

# DATA SCIENCE FOR ENGINEERS

Week 3

Session Co-Ordinator : Abhijit Bhakte



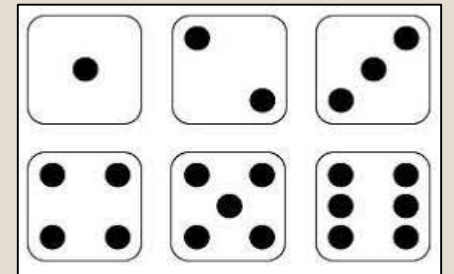
# Random Phenomenon

➤ The phenomenon/experiment whose outcomes are not predictable with certainty are called **random phenomenon**.



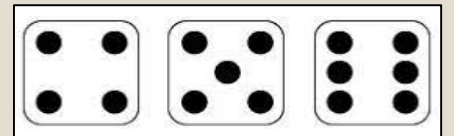
➤ **Sample space:** The set of all possible outcome of an experiment.

Ex  $\rightarrow S = \{1, 2, 3, 4, 5, 6\}$



➤ **Event:** Any subset of sample space

Ex  $\rightarrow$  outcome is greater than 3.  $E = \{4, 5, 6\}$



# Probability

- Probability is a measure that assign real value to every outcome of a random phenomenon
- The probability is ratio of number of ways an event can happen to the number of ways sample space can happen

$$P(E) = \frac{n(E)}{n(S)}$$

- Axioms of probability

- $0 \leq P(E) \leq 1$  (Probability is non-negative and less than one)
- $P(S) = 1$  (Probability of entire sample space is 1)
- $P(A \cup B) = P(A) + P(B)$  (For two mutually exclusive events)

**Q)** What is the probability of rolling an even number on a fair six-sided die?

- A) 1/6
- B) 1/3
- C) 1/2**
- D) 2/3

**Q)** If two fair coins are flipped, what is the probability of getting exactly one head?

- A) 1/4
- B) 1/2**
- C) 3/4
- D) 1/3

**Explanation:**

Even numbers ( $E$ ) =  $\{2, 4, 6\}$  ;  $n(E) = 3$   
Sample space( $S$ ) =  $\{1, 2, 3, 4, 5, 6\}$  ;  $n(S) = 6$

Therefore Prob of even no is  $= \frac{n(E)}{n(S)} = \frac{3}{6} = \frac{1}{2}$

**Explanation:**

Getting exactly one head ( $E$ ) =  $\{HT, TH\}$  ;  
 $n(E) = 2$

Sample space( $S$ ) =  $\{HH, HT, TH, TT\}$  ;  
 $n(S) = 4$

Therefore  $P(E) = \frac{n(E)}{n(S)} = \frac{2}{4} = \frac{1}{2}$

**Q)** A deck of playing cards contains 52 cards. What is the probability of drawing a red card (heart or diamond)?

- a) 1/2
- b) 39/52
- c) 1/4
- d) 1/3

**Q)** A bag contains 5 red balls and 3 green balls. What is the probability of drawing a red ball and then a green ball (without replacement)?

- a) 5/24
- b) 15/56
- c) 5/14
- d) 5/16

**Explanation:**

Total no of red cards in deck:  $n(R) = 13D + 13H = 26$

Total cards in deck:  $n(S) = 52$

Therefore Prob of even no is  $= \frac{n(R)}{n(S)} = \frac{26}{52} = \frac{1}{2}$

**Explanation:**

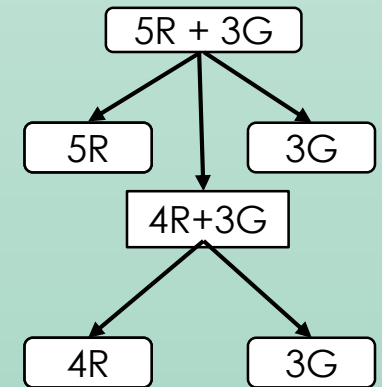
Prob of drawing red ball is:

$$P(R) = \frac{5}{8}$$

Prob of drawing green ball is:

$$P(R) = \frac{3}{7}$$

Combine prob =  $P(R) * P(R) = \frac{5}{8} * \frac{3}{7} = \frac{15}{56}$



**Q)** If out of all possible jumbles of the 'BIRD', a random word is picked, what is the probability, that this will start with a 'B'.

A) 1/3

**B) 1/4**

C) 3/4

D) 2/3

**BIRD**

↓

BIDR

BRID

BRDI

BDRI

BDIR

.

.

### Explanation

- $n(S) = \text{all possible jumbles of BIRD} = 4! = 4 \times 3 \times 2 \times 1$
- $n(E) = \text{jumbles starting with 'B'} = 3! = 3 \times 2 \times 1$

$$P(E) = \frac{n(E)}{n(S)} = \frac{3!}{4!} = \frac{1}{4} = \mathbf{0.25}$$

# Events

- **Mutually Exclusive Event:** The occurrence of one event implies that other event does not occur

Ex→ Coin toss gives either Head or Tail not both

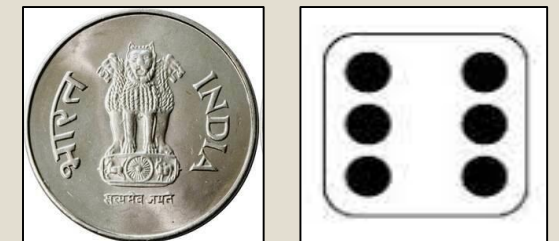
- $P(A \cap B) = 0$
- $P(A \cup B) = P(A) + P(B) + P(A \cap B) = P(A) + P(B)$



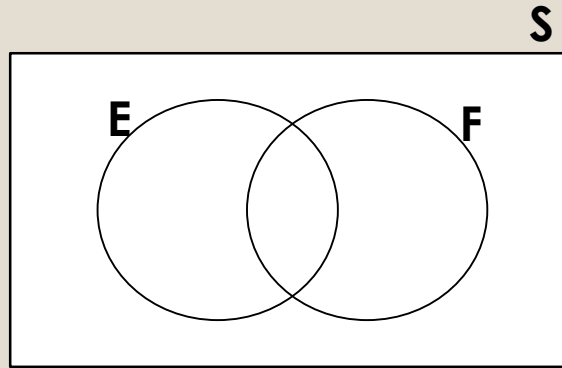
- **Independent Event:** two events are independent if occurrence of one has no influence on other

Ex→ Landing on heads after tossing a coin AND rolling a 5 on a single 6-sided die

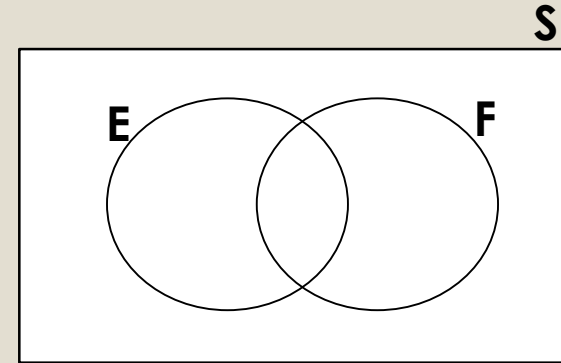
- $P(A \cap B) = P(A) \cdot P(B)$
- $P(A \cup B) = P(A) + P(B) + P(A \cap B) = P(A) + P(B) + P(A) \cdot P(B)$



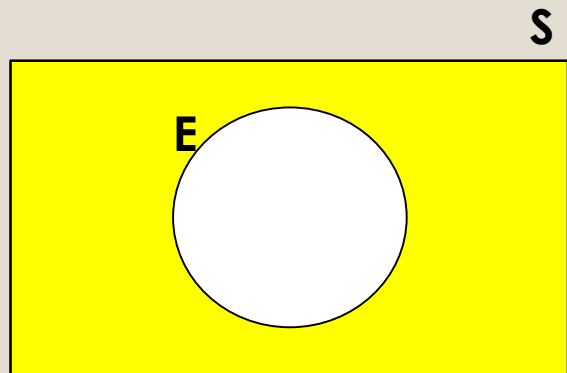
# Venn Diagram



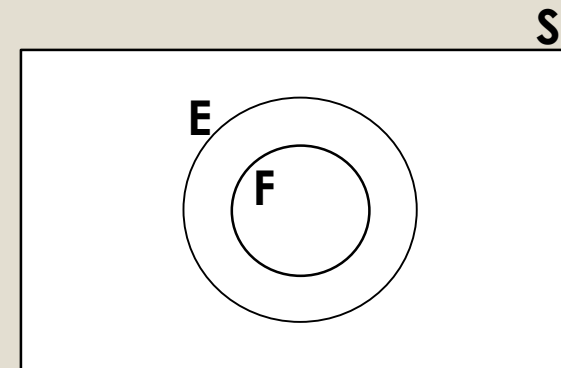
Colour region:  $E \cup F$



Colour region:  $E \cap F$



Colour region:  $E^c$



$E \subseteq F$



**Q)** If two events A and B are mutually exclusive, what can be said about their intersection?

- A)** The intersection of A and B is the empty set.
- B) The intersection of A and B contains only one element.
- C) The intersection of A and B is equal to A.
- D) The intersection of A and B is equal to B.

**Explanation:**

Mutually exclusive events have no common outcomes (coin toss and dice rolling), so the intersection of two mutually exclusive events is the empty set ( $\emptyset$ )

**Q)** If events A and B are independent, which of the following is true about their joint probability?

- A)**  $P(A \text{ and } B) = P(A) \times P(B)$
- B)  $P(A \text{ and } B) = P(A) + P(B)$
- C)  $P(A \text{ and } B) = P(A) - P(B)$
- D)  $P(A \text{ and } B) = P(A) / P(B)$

**Explanation:**

For independent events A and B, the joint probability of both events occurring is given by the product of their individual probabilities:  $P(A \text{ and } B) = P(A) \times P(B)$ .

**Q)** Are mutually exclusive events always dependent?

- a) Yes
- b) No

**Explanation:**

Mutually exclusive events cannot happen together. If one event occurs, it eliminates the possibility of the other event occurring.

**Q)** If events A and B are independent, and  $P(A) = 0.4$  and  $P(B) = 0.6$ , what is  $P(A \cup B)$ ?

- a) 0.2
- b) 0.6
- c) 0.8**
- d) 0.98

**Explanation:**

For independent events, the probability of either event happening is given by

$$P(A \cup B) = P(A) + P(B) - P(A) * P(B)$$

$$P(A \cup B) = 0.4 + 0.6 - (0.4 * 0.6) = 0.8.$$

# Conditional Probability

- If two event A & B are not independent, then information available about the outcome of event A can influence the predictability of event B

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Example → Two fair coins are toss

**Event A:** First toss is Head = {HT, HH}

$$P(A) = n(A)/n(S) = 2/4 = 0.5$$

**Event B:** Two successive head = {HH}

$$P(B) = n(B)/n(S) = 1/4 = 0.25$$

If the event A is given then probability of event B is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.25}{0.5} = \mathbf{0.5}$$

**Q)** If events A and B are independent, which of the following is true about their conditional probabilities?

- A)**  $P(A | B) = P(A)$
- B)  $P(A | B) = P(B)$
- C)  $P(A | B) = P(A) + P(B)$
- D)  $P(A | B) = P(A) \times P(B)$

**Q)** In a group of people, 40% like ice cream, 30% like chocolate, and 20% like both ice cream and chocolate. What is the probability that a randomly selected person likes ice cream given that they like chocolate?

- A)  $1/2$
- B)**  $2/3$
- C)  $4/5$
- D)  $1/3$

**Explanation:**

For independent events A and B, the occurrence of event B does not affect the probability of event A.

With formula: 
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

**Explanation:**

$$P(I) = 0.4$$

$$P(C) = 0.3$$

$$P(C \cap I) = 0.2$$

Randomly selected person likes ice cream given that they like chocolate

$$P(I|C) = \frac{P(C \cap I)}{P(C)} = \frac{0.2}{0.3} = \frac{2}{3}$$

# Random Variable

- A random variable (RV) is a map from sample space to a real line such that there is a unique real number corresponding to every outcome of sample space.
- Ex → coin toss sample space [H T] map to [0 1]

## ➤ **Discrete random variable:**

A variable that take one value from a discrete set of values

Ex → dice rolling [ 1 2 3 4 5 6 ]

## ➤ **Continuous random variable:**

The variable that can take continuous range of values

Ex → Temperature of the week [ 28.3, 21.0, 25.9, 26.1, 32.6, 30.0, 29.8 ]

**Q)** Which of the following is an example of a discrete random variable?

- a) Height of individuals in a population.
- b) Temperature in degrees Celsius.
- c)** Number of heads in three coin tosses.
- d) Time taken to complete a marathon.

**Explanation:**

A discrete random variable takes on distinct, separate values with gaps in between, such as the number of heads in a coin toss, which can only be 0, 1, 2, or 3.

**Q)** The probability distribution of a discrete random variable must satisfy which of the following?

- a) The probabilities must be negative.
- b)** The sum of the probabilities must be exactly 1.
- c) The probabilities must be greater than 1.
- d) The probabilities must be integers.

**Explanation:**

The probabilities assigned to each possible value of a discrete random variable must add up to 1, representing the entire probability space.

This is one of the axiom in probability

# Probability mass/density function

- Probability mass/density function help to assign the probability to every outcome of a sample space

## Probability Mass Function (PMF)

- PMF use for **discrete** random variable
- $P(x) = P[X = x]$
- Ex → In coin toss

$$P[X = 0] = 0.5, P[X = 1] = 0.5$$

## Probability Density Function (PDF)

- PDF use for **continuous** random variable
- $P(a < x < b) = \int_a^b f(x) dx$
- Ex → height between than 20 and 40 C

$$p(20 < T < 40) = \int_{20}^{40} f(x) dx$$

# Probability mass/density function

- Probability mass/density function help to assign the probability to every outcome of a sample space

## Probability Mass Function (PMF)

- PMF use for discrete random variable
- $P(x) = P[X = x]$
- Ex → In coin toss

$$P[X = 0] = 0.5, P[X = 1] = 0.5$$

## Probability Density Function (PDF)

- PDF use for continuous random variable
- $P(a < x < b) = \int_a^b f(x) dx$
- Ex → height between than 20 and 40 C

$$p(20 < T < 40) = \int_{20}^{40} f(x) dx$$



**Que:** The box contain 20 defective items and 80 non-defective items. If two items are selected at random without replacement , what will be the probability that both items are defective ?

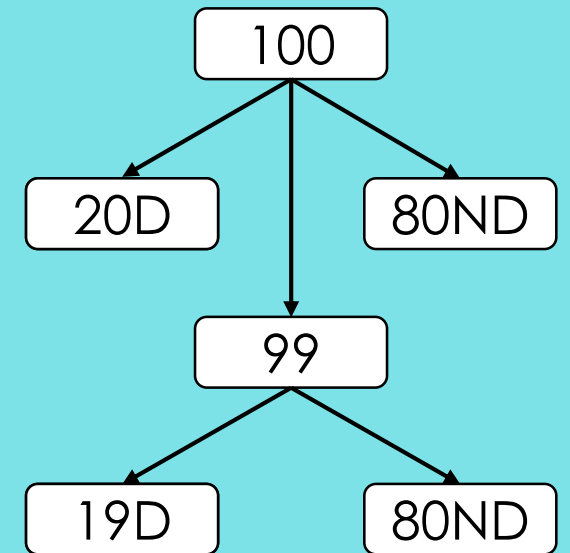
- A) 26/495
- B) 36/495
- C) 23/495
- D) 19/495**

### Explanation

- $P(\text{both items are defective}) = ?$
- Probability of selecting first defective item:

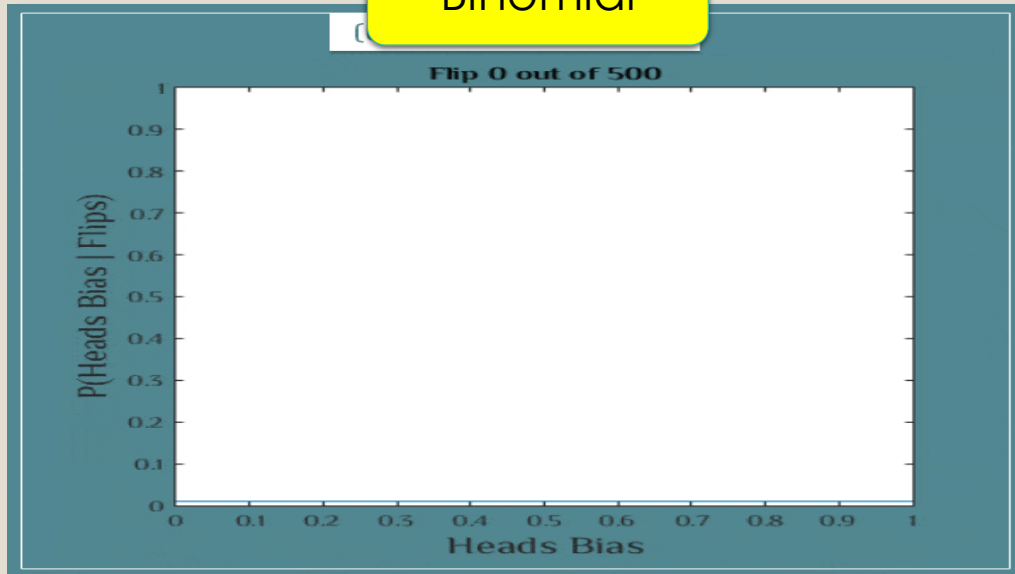
$$P(1D) = \frac{20}{100}$$

- Now we left with total 99 items with 19 defective
- Therefore, probability of selecting second defective item:  $P(2D) = \frac{19}{99}$
- Combine prob=  $P(1D) * P(2D) = \frac{20}{100} * \frac{19}{99} = \frac{19}{495}$



# Distributions

Binomial



$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

where

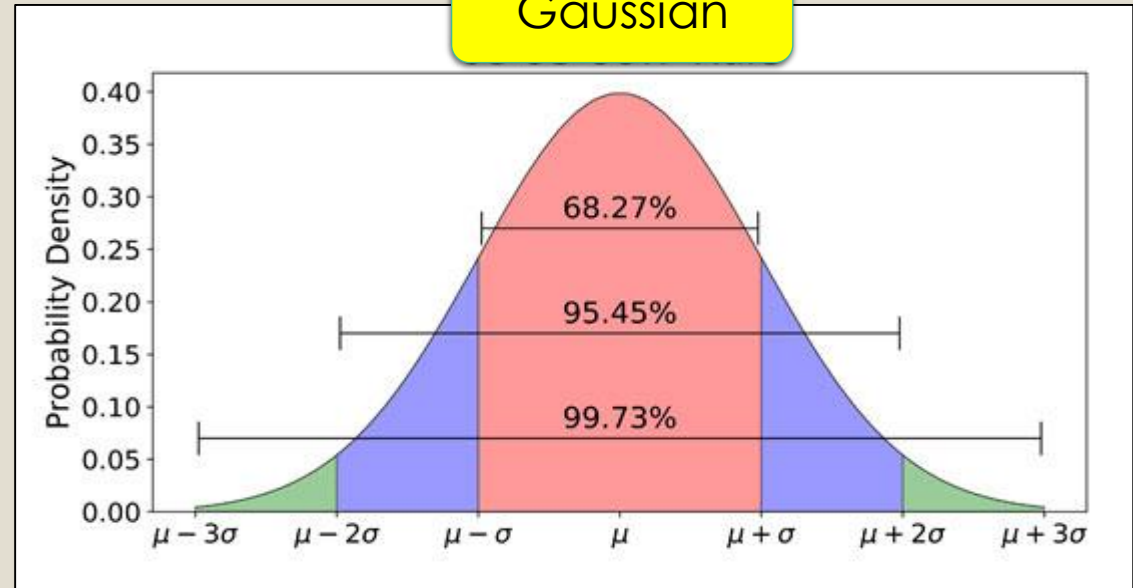
$n$  = the number of trials (or the number being sampled)

$x$  = the number of successes desired

$p$  = probability of getting a success in one trial

$q = 1 - p$  = the probability of getting a failure in one trial

Gaussian

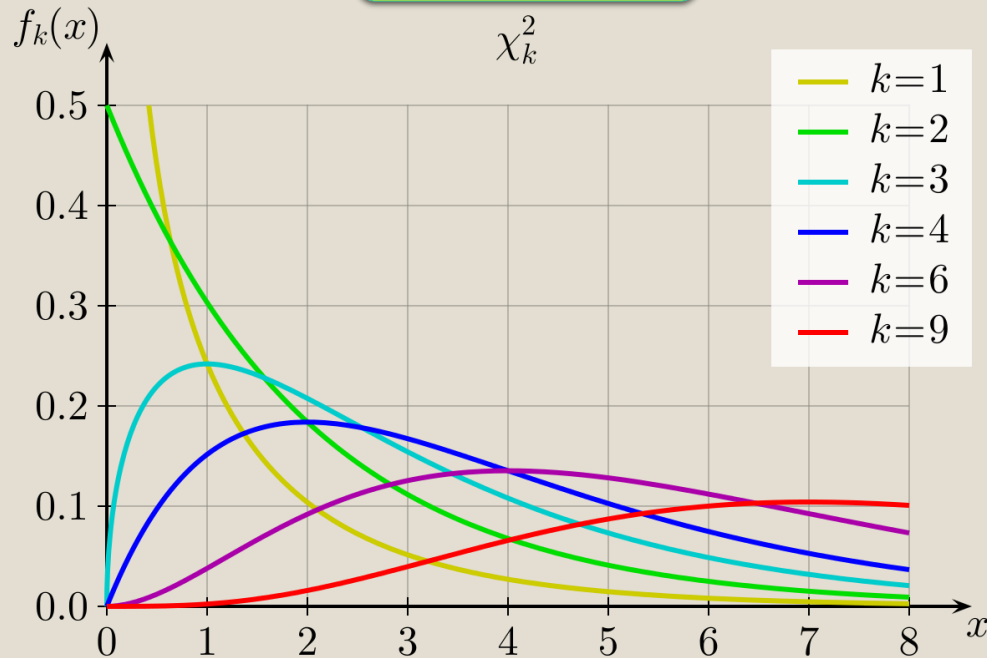


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where  $\sigma$  is the standard deviation and  $\mu$  the mean

# Distributions

## Chi-square



$$f(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right) 2^{k/2}} x^{\frac{k}{2}-1} \cdot e^{-x/2}$$

- The function is characterize by one parameter i.e., degree of freedom ( $k$ )
- The range of distribution is 0 to Inf
- Mean= $k$ ; variance=  $2k$
- This help in hypothesis testing (discuss next)

**Q)** What type of events does the binomial distribution model?

- A) Continuous events
- B) Discrete events**
- C) Events with a normal distribution
- D) Events with a uniform distribution

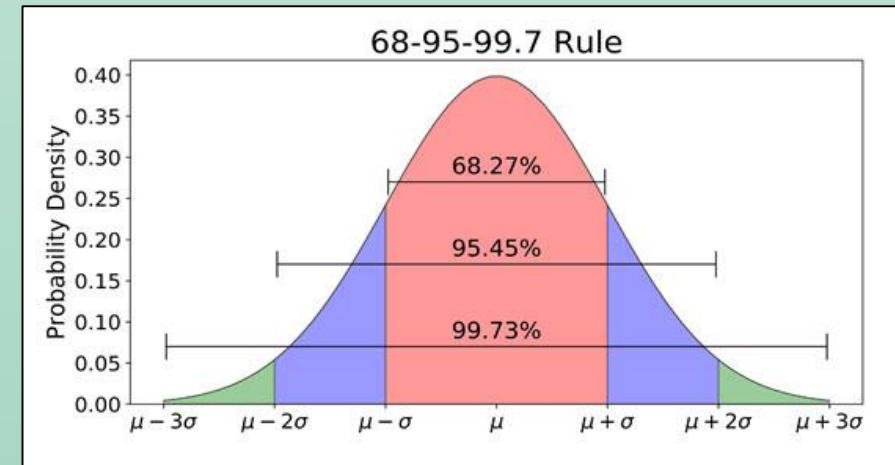
**Explanation:**

The binomial distribution models the number of successes in a fixed number of independent trials, where each trial has two possible outcomes (success or failure). Therefore it is used for discrete events.

**Q)** What is the shape of the Gaussian distribution?

- A) U-shaped
- B) Skewed to the left
- C) Bell-shaped**
- D) Skewed to the right

**Explanation:**



**Q)** What are the parameters that define a chi square distribution?

- A) Mean and standard deviation
- B) Probability of success and number of trials
- C) Degree of freedom**
- D) Mode and median

**Q)** What percentage of data falls within one standard deviation of the mean in a normal distribution?

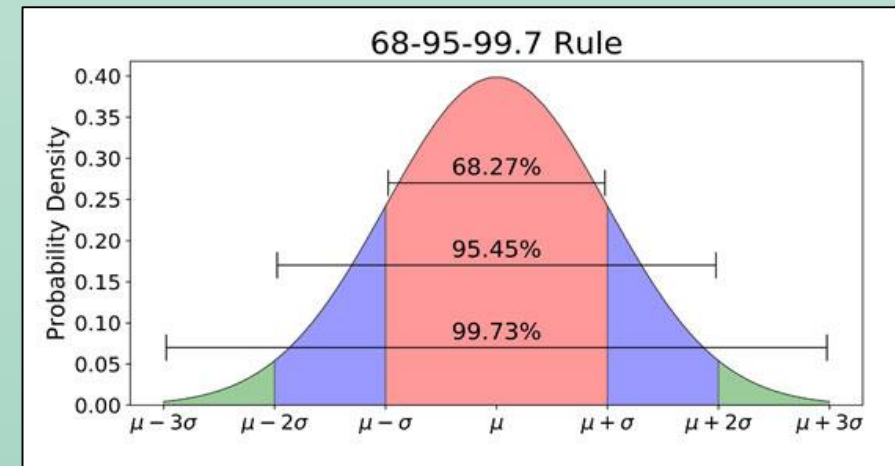
- A) 25%
- B) 50%
- C) 68%**
- D) 95%

### Explanation:

As shown in density function ( $k=\text{dof}$ ) is the only parameter use to generate distribution

$$f(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right) 2^{k/2}} x^{\frac{k}{2}-1} \cdot e^{-x/2}$$

### Explanation:



# Expected Value

- It's a way to understand what might happen on average if you repeat something many times.
- Ex → if coin tossed, probability of head
- For discrete distribution:

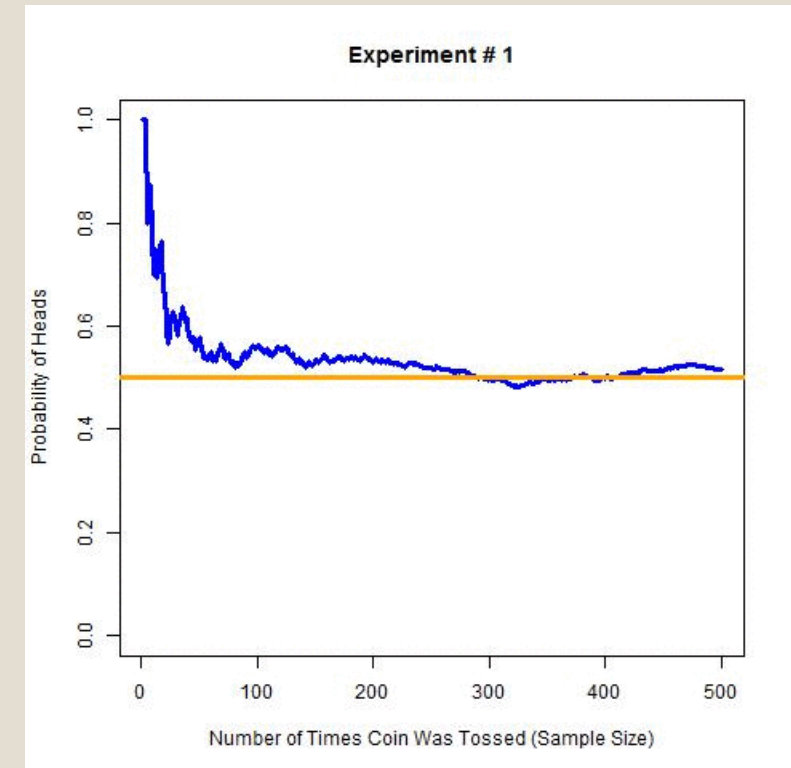
$$E[X] = \sum_{i=1}^n x_i \cdot p(x_i)$$

- For continuous distribution:

$$E[X] = \int_{-\infty}^{\infty} x_i \cdot f(x_i) dx$$

- Mean:  $\mu = E[X]$

- Variance:  $\sigma^2 = E[(x - \mu)^2] = E[X^2] - \mu^2$



**Q)** If you roll a fair six-sided die, what is the expected value of a single roll?

- A)** 3.5
- B) 6
- C) 4.5
- D) 2.5

**Q)** In a normal (Gaussian) distribution, where is the expected value located?

- A) At the mode of the distribution
- B) At the median of the distribution
- C)** At the mean of the distribution

**Explanation:**

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

Using formulae  $E[X] = \sum_{i=1}^n x_i \cdot p(x_i)$

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2} = 3.5$$

**Explanation:**

Expected value generally represents the mean of the distribution

# Properties of expectation and variance

## ➤ Expectation

- $E(ax_1 + b) = a E(x_1) + b$
- $E(ax_1 + bx_2) = a E(x_1) + b E(x_2)$

## ➤ Variance

- $V(ax_1 + b) = a^2 V(x_1)$
- $V(ax_1 + bx_2) = a^2 V(x_1) + b^2 V(x_2) + 2ab \operatorname{cov}(x_1, x_2)$



**Q)** If  $\text{Var}(X) = 9$  and  $\text{Var}(Y) = 16$ , what is the variance of  $2X - 3Y$ , if  $X$  and  $Y$  are independent?

- A) 120
- B) 150
- C) 180**
- D) 100

**Q)** Let  $X$  and  $Y$  be two independent random variables with expectations  $E(X) = 5$  and  $E(Y) = 3$ . What is the expectation of the random variable  $Z = 2X - 3Y$ ?

- A) 4
- B) 1**
- C) -4
- D) -6

**Explanation:**

- Using formulae

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab \text{cov}(X, Y)$$

$$V(2X - 3Y) = 2^2 V(X) + (-3)^2 V(Y) + 0$$

$$V(2X - 3Y) = 4 * 9 + 9 * 16 + 0 = 180$$

**Explanation:**

- Using formulae

$$E(aX + bY) = a E(X) + b E(Y)$$

$$E(2X - 3Y) = 2 E(X) - 3 E(Y)$$

$$E(2X - 3Y) = 2 * 5 - 3 * 3 = 1$$

# Covariance & Correlation

## Covariance

- Covariance indicates the **direction** of the **linear relationship** between variables
- Covariance values are **not standardized**.
- Value can be anything

$$\text{cov}(X, Y) = \sum_{i=1}^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}.$$

## Correlation

- Correlation measures **both** the **strength** and **direction** of the **linear relationship** between two variables
- Correlation values are **standardized**
- Value lie between -1 and +1

$$\text{Correlation} = \frac{\text{Cov}(x, y)}{\sigma_x * \sigma_y}$$

# Properties of joint pdf

- Joint pdf of two random variables  $x$  and  $y$ :  $f(x,y)$

$$P(x \leq a, y \leq b) = \int_{-\infty}^b \int_{-\infty}^a f(x, y) dx dy$$

- **Covariance between  $x$  and  $y$**

$$\sigma_{x,y} = E[(x - \mu_x)(y - \mu_y)]$$

- **Correlation between  $x$  and  $y$ :**

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$$

**Q)** If two random variables are independent, what can we say about their covariance and correlation?

- a) Covariance is zero, and correlation is zero.
- b) Covariance is zero, and correlation can be any value.
- c) Covariance can be any value, and correlation is zero.
- d) Covariance can be any value, and correlation is one.

**Explanation:**

- When two random variables are independent, they have no linear relationship, so both the covariance and correlation will be zero.

**Q)** The correlation coefficient between X and Y is 0.6. The covariance between them is 25. Then find the product between variance of both the variables?

- a) 39.33
- b) 41.66
- c) 40.11
- d) 36.99

**Explanation:**

- Using formulae

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$$

$$0.6 = \frac{25}{\sigma_x \sigma_y}$$

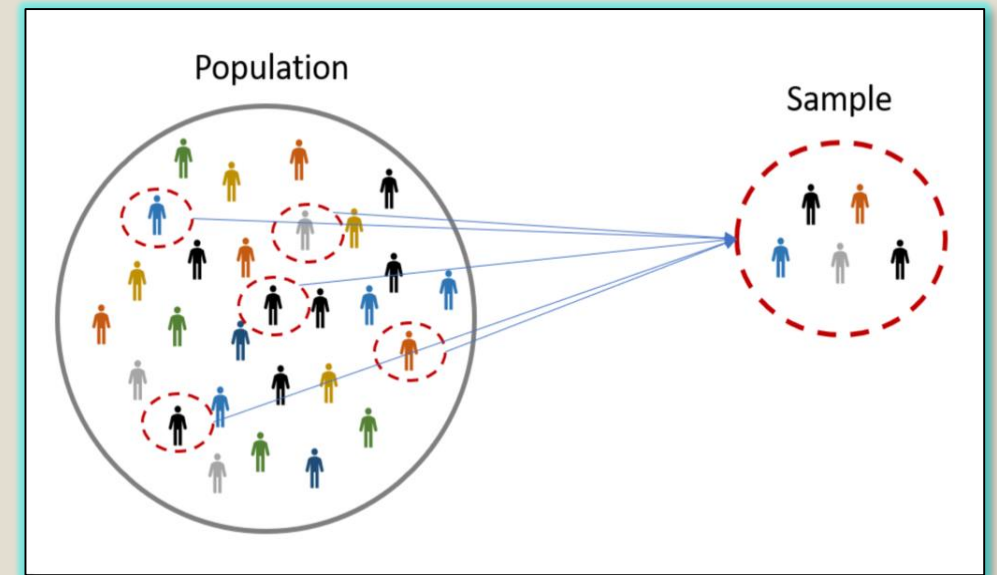
$$\sigma_x \sigma_y = \frac{25}{0.6} = 41.667$$

# R studio to calculate probability

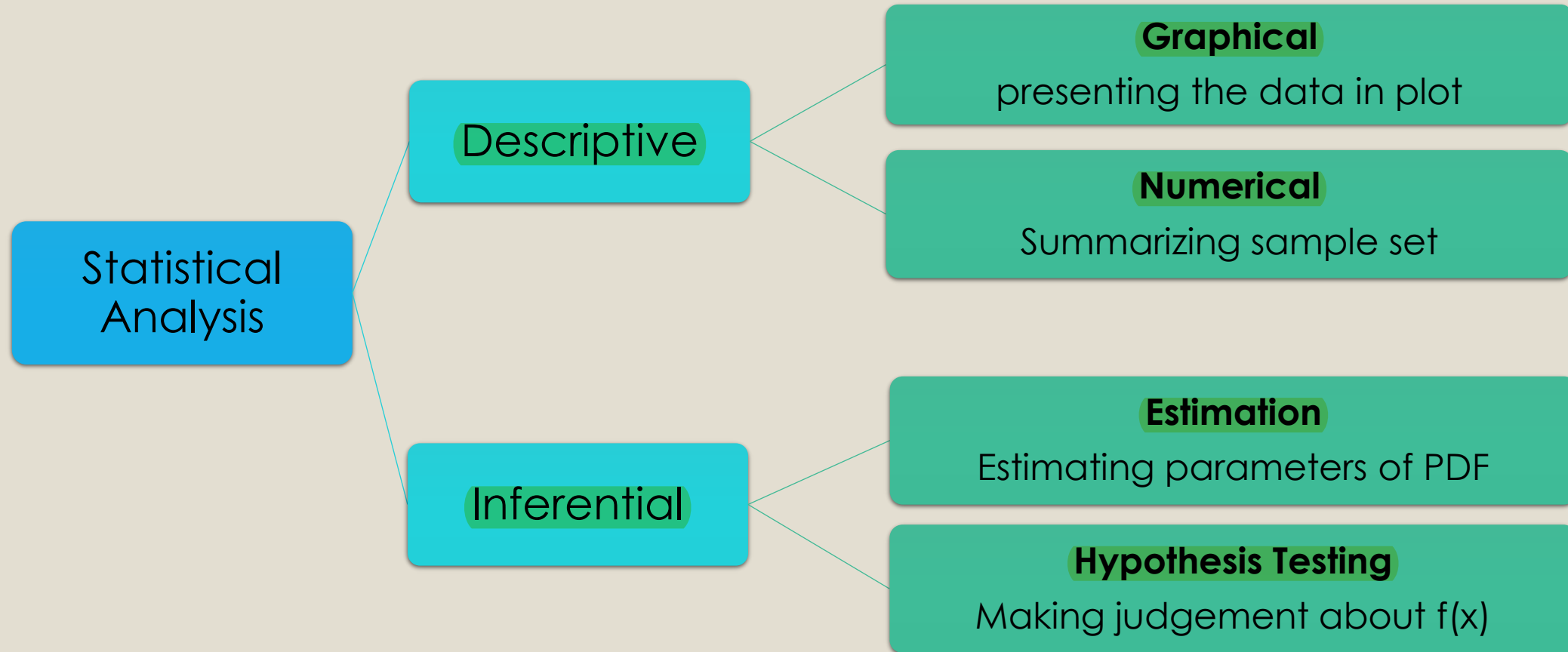
- **`pnorm(X, mean, std, 'lower tail'=T/F)`** : it gives the **probability** of the PDF of normal distribution for provided X random variable
- **`pchisq`**: chi-square distribution (parameter is DOF)
- **`pbinom`**: *binomial distribution (parameters are trials and success probability)*
- **`punif`**: *uniform distribution (min and max value)*
  
- **`qnorm(p, mean, std, 'lower tail'=T/F)`** : it gives the **random variable** of the PDF of normal distribution for provided probability p
- **`qchisq`**: chi-square distribution (parameter is DOF)
- **`qbinom`**: *binomial distribution (parameters are trials and success probability)*
- **`qunif`**: *uniform distribution (min and max value)*

# Statistical sampling

- **Population:** Set of all possible outcome of random experiment
- **Sample set:** Finite set of observation obtained from experiment
- Sampling **help to make inferences** about the population
- The inference may be uncertain because samples might be uncertain

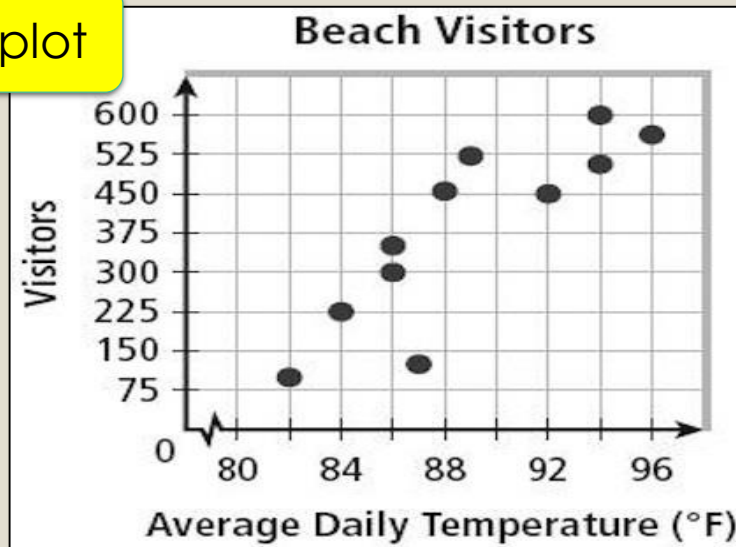


# Statistical Analysis

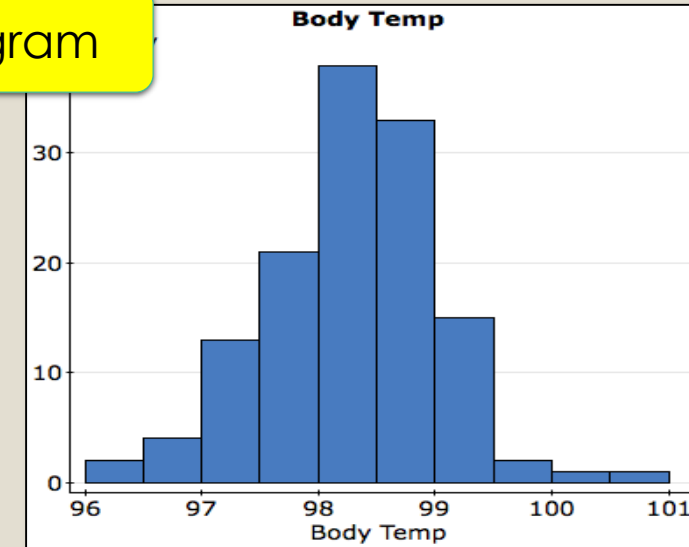


# Graphical statistics

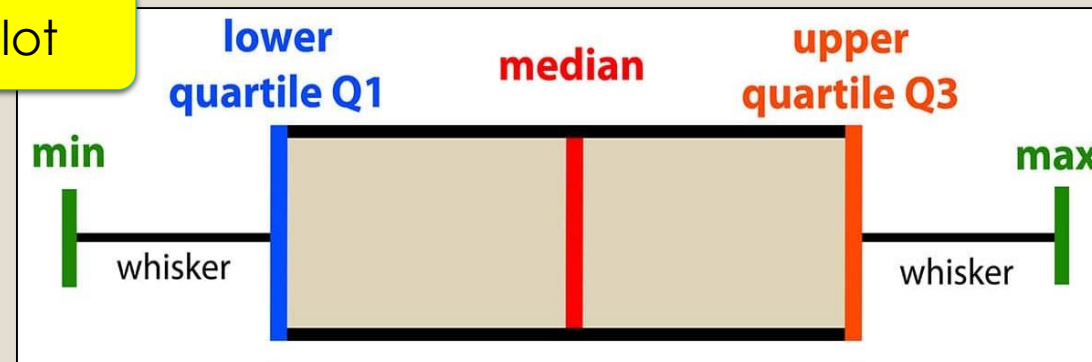
Scatter plot



Histogram



Box plot





# Numerical Statistics

**Mean:** Mean is **average** or norm  $\rightarrow \frac{1+3+4+6+6+7+8}{7} = 5$

**Median:** Median is **middle** value  $\rightarrow 1\ 3\ 4\ \mathbf{6}\ 6\ 7\ 8$

**Mode:** Mode is **most frequent** value  $\rightarrow 1\ 3\ 4\ \mathbf{6\ 6}\ 7\ 8$

**Range:** **Difference** between **lowest** and **highest** value  $\rightarrow 8 - 1 = 7$

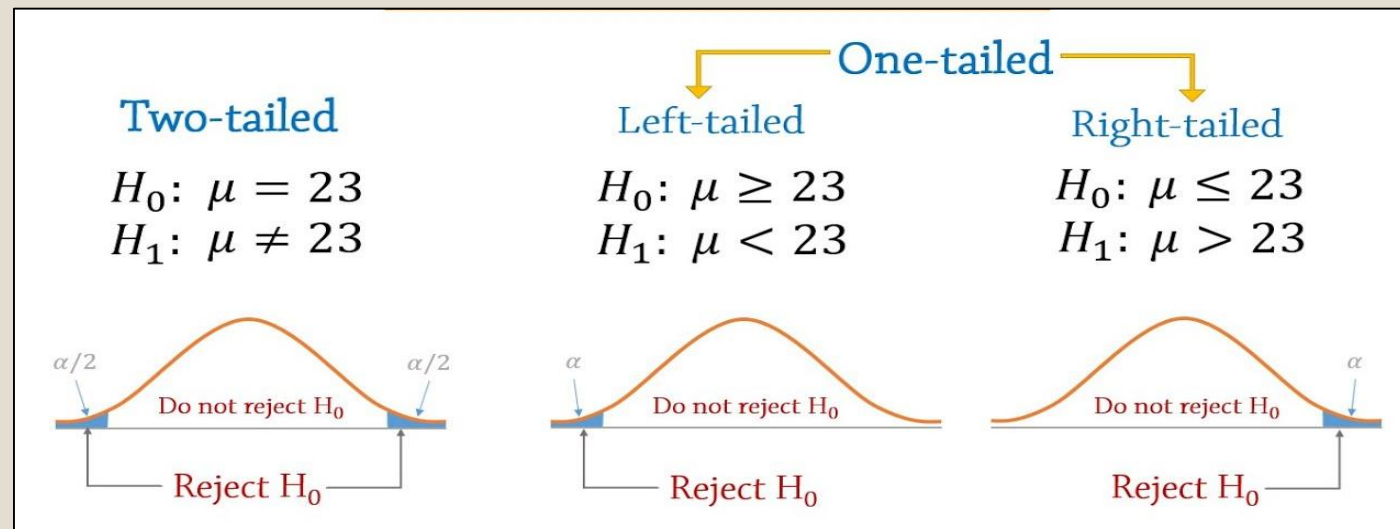
Goals scored in seven  
matched

1 3 4 6 6 7 8

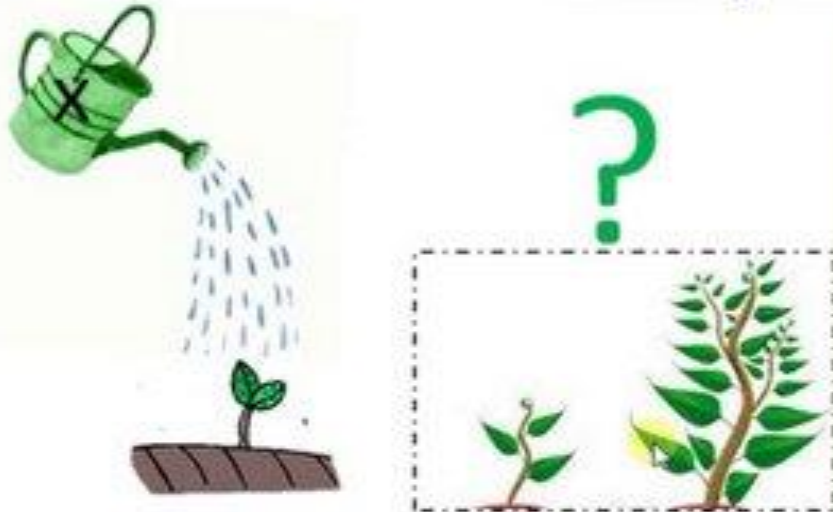
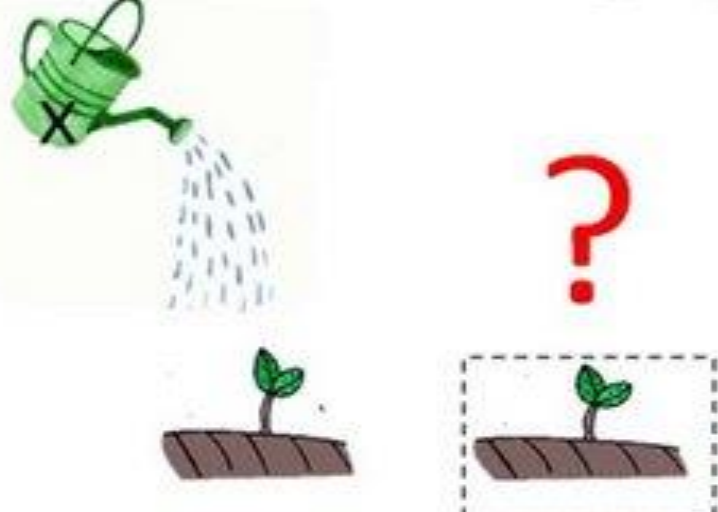


# Hypothesis testing

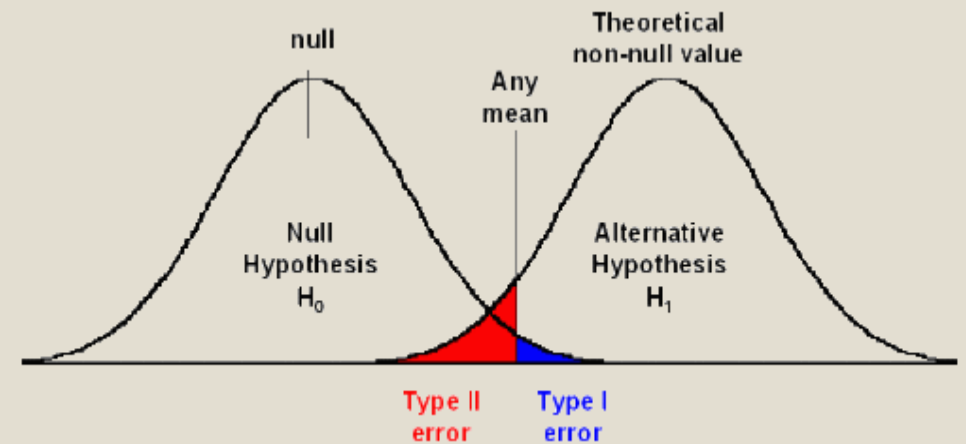
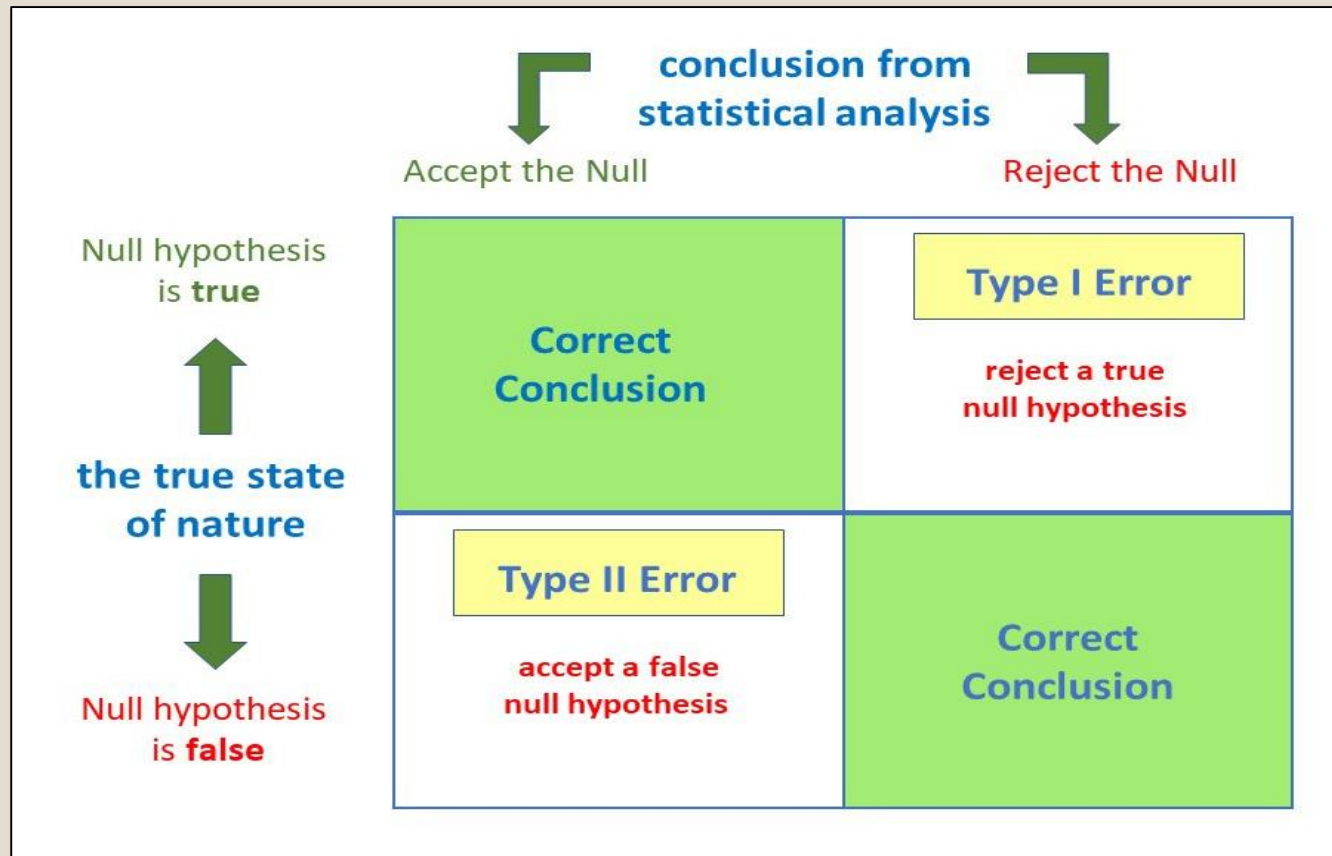
- Hypothesis testing is **used to make decisions**
- Ex: Whether company stock yield profit that desired value
- Ex: Whether the effect of drug A similar to drug B
- Hypothesis testing is generally **converted** to a **test of mean** and **variance** parameter of population



# Null & Alternative hypothesis

	
<p><math>H_1</math>: Application of bio-fertilizer 'x' increase plant growth.</p>	<p><math>H_0</math>: Application of bio-fertilizer 'x' <u>do not</u> increase plant growth.</p>
<p><b>Alternative hypothesis</b></p>	<p><b>Null hypothesis</b></p>
<p>✓ The alternative hypothesis is a hypothesis which the researcher tries to prove.</p>	<p>✓ The null hypothesis is a hypothesis which the researcher tries to disprove, or nullify.</p>

# Errors in hypothesis testing



# z-test vs t-test

## z-test

- Used when population variance is known
- Used for sample size greater than 30
- Based on normal distribution

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\sigma = \text{known std. dev.}$

## T-test

- Used when variance is not known
- Use for sample size less than 30
- Based on student-t distribution

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$s = \text{sample std. dev}$

**Que:** A teacher claims that the mean score of students in his class is greater than 82 with a standard deviation of 20. If a sample of 81 students was selected with a mean score of 90 then check if there is enough evidence to support this claim at a 0.05 significance level

- Number of students =  $n = 81$
- Sample mean =  $\bar{x} = 90$
- Population Std. deviation =  $\sigma = 20$

$$H_0: \mu = 82$$

$$H_1: \mu > 82$$

From z table critical value of  $\alpha$  is 1.645 (this is calculated using Rstudio with *pnorm* & *qnorm*)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = 3.6$$

As  $3.6 > 1.645$  thus, the **null hypothesis is rejected**

**Que:** A researcher wants to test whether the mean weight of a sample of 50 apples is significantly different from a claimed population mean weight of 150 grams. The sample mean weight is 149 grams, and the sample standard deviation is 10 grams. Calculate the t-statistic for this test.

- Number of apples =  $n = 50$
- $DOF = n - 1 = 49$
- Population mean =  $\mu = 150$
- Sample mean =  $\bar{x} = 149$
- Sample Std. deviation =  $s = 10$

$$H_0: \mu = 150$$

$$H_1: \mu \neq 150$$

From t table critical value of  $\alpha$  is  $\pm 1.676$  (95%)  
(this is calculated using Rstudio with `pt` & `qt`)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{149 - 150}{\frac{10}{\sqrt{50}}} = -0.7071$$

As  $-0.7071 > -1.676$  thus, the **null hypothesis is accepted**

