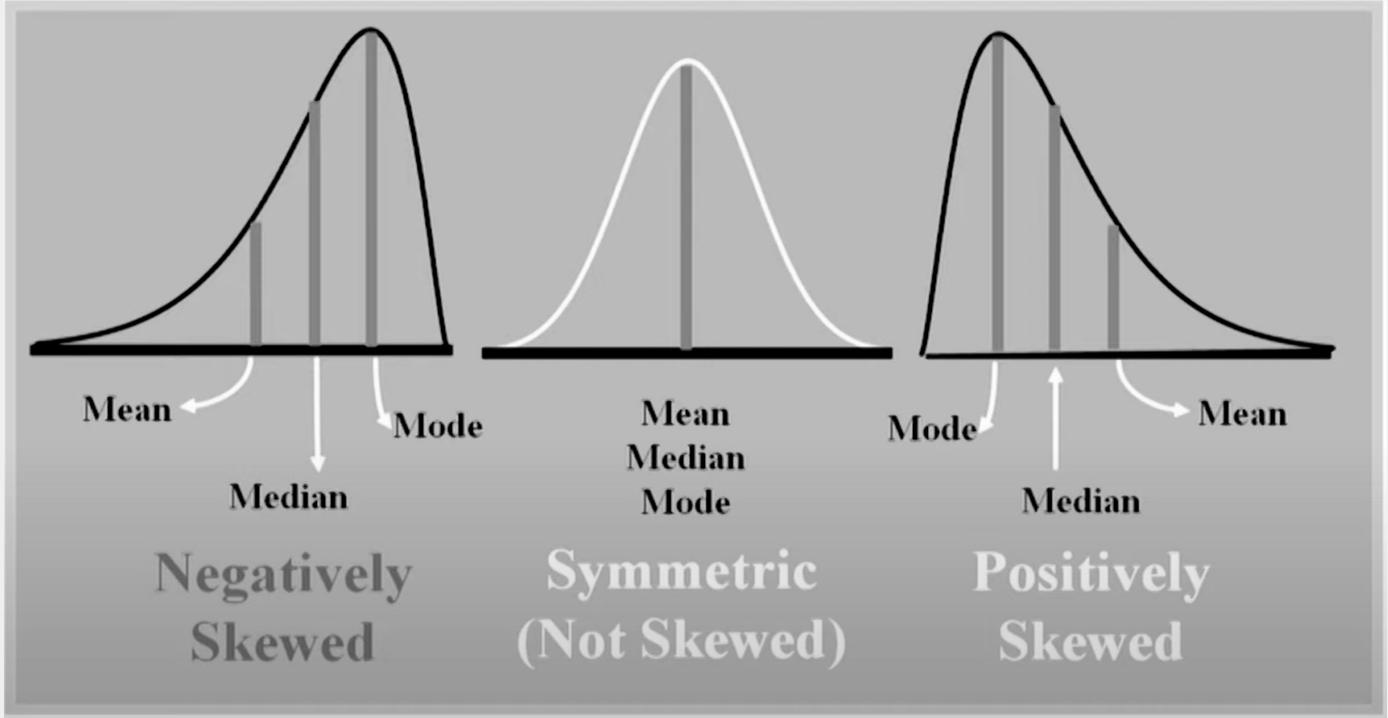
SKEWNESS


Negatively
Skewed
Symmetric
(Not Skewed)
Positively
Skewed

SKEWNESS..

- The skewness of a distribution is measured by comparing the relative positions of the mean, median, and mode.
- Distribution is *symmetrical:*
    - ***Mean = Median = Mode***
- Distribution is *skewed right:*
    - *Median lies between mode and mean, and mode is less than mean*
- Distribution is *skewed left:*
    - ***Median lies between mode and mean, and mode is greater than mean***

SKEWNESS…


Mean
Median
Mode
Negatively
Skewed
Mean
Median
Mode
Symmetric
(Not Skewed)
Mode
Median
Mean
Positively
Skewed

Coefficient of Skewness..

- Summary measure of the skewness.

$$\boxed{S = \frac{3\left(\mu - M_d\right)}{\sigma}}$$

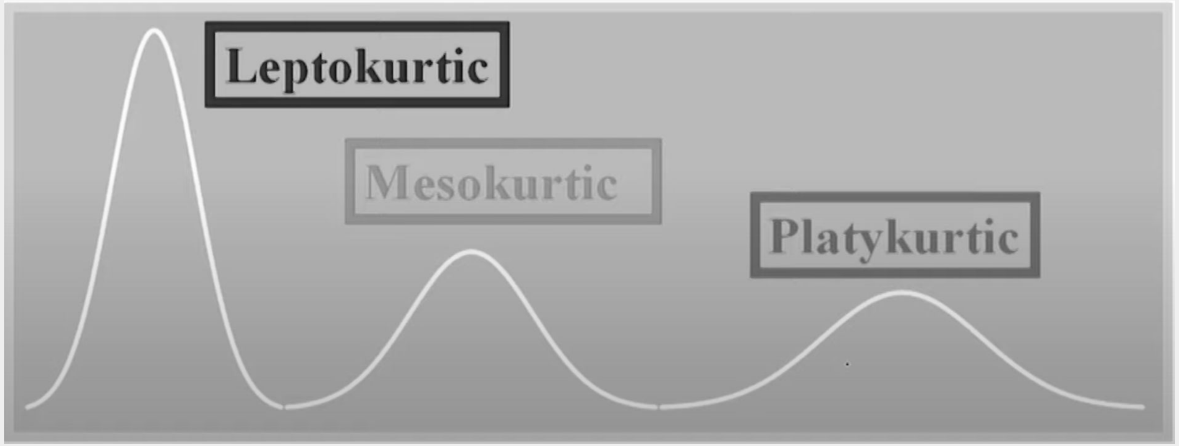
- If S < 0, the distribution is <u>negatively skewed</u> (Skewed to the left)
- If S = 0, the distribution is <u>symmetric</u> (Not skewed)
- If S > 0, the distribution is <u>positively skewed</u> (Skewed to the right)

Coefficient of skewness

$\mu_1 = 23$	$\mu_2 = 26$	$\mu_3 = 29$
$M_{d1} = 26$	$M_{d2} = 26$	$M_{d3} = 26$
$\sigma_1 = 12.3$	$\sigma_2 = 12.3$	$\sigma_3 = 12.3$
$S_1 = \frac{3(\mu_1 - M_{d1})}{\sigma_1}$	$S_2 = \frac{3(\mu_2 - M_{d2})}{\sigma_2}$	$S_3 = \frac{3(\mu_3 - M_{d3})}{\sigma_3}$
$= \frac{3(23-26)}{12.3}$	$= \frac{3(26-26)}{12.3}$	$= \frac{3(29-26)}{12.3}$
$= -0.73$	$= 0$	$= +0.73$

KURTOSIS

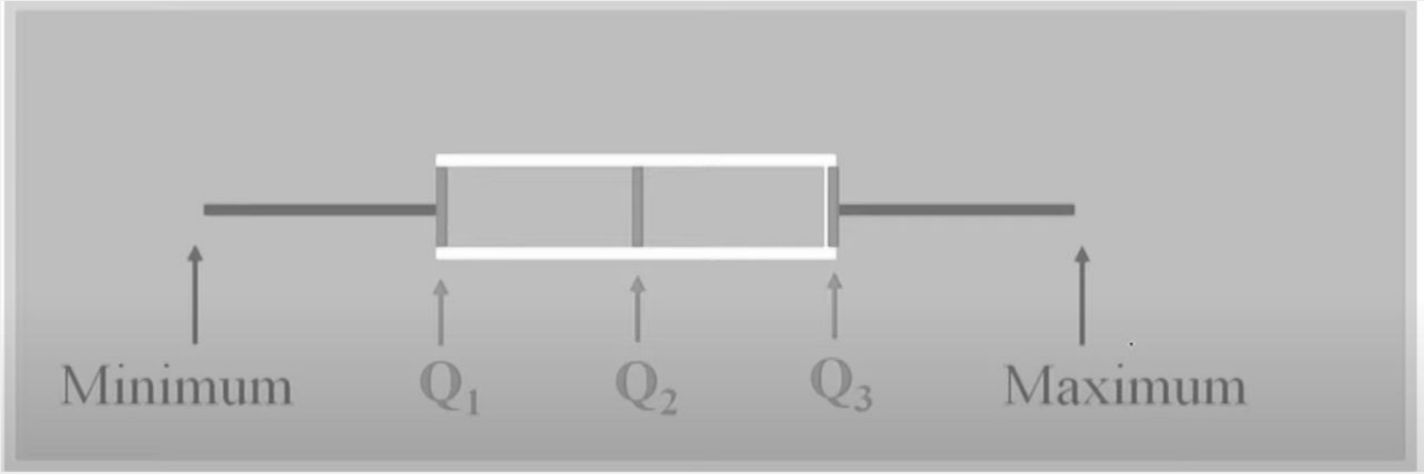
- Peakedness of the distribution
  - Leptokurtic: high and thin
  - Mesokurtic: normal in shape
  - Platykurtic: flat and spread out.



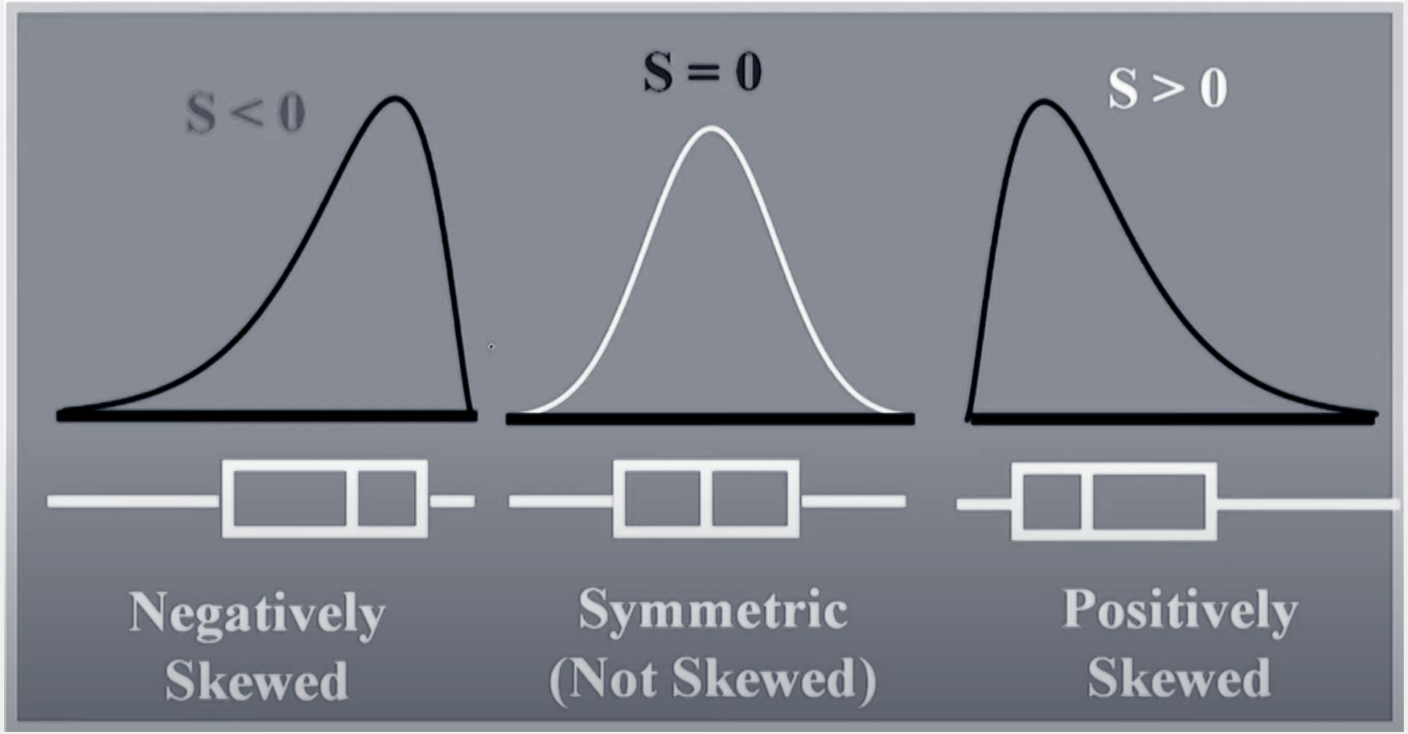
Box and Whisker Plot

- Five specific values are used:
  - Median, Q_2
  - First Quartile, Q_1
  - Third Quartile, Q_3
  - Minimum value in the data set.
  - Maximum value in the data set.

Box and Whisker plot



Skewness: Box and whisker plots, and coefficient of skewness



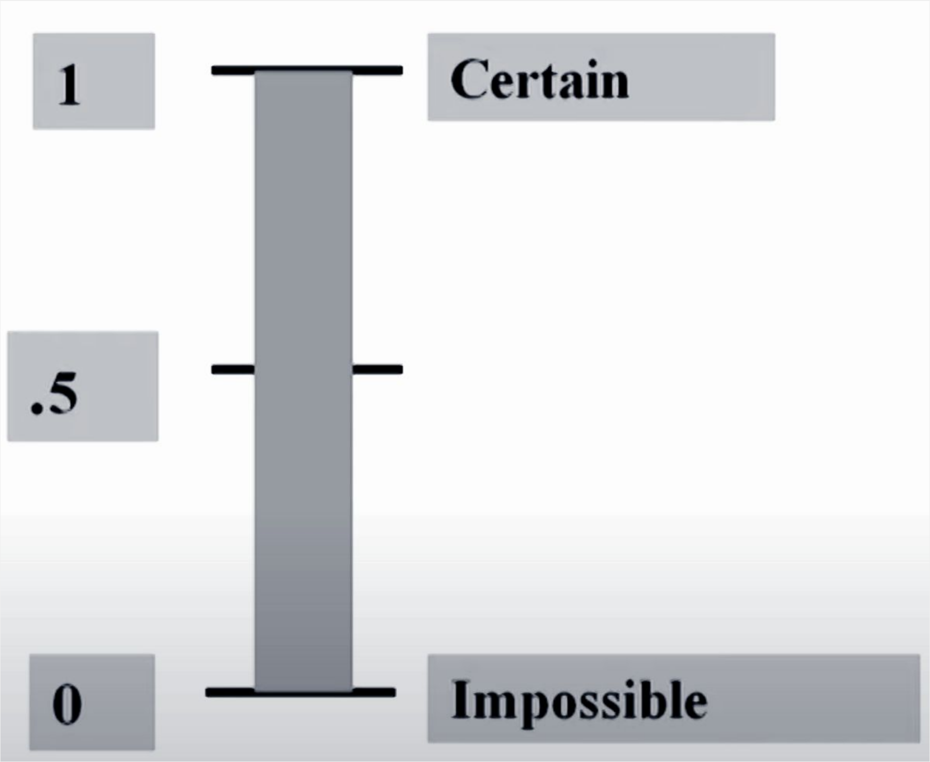
Introduction to Probability – Learning Objectives

- Comprehend the different ways of assigning probability.
- Understand and apply marginal, union, joint and conditional probabilities.
- Solve problems using the laws of probability including the laws of addition, multiplication and conditional probability.
- Revise probabilities using Baye’s Rule.

Probability

- Probability is the numerical measure of the likelihood that an event will occur.
- The probability of an event must be in between 0 and 1, inclusively
    - $0 \leq P(A) \leq 1$ for any event A.
- The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1
    - $P(A) + P(B) + P(C) = 1$
    - A, B, and C are mutually exclusive and collectively exhaustive.

Range of Probability



Methods of assigning Probabilities

- Classical method of assigning probability (rules and laws)
- Relative frequency of occurrence (Cumulated historical data)
- Subjective probability (Personal intuition or reasoning)

Classical Probability

- Number of outcomes leading to the event divided by the total number of outcomes possible.
- Each outcome is equally likely.
- Determined a priori – before performing the experiments.
- Applicable to game of chance.
- Objective – everyone correctly using the method assigns an identical probability.

Classical Probability

$$P(E) = \frac{n_e}{N}$$

Where:

N = total number of outcomes

n_e = number of outcomes in E

Relative frequency Probability

- Based on historical data.
- Computed after the experiment is performed.
- Number of times an event occurred divided by the number of trials
- Objective – everyone correctly using the method assigns an identical probability

Relative frequency Probability

$$P(E) = \frac{n_e}{N}$$

Where :

N = total number of trials

n_e = number of outcomes producing E

Subjective Probability

- Comes from a person intuition or reasoning.
- Subjective – different individual may (correctly) assign different numeric probabilities to the same event.
- Degree of belief
- Useful for unique (single-trial) experiments.
    - New product Introduction.
    - Initial public offering of common stock
    - Site selection decisions.
    - Sporting events.

Probability Terminology

- Experiments.
- Events
- Elementary events
- Sample space
- Unions and intersections.
- Mutually exclusive events
- Independent events
- Collectively exhaustive events
- Complementary events.