Statistics

Expectation and Variance of RVs

Discrete Random Variables:

- **Probability Mass Function (PMF):** Let X be a discrete random variable with possible values $x_1, x_2, x_3, \ldots$ and corresponding probabilities $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$, $p_3 = P(X = x_3)$, ... (where $\Sigma p_i = 1$).
- **Expectation (Mean):**

  $E[X] = \mu = \Sigma [x_i * p_i]$ (sum over all possible values of x)
- **Variance:**

  $Var(X) = E[(X - \mu)^2] = \Sigma [(x_i - \mu)^2 * p_i]$

  $Var(X) = E[X^2] - (E[X])^2$ (a computationally useful form)
- **Expectation of a Function of X:**

  $E[g(X)] = \Sigma [g(x_i) * p_i]$

Expectation and Variance of RVs

General Properties of Expectation:

- **Linearity:** $E[aX + bY] = aE[X] + bE[Y]$, where a and b are constants and X and Y are random variables.
- **Constant:** $E[c] = c$, where c is a constant.

General Properties of Variance:

- **Constant:** $Var(c) = 0$, where c is a constant.
- **Scaling:** $Var(aX) = a^2Var(X)$, where a is a constant.
- **Linear Transformation:** $Var(aX + b) = a^2Var(X)$, where a and b are constants.
- **Independence:** If X and Y are independent random variables, then $Var(X + Y) = Var(X) + Var(Y)$. (This does *not* generally hold if X and Y are dependent).

Expectation and Variance of RVs

Q1. Let Z be a random variable with the following probability distribution:
$P(Z = -1) = 0.2$ $P(Z = 0) = 0.5$ $P(Z = 1) = 0.3$
Define a new random variable $W = Z^2$.

- Find the expected value of W, E[W].
- Find the variance of W, Var(W).

First, we need to find the probability distribution of W.

- If $Z = -1$, then $W = (-1)^2 = 1$. $P(W = 1) = P(Z = -1) = 0.2$
- If $Z = 0$, then $W = (0)^2 = 0$. $P(W = 0) = P(Z = 0) = 0.5$
- If $Z = 1$, then $W = (1)^2 = 1$. $P(W = 1) = P(Z = 1) = 0.3$

Notice that W can only take the values 0 and 1. The probability distribution of W is:

- $P(W = 0) = 0.5$
- $P(W = 1) = 0.2 + 0.3 = 0.5$

Now we can calculate E[W]:

$E[W] = \Sigma [w * P(W = w)]$
$= (0 * 0.5) + (1 * 0.5)$
$= 0 + 0.5 = 0.5$

$Var(W) = E[W^2] - (E[W])^2$
Since W can only be 0 or 1, W^2 will also only be 0 or 1. In fact, $W^2 = W$ in this case. This is because $0^2=0$ and $1^2=1$. So, $E[W^2] = E[W] = 0.5$

$Var(W) = E[W^2] - (E[W])^2$
$= 0.5 - (0.5)^2$
$= 0.5 - 0.25$
$= 0.25$

Expectation and Variance of RVs

Q2. Let X be a random variable with E[X] = 5 and Var(X) = 2. Let Y = 3X - 4.

- Find E[Y].
- Find Var(Y).

We can use the linearity of expectation, which states that E[aX + b] = aE[X] + b, where 'a' and 'b' are constants.

E[Y] = E[3X - 4]

= 3E[X] – 4 (using linearity of expectation)

= 3(5) - 4 (substituting E[X] = 5)

= 15 - 4

= 11

We can use the property of variance that states $Var(aX + b) = a^2Var(X)$, where 'a' and 'b' are constants. Notice that the constant term 'b' does not affect the variance.

Var(Y) = Var(3X - 4)

$= 3^2Var(X)$ (using the property of variance)

= 9(2) (substituting Var(X) = 2)

= 18

Co-Variance of RVs

Q3. Let U and V be two independent standard normal random variables, i.e., U,V ~

N(0,1). Define the new random variables:

R = 5 + 2U - 3UV

S = 2 - U + V

Find cov(R, S)

The covariance between two random variables R and S is defined as:
cov(R, S) = E[(R - E[R])(S - E[S])] = E[RS] - E[R]E[S]

Co-Variance of RVs

Covariance:

- **Definition:** $Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- **Relationship to Variance:** $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$ (This holds in general, whether or not X and Y are independent.)
- **Independence:** If X and Y are independent, $Cov(X, Y) = 0$. (The converse is not necessarily true.)

Standard Deviation:

- The standard deviation of X, denoted σ or $SD(X)$, is the square root of the variance: $\sigma = \sqrt{Var(X)}$. It provides a measure of the spread of the distribution in the original units of the random variable.

Co-Variance of RVs

Q1. Let U and V be two independent standard normal random variables, i.e., U,V ~ N(0,1). Define the new random variables: R = 5 + 2U - 3UV and S = 2 - U + V. Find cov(R, S)

The covariance between two random variables R and S is defined as:
cov(R, S) = E[(R - E[R])(S - E[S])] = E[RS] - E[R]E[S]

First, let's find the expected values of R and S:
•E[R] = E[5 + 2U - 3UV] = 5 + 2E[U] - 3E[UV]
Since U and V are independent, E[UV] = E[U]E[V]. Also, E[U] = E[V] = 0, as U and V are standard normal.
Therefore, E[R] = 5 + 2(0) - 3(0)(0) = 5

•E[S] = E[2 - U + V] = 2 - E[U] + E[V] = 2 - 0 + 0 = 2
Now, let's find E[RS]:
$E[RS] = E[(5 + 2U - 3UV)(2 - U + V)] = E[10 - 5U + 5V + 4U - 2U^2 + 2UV - 6UV + 3U^2V - 3UV^2]$
$= 10 - 5E[U] + 5E[V] + 4E[U] - 2E[U^2] + 2E[UV] - 6E[UV] + 3E[U^2V] - 3E[UV^2]$

Since U and V are standard normal, E[U] = E[V] = 0 and $E[U^2] = E[V^2] = 1$.
Also, since U and V are independent, E[UV] = E[U]E[V] = 0, $E[U^2V] = E[U^2]E[V] = 1 * 0 = 0$, and $E[UV^2] = E[U]E[V^2] = 0 * 1 = 0$.

Therefore, E[RS] = 10 - 5(0) + 5(0) + 4(0) - 2(1) + 2(0) - 6(0) + 3(0) - 3(0) = 10 - 2 = 8
Finally, we can find the covariance: cov(R, S) = E[RS] - E[R]E[S] = 8 - (5)(2) = 8 - 10 = -2

Probability Theory

Events and Sample Space

- Sample Space:
  - for a procedure Sample Space consists of all possible simple events; that is, the sample space consists of all outcomes that cannot be broken down any further
- Event
  - any collection of results or outcomes of a procedure
- Simple Event
  - an outcome or an event that cannot be further broken down into simpler components

- Sample space Ω - set of all possible outcomes of a random experiment
  - Dice roll: {1, 2, 3, 4, 5, 6}
  - Coin toss: {Tails, Heads}
- Event space $\mathcal{F}$ - subsets of elements in a sample space
  - Dice roll: {1, 2, 3} or {2, 4, 6}
  - Coin toss: {Tails}

Events and Sample Space

- A pair of dice are rolled. The sample space has 36 simple events:

1,1 1,2 1,3 1,4 1,5 1,6
2,1 2,2 2,3 2,4 2,5 2,6
3,1 3,2 3,3 3,4 3,5 3,6
4,1 4,2 4,3 4,4 4,5 4,6
5,1 5,2 5,3 5,4 5,5 5,6
6,1 6,2 6,3 6,4 6,5 6,6

where the pairs represent the numbers rolled on each dice.

- Which elements of the sample space correspond to the event that the sum of each dice is 4?

Probability

- The word 'Probability' means the chance of occurring of a particular event.
- It is generally possible to predict the future of an event quantitatively with a certain probability of being correct.
- The probability is used in such cases where the outcome of the trial is uncertain.

$$P(A) = \frac{\text{number of cases favourable to A}}{\text{number of possible outcomes}}$$

- P - denotes a probability.
- A, B, and C - denote specific events.
- P(A) - denotes the probability of event A occurring.

Probability

- Probability of an Event Defined over (Ω, F) s.t.
  - $0 < P(a) < 1$ for all a in F
  - $P(\Omega) = 1$
- Probability of an event which is certain to occur is **one**.
- Probability of an event which is impossible to **zero**.
- If the probability of happening of an event P(A) and that of not happening is P(A'), then $P(A) + P(A') = 1$,

  where, $0 \leq P(A) \leq 1$, $0 \leq P(A') \leq 1$.

Event Relations

- **Equally Likely Events:** Events are said to be equally likely if one of them cannot be expected to occur in preference to others. In other words, it means each outcome is as likely to occur as any other outcome.
  - *Example:* When a die is thrown, all the six faces, i.e., 1, 2, 3, 4, 5 and 6 are equally likely to occur.
- **Mutually Exclusive or Disjoint Events:** Events are called mutually exclusive if they cannot occur simultaneously.
  - *Example:* Suppose a card is drawn from a pack of cards, then the events getting a jack and getting a king are mutually exclusive because they cannot occur simultaneously.

Event Relations

- **Exhaustive Events:** The total number of all possible outcomes of an experiment is called exhaustive events.
  - Example: In the tossing of a coin, either head or tail may turn up. Therefore, there are two possible outcomes. Hence, there are two exhaustive events in tossing a coin.
- **Dependent Event:** Events are said to be dependent if occurrence of one affect the occurrence of other events.
- **Independent Events:** Events A and B are said to be independent if the occurrence of any one event does not affect the occurrence of any other event.

$$P(A \cap B) = P(A)\,P(B).$$

Event Relations

Example: A coin is tossed thrice, and all 8 outcomes are equally likely

A: "The first throw results in heads."

B: "The last throw results in Tails."

Prove that event A and B are independent.

Solution:

Sample Space: [HHH, HHT, HTH, THH, TTT, TTH, THT, HTT]

A: [HHH, HHT, HTH, HTT]

B: [HHT, TTT, THT, HTT]

A∩B: [HHT, HTT]

$P(A) = \frac{4}{8} = \frac{1}{2}$ $P(B) = \frac{4}{8} = \frac{1}{2}$

$P(A \cap B) = \frac{1}{2} x \frac{1}{2} = \frac{1}{4}$

Event Relations

Theorem 1: If A and B are two mutually exclusive events, then

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

Proof: Let the n=total number of exhaustive cases

n_1= number of cases favorable to A.

n_2= number of cases favorable to B.

Now, we have A and B two mutually exclusive events. Therefore, $n_1 + n_2$ is the number of cases favorable to A or B.

$$P(A \cup B) = \frac{\text{favorable cases}}{\text{Total number of exhaustive cases}} = \frac{n_1 + n_2}{n} = \frac{n_1}{n} + \frac{n_2}{n}$$

But we have, $P(A) = \frac{n_1}{n}$ and $P(B) = \frac{n_2}{n}$

Hence, $P(A \cup B) = P(A) + P(B)$.

Event Relations

Theorem2: If A and B are two events that are not mutually exclusive, then

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

Proof: Let n = total number of exhaustive cases

n_1 =number of cases favorable to A

n_2 = number of cases favorable to B

n_3 = number of cases favorable to both A and B

But A and B are not mutually exclusive. Therefore, A and B can occur simultaneously. So,$n_1+n_2-n_3$ is the number of cases favorable to A or B.

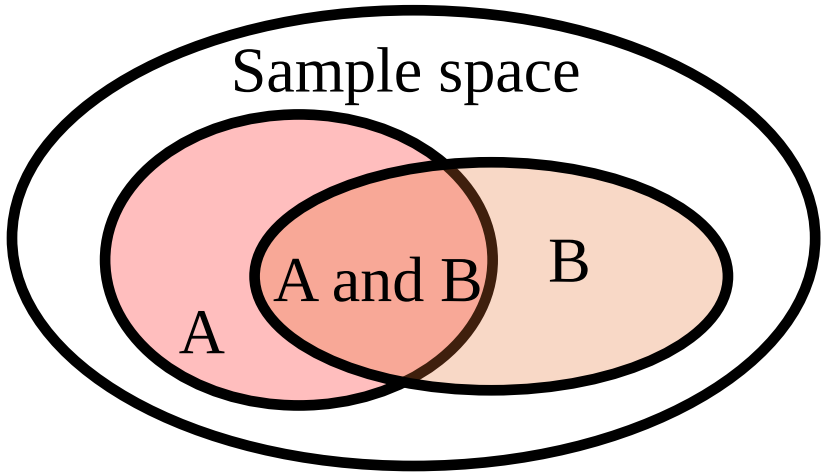
Therefore, $P(A \cup B)=\frac{n_1+n_2-n_3}{n}=\frac{n_1}{n}+\frac{n_2}{n}-\frac{n_3}{n}$

But we have, $P(A)=\frac{n_1}{n}$, $P(B)=\frac{n_2}{n}$ and $P(A\cap B)=\frac{n_3}{n}$

Hence, $P(A \cup B)=P(A)+P(B)- P(A\cap B)$.

Conditional Probability

- The probability of an event A based on the occurrence of another event B is termed conditional Probability. It is denoted as **P(A|B)** and represents the probability of A when event B has already happened.
- P(A | B) = P(A∩B)/ P(B)
- If the two events are independent:
  - ***P(A∩B) = P(A) ∗P(B)***
  - ***P(A/B) = P(A)***



Joint Probability:

The probability of two more events occurring together and at the same time is measured it is termed as Joint Probability.

Joint probability for two events A and B is denoted as, P(A∩B).

Bayes Theorem

- Bayes theorem, also known as the Bayes Rule, is used to determine the conditional probability of event A when event B has already happened.
- “The conditional probability of an event A, given the occurrence of another event B, is equal to the product of the probability event of B given A and the probability of A divided by the probability of event B.” i.e.

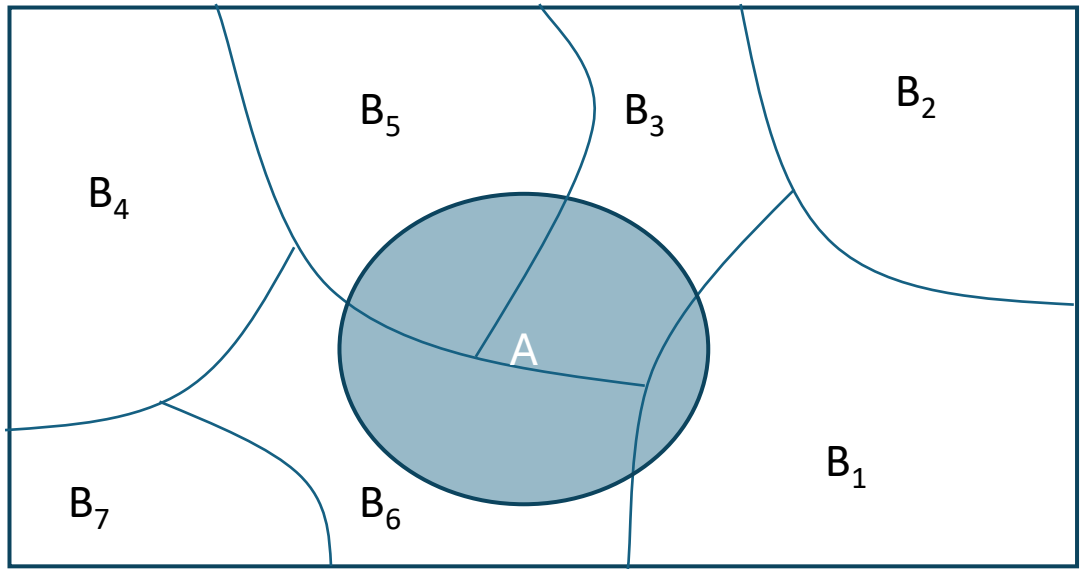
$$P(A|B) = P(B|A)P(A) / P(B) \quad \text{given} \quad P(B) \neq 0$$

- where,
  - P(A) and P(B) are the probabilities of events A and B
  - P(A|B) is the probability of event A when event B happens
  - P(B|A) is the probability of event B when event A happens

Theorem of Total Probability

- Let E1, E2,.....En are mutually exclusive and exhaustive events associated with a random experiment and lets E be an event that occurs with some Ei.
- Then,

$$P(E) = {}^{n}\sum_{i=1} P(E/E_i) \,.\, P(E_i)$$



$$p(A) = \sum P(B_i) P(A \mid B_i)$$

Questions

Q1. Two dice are thrown. The events A, B, C, D, E, F

A = getting even number on first die.

B= getting an odd number on the first die.

C = getting a sum of the number on dice ≤ 5

D = getting a sum of the number on dice > 5 but less than 10.

Show that:

1. A, B are a mutually exclusive event and Exhaustive Event.
2. A, C are not mutually exclusive.
3. C, D are a mutually exclusive event but not Exhaustive Event.
4. A'∩B' are a mutually exclusive and exhaustive event.

Questions

Q2. A bag contains 5 green and 7 red balls. Two balls are drawn. Find the probability that one is green and the other is red.

Q3. Find the probability of drawing a heart on each of two consecutive draws from well shuffled-packs of cards if the card is not replaced after the draw.

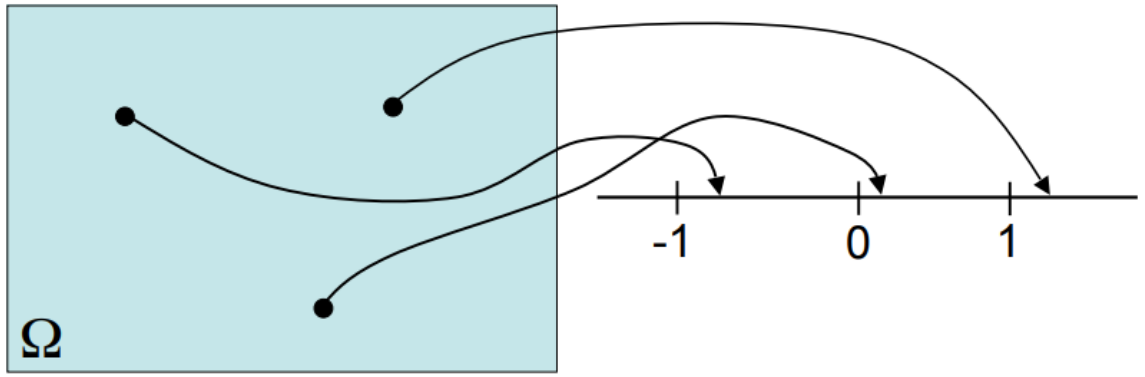
Q4. “X+Y=6 or X+Y=7” – given this (and only this), what is the probability of Y=5?

Q5. There are three urns containing 3 white and 2 black balls; 2 white and 3 black balls; 1 black and 4 white balls respectively. There is an equal probability of each urn being chosen. One ball is equal probability chosen at random. what is the probability that a white ball is drawn?

Probability Theory

Random Variable

- A random variable is a numerical quantity that is generated by a random experiment.
- A RV is any rule (i.e., function) that associates a number with each outcome in the sample space.



Example 1 : Machine Breakdowns

- Sample space : $S = \{electrical, mechanical, misuse\}$
- Each of these failures may be associated with a repair cost
- State space : $\{50, 200, 350\}$
- Cost is a random variable : 50, 200, and 350

Random Variable

- We will denote random variables by capital letters, such as X or Z, and the actual values that they can take by lowercase letters, such as x and z.
- A RV is called **discrete** if its possible values form a finite or countable set.
- A RV is called **continuous** if its possible values contain a whole interval of numbers.

Experiment	Number X	Possible Values of X	Type of RV
Roll two fair dice	Sum of the number of dots on the top faces	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	**discrete**
Flip a fair coin repeatedly	Number of tosses until the coin lands heads	1, 2, 3,4, …	**discrete**
Measure the voltage at an electrical outlet	Voltage measured	$118 \leq x \leq 122$	**continuous**
Operate a light bulb until it burns out	Time until the bulb burns out	$0 \leq x < \infty$	**continuous**

Probability Distribution

- The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable.
- The probabilities of a RV X must satisfy the following two conditions:
  - Each probability P(x) must be between 0 to 1: **$0 \leq P(x) \leq 1$**.
  - The sum of all the possible probabilities is 1: **$\sum P(x)=1$**.
- For a **discrete random variable**, *x*, the probability distribution is defined by a **probability mass function**, denoted by *f(x)*.
- This function provides the probability for each value of the random variable.

Probability Mass Function (p.m.f.)

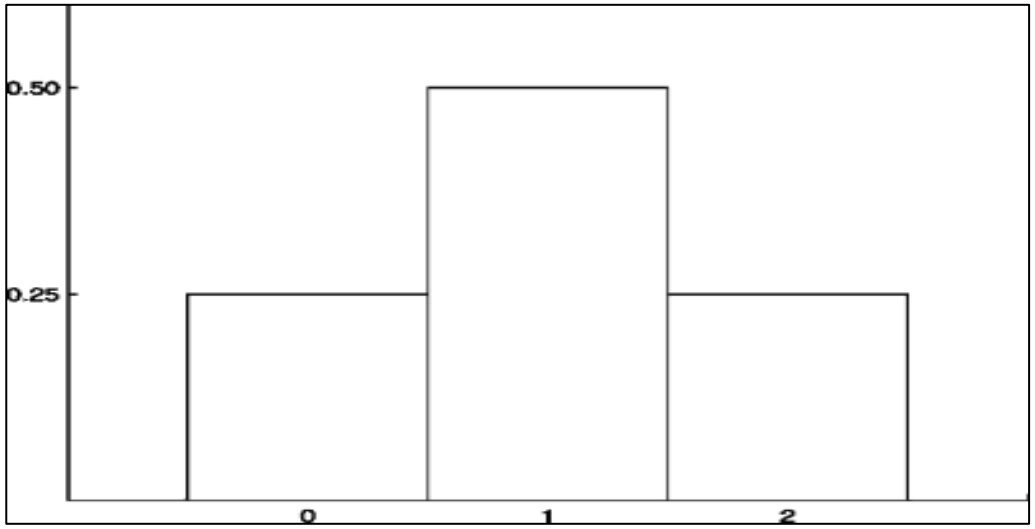
- A set of probability value p_i assigned to each of the values taken by the discrete random variable x_i
- $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$
- Probability : $P(X = x_i) = p_i$

Probability Mass Function

Q1. A fair coin is tossed twice. Let X be the number of heads that are observed.

- Construct the probability distribution of X.
- Find the probability that at least one head is observed.

The possible values that X can take are 0, 1, and 2.



- S={hh,ht,th,tt} are equally likely outcomes
  - X=0 to {tt}, X=1 to {ht,th}, and X=2 to {hh}.
- The probability distribution of X, is given by

x	0	1	2
P(x)	0.25	0.50	0.25

"At least one head" is the event X≥1, which is the union of the mutually exclusive events X=1 and X=2.

- Thus, P(X≥1) = P(1)+P(2) =0.50+0.25 =0.75

Probability Mass Function

Q2: A pair of fair dice is rolled. Let X denote the sum of the number of dots on the top faces.

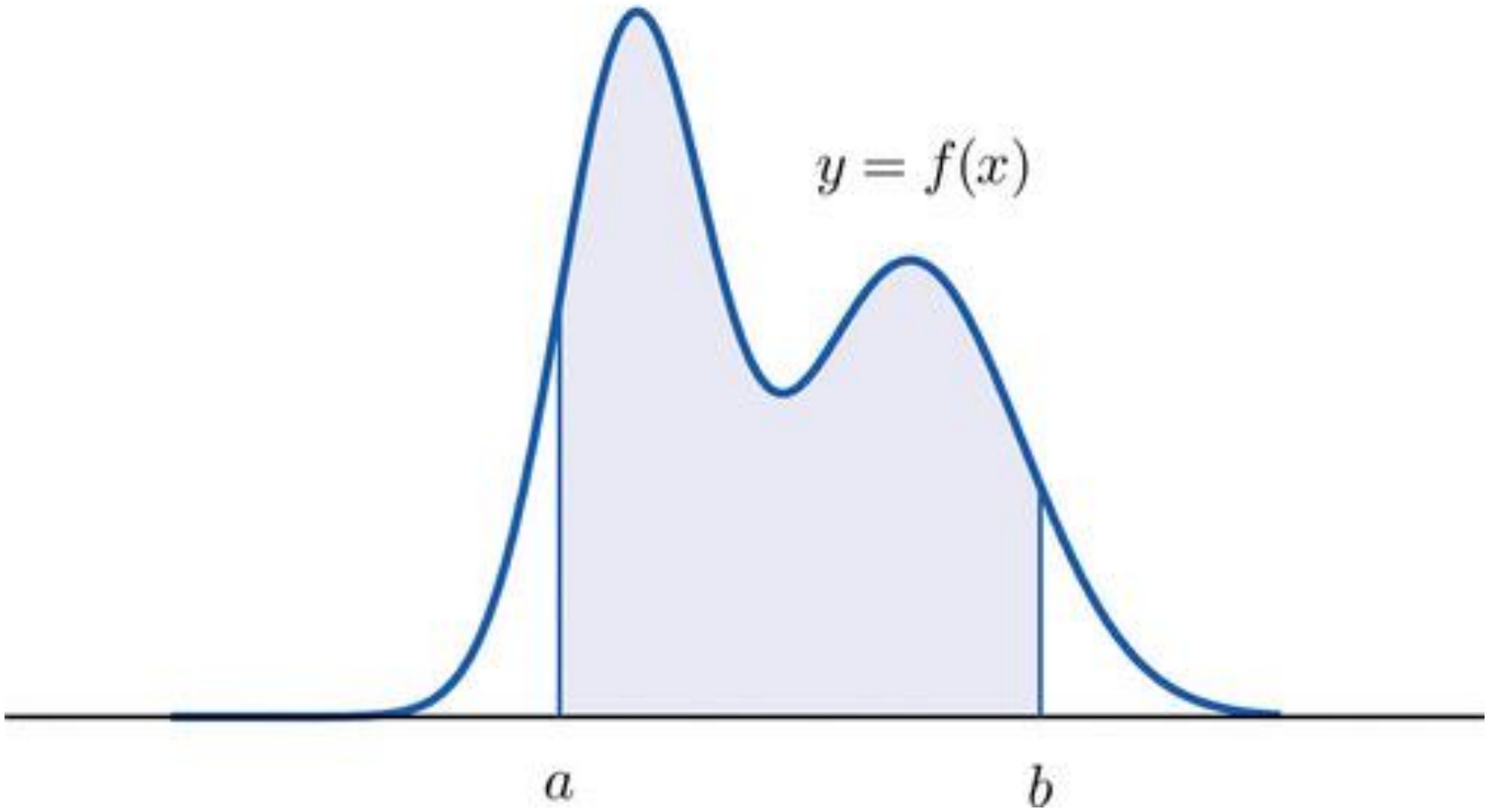
- Construct probability distribution of X for a pair of fair dice.
- Find $P(X \geq 9)$ Ans: 10/36
- Find the probability that X takes an even value. Ans: 0.5

Probability Density Function

- The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a function f(x), called a **Probability Density Function**.
- The probability that X assumes a value in interval [a,b] is equal to the area of the region that is bounded above by the graph of the equation $y=f(x)$, bounded below by the x-axis, and bounded on the left and right by the vertical lines through a and b.
- Density Function $f(x)$ must satisfy the following two conditions:
  - For all numbers x, $f(x) \geq 0$, so that the graph of $y=f(x)$ never drops below the x-axis.
  - The area of the region under the graph of $y=f(x)$ and above the x-axis is 1.

Probability Density Function

$P(a < X < b) = \text{area of shaded region}$



Probability Density Function

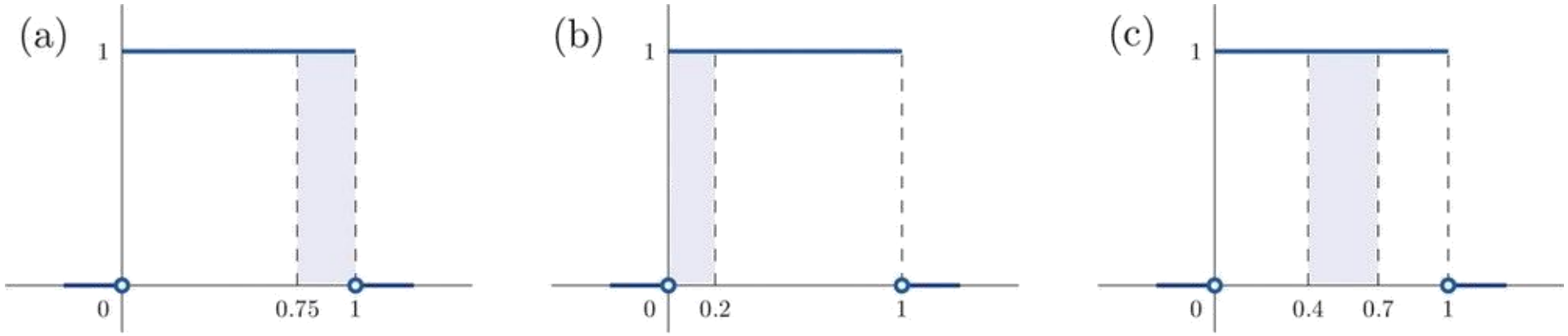
Q3. A random variable X has the uniform distribution on the interval $[0,1]$: the density function is $f(x)=1$ if x is between 0 and 1 and $f(x)=0$ for all other values of x (a uniform distribution).

- Find $P(X>0.75)$, the probability that X assumes a value greater than 0.75
- Find $P(X\leq 0.2)$, the probability that X assumes a value less than or equal to 0.2
- Find $P(0.4<X<0.7)$, the probability that X assumes a value between 0.4 and 0.7

Q4. A man arrives at a bus stop at a random time (that is, with no regard for the scheduled service) to catch the next bus. Buses run every 30 minutes without fail, hence the next bus will come any time during the next 30 minutes with evenly distributed probability (a uniform distribution). Find the probability that a bus will come within the next 10 minutes.

Probability Density Function

- P(X>0.75) is the area of the rectangle of height 1 and base length $1-0.75=0.25$, hence is $\text{base}\times\text{height}=(0.25)\cdot(1) = 0.25$
- P(X≤0.2) is the area of the rectangle of height 1 and base length $0.2-0=0.2$, hence is $\text{base}\times\text{height}=(0.2)\cdot(1)=0.2$
- P(0.4<X<0.7) is the area of the rectangle of height 1 and length $0.7-0.4=0.3$, hence is $\text{base}\times\text{height}=(0.3)\cdot(1)=0.3$



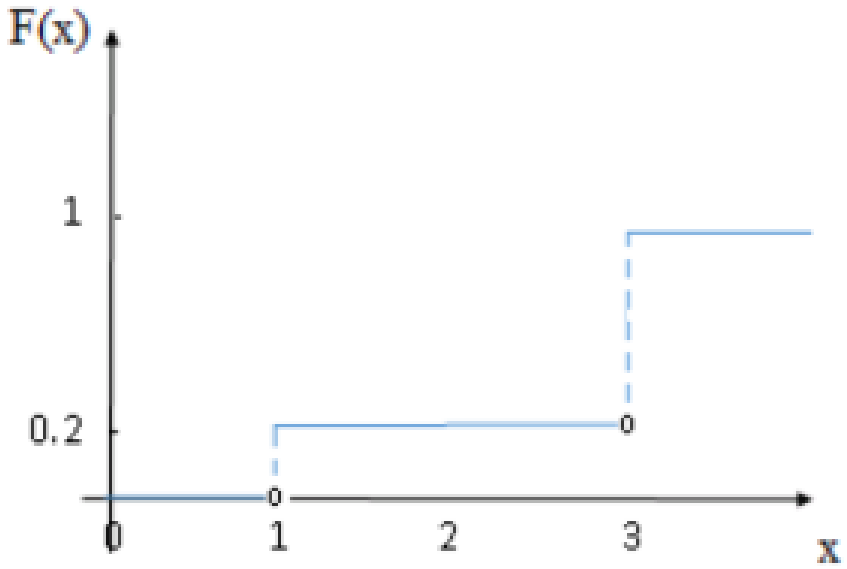
CDF of Random Variable

- Suppose that X is a random variable with values in $\mathbb{R}$. The Cumulative Distribution Function of X is the function F:$\mathbb{R}$→[0,1] is defined by

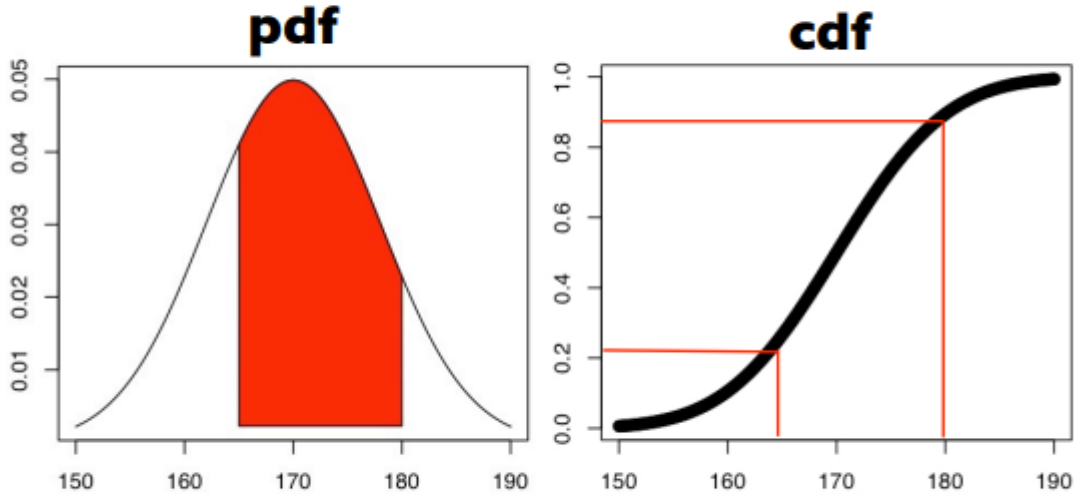
$$F(x)=P(X \leq x),\ x \in \mathbb{R}$$

- **Discrete RV**

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p(x_i),$$



Continuous rv: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$



$$P(a \leq X \leq b) = F(b) - F(a)$$

Probability Theory

Expectation of Random Variable

- The expected value, or mean, of a random variable x denoted by E(x) or μ is a weighted average of the values the random variable may assume.
- In the discrete case the weights are given by the probability mass function, and in the continuous case the weights are given by the probability density function.

- The expectation of a discrete random variable X is given by:

  $\mu = E(X) = \sum x f(x)$, where *f(x)* is the PMF of x.

- The expectation of a continuous random variable X is given by:

  $\mu = E(X) = \int x\, f(x) dx$, where *f(x)* is the PDF of x.

Variance of Random Variable

- The **variance** of a random variable, denoted by Var(x) or σ^2, is a weighted average of the squared deviations from the mean.
  - For discrete RV: $\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x)$
  - For continuous RV: $\text{Var}(x) = \sigma^2 = \int(x - \mu)^2 f(x)dx$
- The **standard deviation**, denoted σ, is the positive square root of the variance.

Median of Random Variable

- The **median** of the discrete random variable X, is the value of x for which $P(X \leq x)$ is greater than or equal to 0.5 and $P(X \geq x)$ is greater than or equal to 0.5.

Median of Random Variable

- Let X be a continuous rv with probability density function, f(x). The median of X can be obtained by solving for c in the equation below:

$$\int_{-\infty}^{c} f(x)dx=0.5$$

- That is, it is the value for which the area under the curve from negative infinity to c is equal to 0.50.

Quantiles of Random Variable

- Let X is a real-valued random variable with Cumulative Distribution Function F.
- For $p \in (0,1)$, a value of x is called **a quantile of order *p*** for the distribution if

  $$F(x-)=P(X < x) \leq p \text{ and } F(x)=P(X \leq x) \geq p .$$

- A quantile of order *p* is a value where the graph of the cumulative distribution function crosses *p*.
- Median is also called 50^{th} percentile.

Questions

Q1. A service organization in a large town organizes a raffle each month. One thousand raffle tickets are sold for $1 each. Each has an equal chance of winning. First prize is $300, second prize is $200, and third prize is $100. Let, X denote the net gain from the purchase of one ticket.

- Construct the probability distribution of X
- Find the probability of winning any money in the purchase of one ticket.
- Find the expected value of X, and interpret its meaning.

Questions

a) If a ticket is selected as the first prize winner, the net gain to the purchaser is the $300 prize less the $1 that was paid for the ticket, hence $X=300-11=299$. There is one such ticket, so $P(299)=0.001$

Applying the same “income minus outgo” principle to the second and third prize winners and to the 997 losing tickets yields the probability distribution:

x	299	199	99	−1
P(x)	0.001	0.001	0.001	0.997

b) Let W denote the event that a ticket is selected to win one of the prizes. Using the table

$$P(W) = P(299)+P(199)+P(99) = 0.003$$

c) $E(X) = (299)\cdot(0.001) + (199)\cdot(0.001) + (99)\cdot(0.001) + (-1)\cdot(0.997)$

$= -0.4$

Questions

Q2. A discrete rv X has following probability distribution:

x	−1	0	1	4
P(x)	0.2	0.5	c	0.1

Find

a) c — 0.2

b) P(0) — 0.5

c) P(X>0) — 0.3

d) P(X≥0) — 0.8

e) P(X≤−2) — 0

f) The mean μ of X — 0.4

g) The variance σ^2 of X — 1.84

h) The standard deviation σ of X — 1.3565

Questions

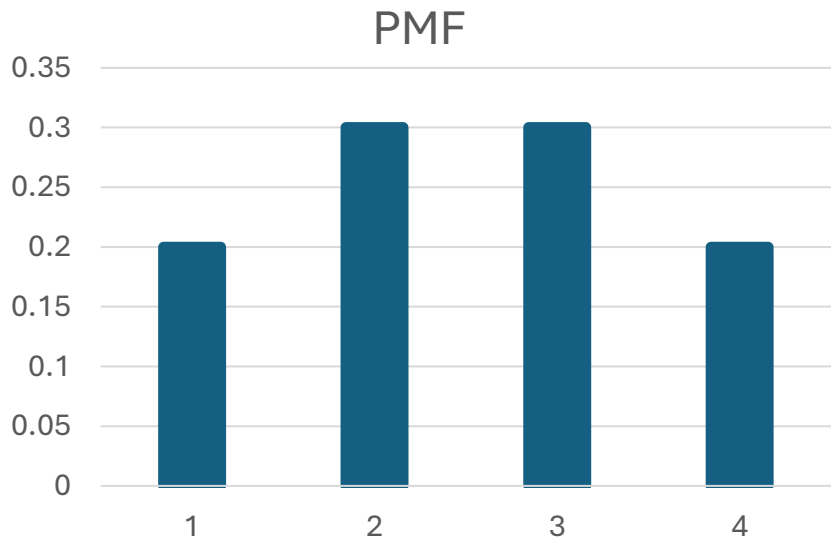
Q3. Given the following probability density function of a discrete random variable,

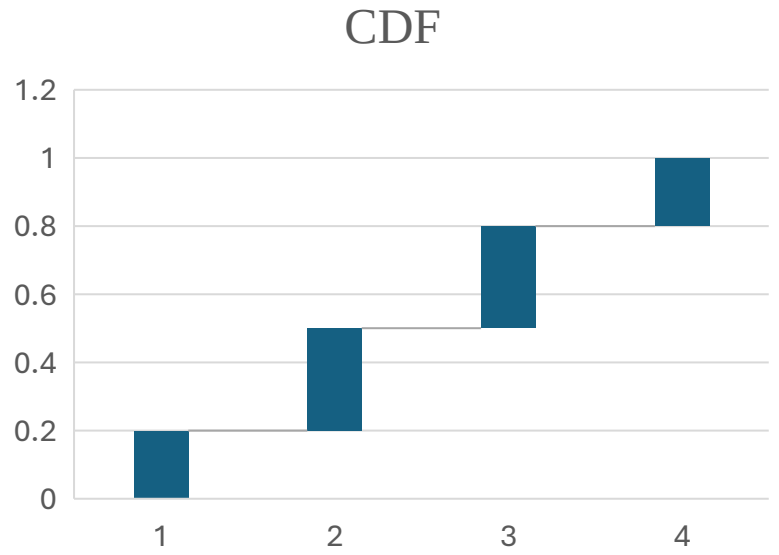
calculate the median of the distribution: $f(x) = \begin{cases} 0.2 & x = 1, 4 \\ 0.3 & x = 2, 3 \end{cases}$

- Solution: The median of the distribution above is 2 because;

  $P(X \leq 2) = P(X=1) + P(X=2) = 0.2 + 0.3 = 0.5$ and,

  $P(X \geq 2) = P(X=2) + P(X=3) + P(X=4) = 0.3 + 0.3 + 0.2 = 0.8$

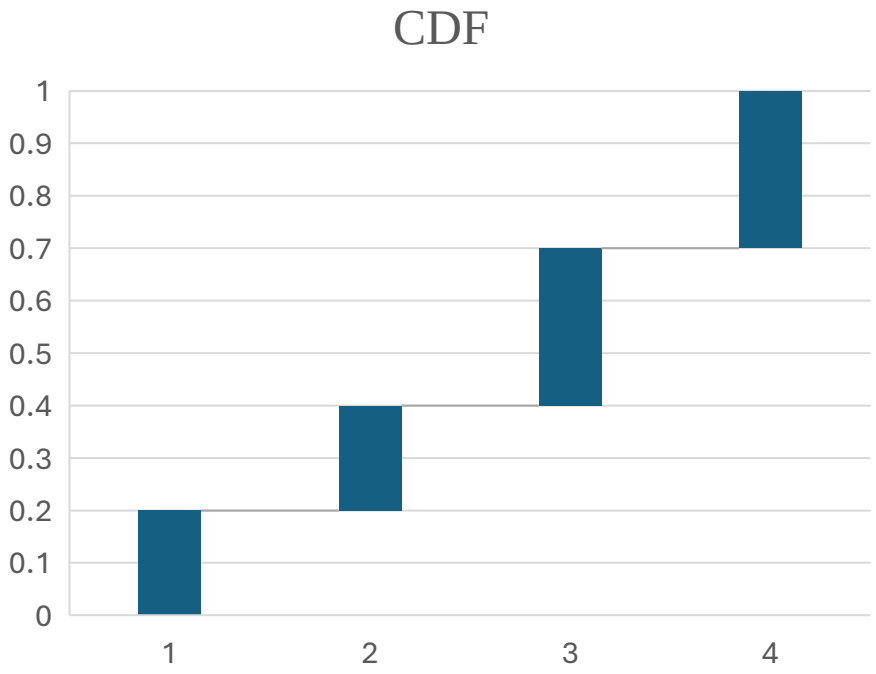




Questions

Q4. Given the following probability density function of a discrete random variable, calculate the 75th Percentile of the distribution:

$$f(x) = \begin{cases} 0.2 & x = 1, 4 \\ 0.3 & x = 2, 3 \end{cases}$$



Solution: The 75th percentile of the distribution is 4 because;

$P(X < 3) = P(X=1)+P(X=2)$

$= 0.2+0.3 = 0.5$ and,

$P(X \leq 3) = P(X=1)+P(X=2)+P(X=3)$

$= 0.2+0.3+0.3=0.8$

Probability Theory

Joint Distribution of RVs

- In real life, we are often interested in two (or more) random variables at the same time. For example,
  - we might measure the height and weight of an object, or
  - frequency of exercise and rate of heart disease in adults,
  - level of air pollution and rate of respiratory illness in cities,
  - number of Facebook friends and age of Facebook members
- Joint distribution allows us to compute probabilities of events involving both variables and understand the relationship between the variables.

Joint Distribution of Discrete RVs

- Suppose X and Y are two discrete random variables.
  - X takes values $\{x_1, x_2, \ldots, x_n\}$ and Y takes values $\{y_1, y_2, \ldots, y_m\}$. The ordered pair (X, Y) take values in the product$\{(x_1, y_1),(x_1, y_2), \ldots(x_n, y_m)\}$.
- The joint probability mass function (joint pmf) of X and Y is the function $p(x_i, y_j)$ giving the probability of the joint outcome $X = x_i$, $Y = y_j$.

Joint probability mass function must satisfy two properties:

1. $0 \leq p(x_i, y_j) \leq 1$
2. The total probability is 1.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(xi, yj) = 1$$

$X \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\cdots$	$p(x_1, y_j)$	$\cdots$	$p(x_1, y_m)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\cdots$	$p(x_2, y_j)$	$\cdots$	$p(x_2, y_m)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$	$\cdots$	$p(x_i, y_j)$	$\cdots$	$p(x_i, y_m)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$	$\cdots$	$p(x_n, y_j)$	$\cdots$	$p(x_n, y_m)$

Joint Distribution of Discrete RVs

Q1. Roll two dice. Let X be the value on the first die and let T be the total on both dice. Draw the joint probability table.

$X \backslash T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

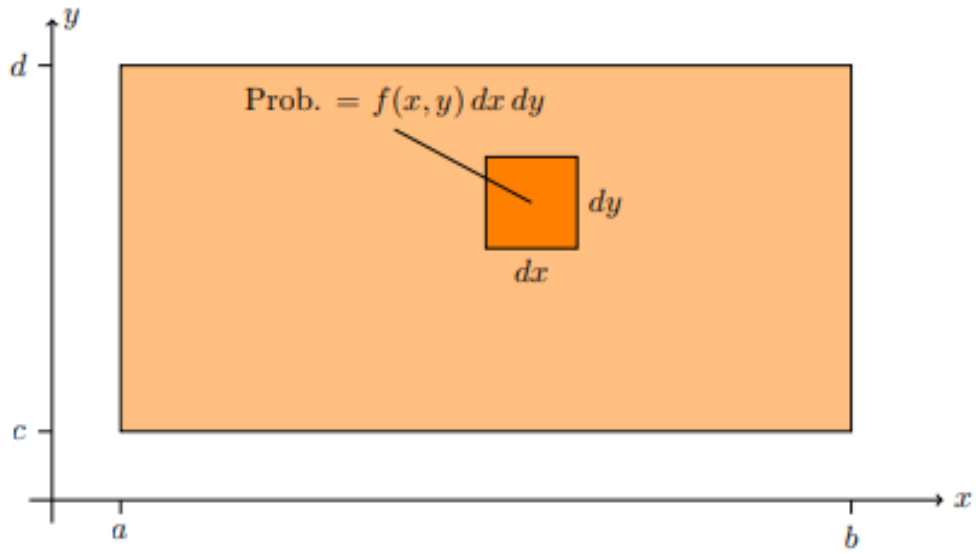
Q2. Roll two dice. Let X be the value on the first die and let Y be the value on the second die. Then both X and Y take values 1 to 6 and the joint pmf is p(i, j) = 1/36 for all i and j between 1 and 6. Draw the Joint probability table and find the probability of event B= ‘ X-Y ≥ 2’.

Joint Distribution of Continuous RVs

If X takes values in [a, b] and Y takes values in [c, d] then the pair (X, Y) takes values in the product [a, b] × [c, d].

- The joint probability density function (joint pdf) of X and Y is a function f(x, y) giving the probability density at (x, y).
- That is, the probability that (X, Y) is in a small rectangle of width dx and height dy around (x, y) is f(x, y)dxdy.
- A joint PDF must satisfy:
  1. 0 ≤ f(x, y)
  2. The total probability is 1.

$$\int_c^d \int_a^b f(x, y)dxdy = 1$$



Joint Cumulative Distributions RVs

Suppose X and Y are jointly-distributed random variables. We will use the notation '$X \leq x$, $Y \leq y$' to mean the event '$X \leq x$ and $Y \leq y$'. The joint cumulative distribution function (joint cdf) is defined as

$$F(x, y) = P(X \leq x, Y \leq y)$$

Continuous case: If X and Y are continuous random variables with joint density $f(x, y)$ over the range $[a, b] \times [c, d]$ then the joint cdf is given by the double integral

$$F(x, y) = \int_c^y \int_a^x f(u, v)\, du\, dv.$$

To recover the joint pdf, we differentiate the joint cdf. Because there are two variables we need to use partial derivatives:

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y).$$

Discrete case: If X and Y are discrete random variables with joint pmf $p(x_i, y_j)$ then the joint cdf is give by the double sum

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j).$$

Joint Distribution of Continuous RVs

Q3. Let X & Y both take values in [0,1] with density f(x, y) = 4xy.

i. Show f(x, y) is a valid joint PDF,

ii. Visualize the event A = 'X < 0.5 and Y > 0.5' and find its probability.

To show $f(x, y)$ is a valid joint pdf we must check that it is positive (which it clearly is) and that the total probability is 1.

$$\text{Total probability} = \int_0^1 \int_0^1 4xy\,dx\,dy = \int_0^1 \left[2x^2y\right]_0^1 dy = \int_0^1 2y\,dy = 1. \quad \text{QED}$$

The event A is just the upper-left-hand quadrant. Because the density is not constant we must compute an integral to find the probability.

$$P(A) = \int_0^{.5} \int_{.5}^1 4xy\,dy\,dx = \int_0^{.5} \left[2xy^2\right]_{.5}^1 dx = \int_0^{.5} \frac{3x}{2}\,dx = \boxed{\frac{3}{16}}.$$

Q4. Let X & Y both take values in [0,1] with density f(x, y) = 4xy. Find Joint CDF of X and Y.

Marginal Density RVs

Given a joint density for X and Y, we define the marginal density of X to be

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

and the marginal density of Y to be

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

As usual, we restrict the integral to the region where f is positive when that is not the entire plane.

Example Consider

$$f(x,y) = \begin{cases} x+y & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal density of X is given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \int_0^1 x+y\,dy \\ &= x + 1/2 \end{aligned}$$

Marginal Distributions RVs

Q5. Suppose (X, Y) takes values on the unit square [0, 1] × [0, 1] with joint pdf $f(x, y) = \frac{3}{2}(x^2 + y^2)$. Find the marginal pdf $f_X(x)$ and use it to find P(X < 0.5).

$$f_X(x) = \int_0^1 \frac{3}{2}(x^2 + y^2)\, dy = \left[\frac{3}{2}x^2 y + \frac{y^3}{2}\right]_0^1 = \boxed{\frac{3}{2}x^2 + \frac{1}{2}}.$$

$$P(X < 0.5) = \int_0^{0.5} f_X(x)\, dx = \int_0^{0.5} \frac{3}{2}x^2 + \frac{1}{2}\, dx = \left[\frac{1}{2}x^3 + \frac{1}{2}x\right]_0^{0.5} = \boxed{\frac{5}{16}}.$$

Independence in RVs

- Events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- The joint distribution (or density or mass) of Independent RVs is the product of the marginals.

Definition: Jointly-distributed random variables X and Y are independent if their joint cdf is the product of the marginal cdf's:

$$F(X, Y) = F_X(x)F_Y(y).$$

For discrete variables this is equivalent to the joint pmf being the product of the marginal pmf's.:

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j).$$

For continous variables this is equivalent to the joint pdf being the product of the marginal pdf's.:

$$f(x, y) = f_X(x)f_Y(y).$$

Independence in RVs

Example 12. For discrete variables independence means the probability in a cell must be the product of the marginal probabilities of its row and column. In the first table below this is true: every marginal probability is $1/6$ and every cell contains $1/36$, i.e. the product of the marginals. Therefore X and Y are independent.

$X \backslash Y$	1	2	3	4	5	6	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(y_j)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Example 13. For continuous variables independence means you can factor the joint pdf or cdf as the product of a function of x and a function of y.

(i) Suppose X has range $[0, 1/2]$, Y has range $[0, 1]$ and $f(x, y) = 96x^2y^3$ then X and Y are independent. The marginal densities are $f_X(x) = 24x^2$ and $f_Y(y) = 4y^3$.

(ii) If $f(x, y) = 1.5(x^2 + y^2)$ over the unit square then X and Y are not independent because there is no way to factor $f(x, y)$ into a product $f_X(x)f_Y(y)$.

(iii) If $F(x, y) = \frac{1}{2}(x^3y + xy^3)$ over the unit square then X and Y are not independent because the cdf does not factor into a product $F_X(x)F_Y(y)$.

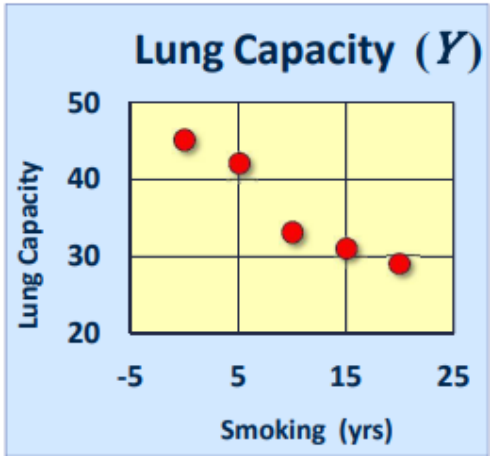
Probability Theory

Covariance of RVs

- Random Variables may change in relation to each other. Covariance is a measure of association of two variables.
- If positive, then both variables increase or decrease together. If negative, then they vary in opposite manner.
- Covariance measures how much the movement in one variable predicts the movement in a corresponding variable
- **Example:** investigate relationship between cigarette smoking and lung capacity as shown in figure.

N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

- Variables smoking and lung capacity *covary* inversely, like



Covariance of RVs

- Average product of deviation measures extent to which variables co-vary, the degree of linkage between them.

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

Deviation of data 1 from mean

Deviation from mean of data 2

Cigs (X)				Cap (Y)
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
		$\sum$= -215		

Evaluation yields,

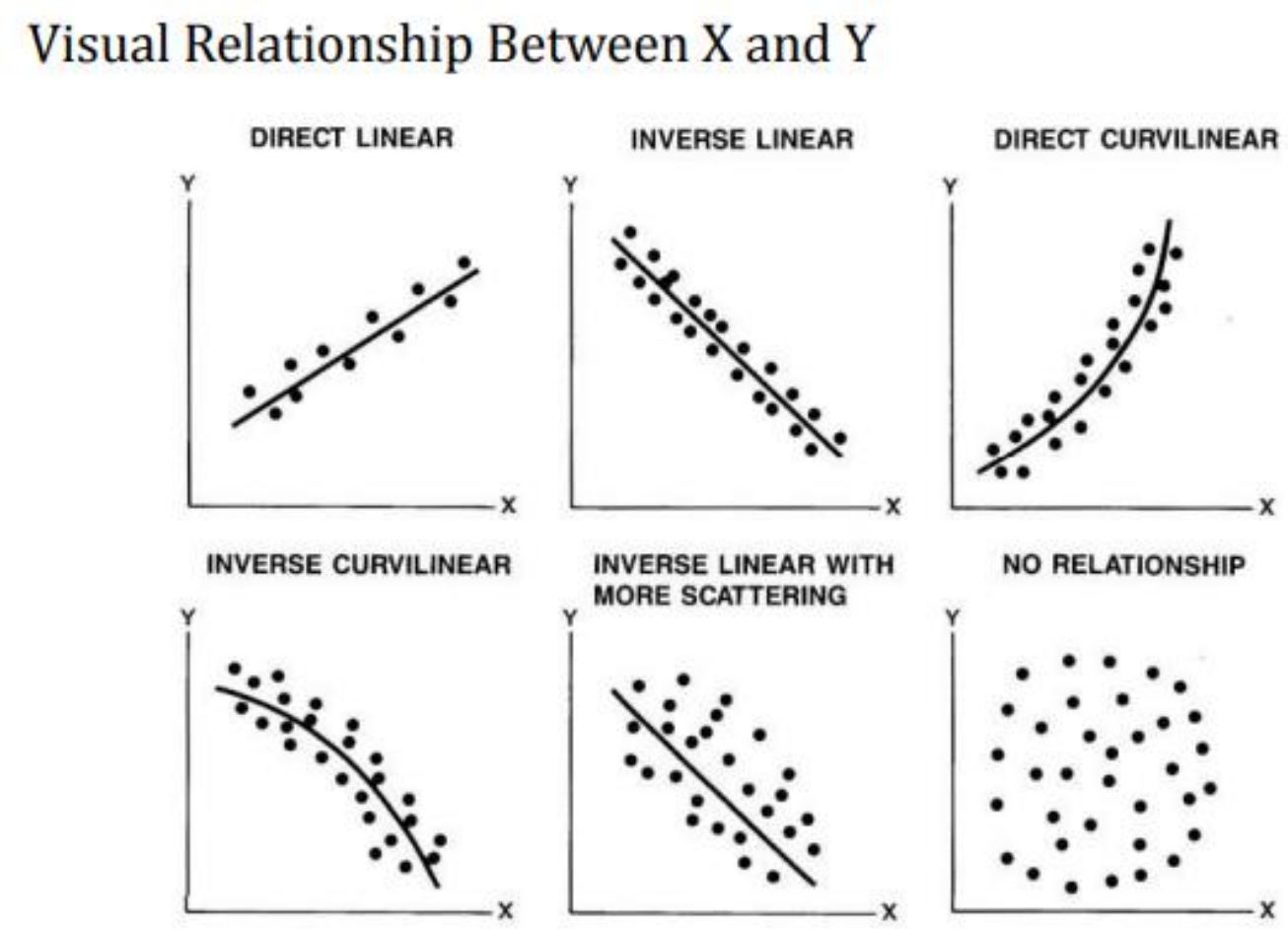
$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

Correlation of RVs

- A measure which determines the standard change in one variable due to change in the other variable.
- Correlation is of two types, i.e. positive correlation or negative correlation.
- Correlation can take any value between -1 to +1, where in values close to +1 represents strong positive correlation and values close to -1 is an indicator of strong negative correlation.
- Measures of correlation:
  - Scatter diagram
  - Rank correlation coefficient

Correlation of RVs

- Correlation using scatter plot



Correlation of RVs

- Correlation coefficient

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

- Covariance, Cov(X,Y) is dependent upon the units of X & Y.
- Correlation, Corr(X,Y), scales covariance by the standard deviations of X & Y so that it lies between 1 & –1

$$Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Where σ is the Standard deviation

Common Distributions of RVs

- Uniform distribution
- Poisson distribution
- Normal distribution
- Standard normal distribution

Common Distributions of RVs

The Uniform Distribution

A random variable X is said to be *uniformly distributed* in $a \leq x \leq b$ if its density function is

$$f(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and the distribution is called a *uniform distribution*.

The distribution function is given by

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \leq x < b \\ 1 & x \geq b \end{cases}$$

The mean and variance are, respectively,

$$\mu = \frac{1}{2}(a+b), \qquad \sigma^2 = \frac{1}{12}(b-a)^2$$

Common Distributions of RVs

The Poisson Distribution

Let X be a discrete random variable that can take on the values 0, 1, 2, . . . such that the probability function of X is given by

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots \tag{13}$$

where λ is a given positive constant. This distribution is called the *Poisson distribution* (after S. D. Poisson, who discovered it in the early part of the nineteenth century), and a random variable having this distribution is said to be *Poisson distributed*.

Mean	$\mu = \lambda$
Variance	$\sigma^2 = \lambda$
Standard deviation	$\sigma = \sqrt{\lambda}$

When p is small and n is fixed, Mean = λ = np, where

- **n** is the Number of Trails
- **p** is Probability of Success

Common Distributions of RVs

The Normal Distribution

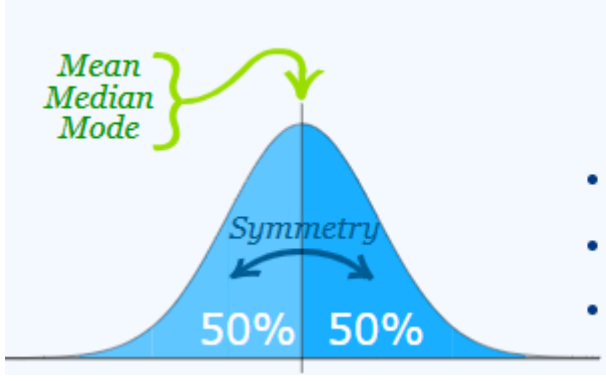
One of the most important examples of a continuous probability distribution is the *normal distribution*, sometimes called the *Gaussian distribution*. The density function for this distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty \tag{4}$$

where μ and σ are the mean and standard deviation, respectively. The corresponding distribution function is given by

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(v-\mu)^2/2\sigma^2}\, dv \tag{5}$$

If X has the distribution function given by (5), we say that the random variable X is *normally distributed* with mean μ and variance σ^2.



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

Common Distributions of RVs

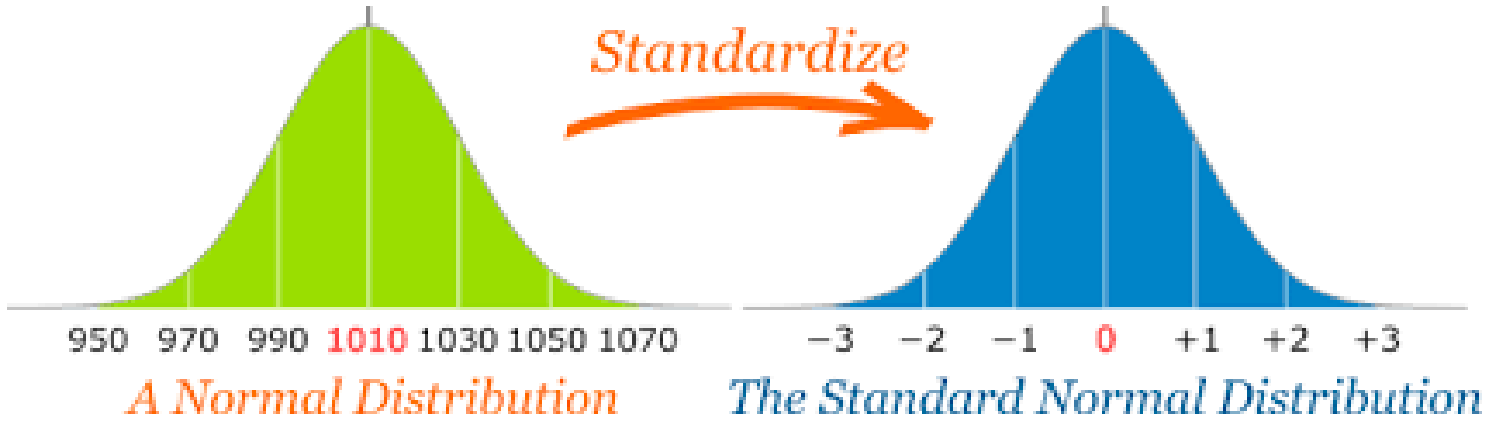
Standard normal distribution, also known as the z-distribution

- In this distribution, the **mean (average) is 0** and the **standard deviation (a measure of spread) is 1**.
- This creates a **bell-shaped curve** that is symmetrical around the mean ie. 0.
- The random variable of a standard normal distribution is known as the standard score or a z-score.

$$z = (X - \mu) / \sigma$$

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Where
$-\infty < z < \infty$



Central Limit Theorem

When large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.

Let us assume we have a random variable X.

Let σ be its standard deviation and μ is the mean of the random variable.

Now as per the Central Limit Theorem, the sample mean $\bar{X}$ will approximate to the normal distribution which is given as $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.

Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample Mean = Population Mean = μ

$$\text{Sample Standard Deviation} = \frac{\text{Standard Deviation}}{n}$$

OR

$$\text{Sample Standard Deviation} = \frac{\sigma}{\sqrt{n}}$$

Statistics

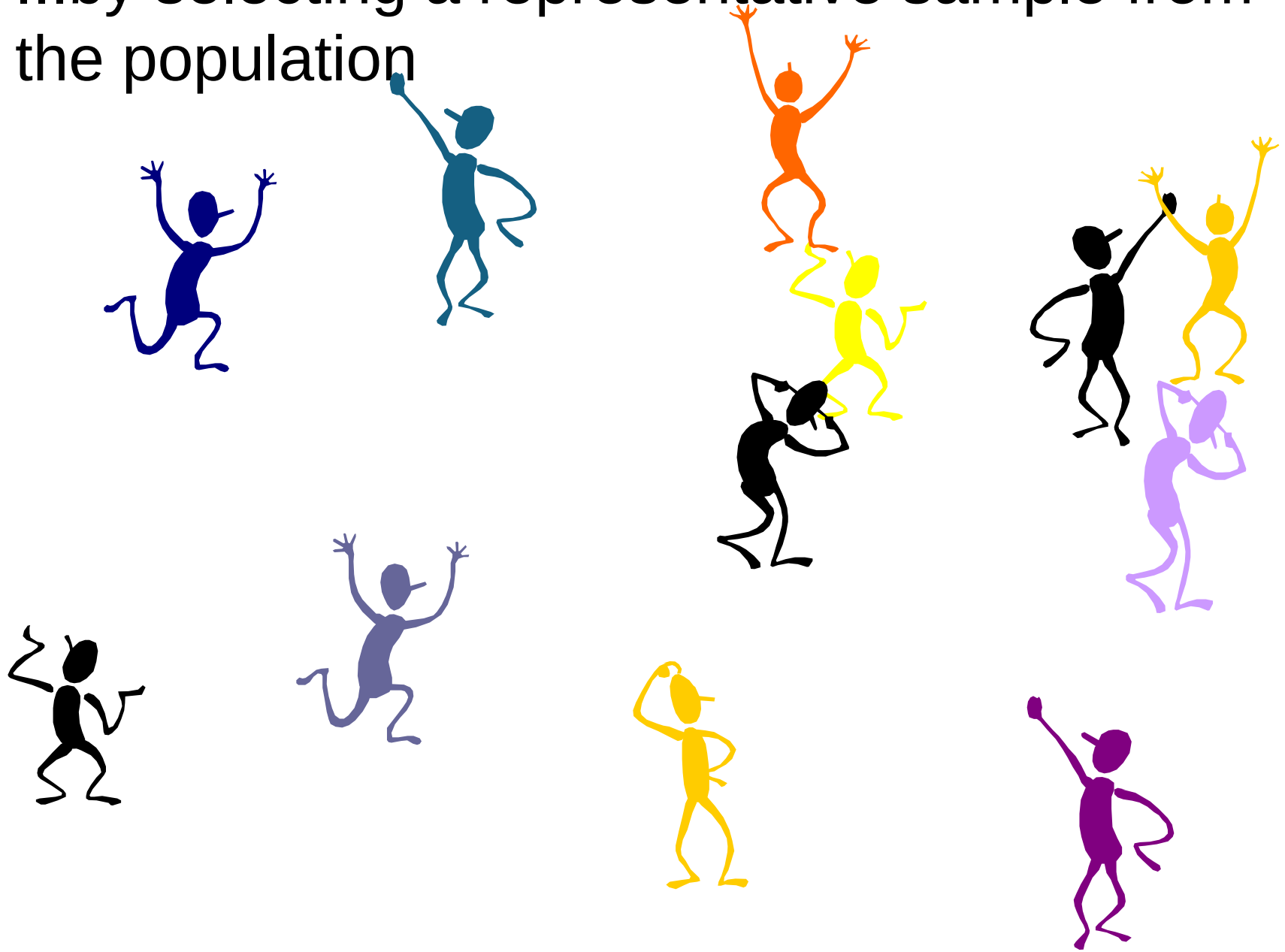
Sampling

- **Population** – A group that includes all the cases (individuals, objects, or groups) in which the researcher is interested.
- **Sample** – A relatively small subset from a population.

- **Simple Random Sample** – A sample designed in such a way as to ensure that
  (1) every member of the population has an equal chance of being chosen and
  (2) (2) every combination of N members has an equal chance of being chosen.
- This can be done using a computer, calculator, or a table of random numbers

Population inferences can be made...

...by selecting a representative sample from the population



How about family of three?

Num Girls	child #1	child #2	child #3
0	B	B	B
1	B	B	G
1	B	G	B
1	G	B	B
2	B	G	G
2	G	B	G
2	G	G	B
3	G	G	G

Probability distributions: Permutations

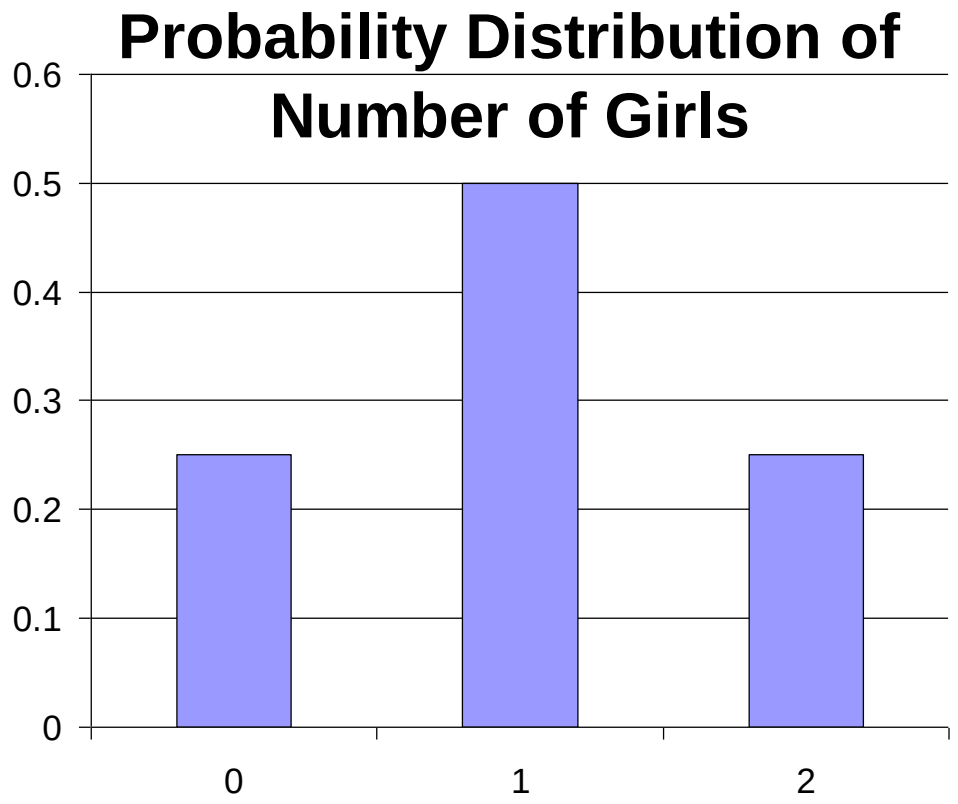
What is the probability distribution of number of girls in families with two children?

2 GG

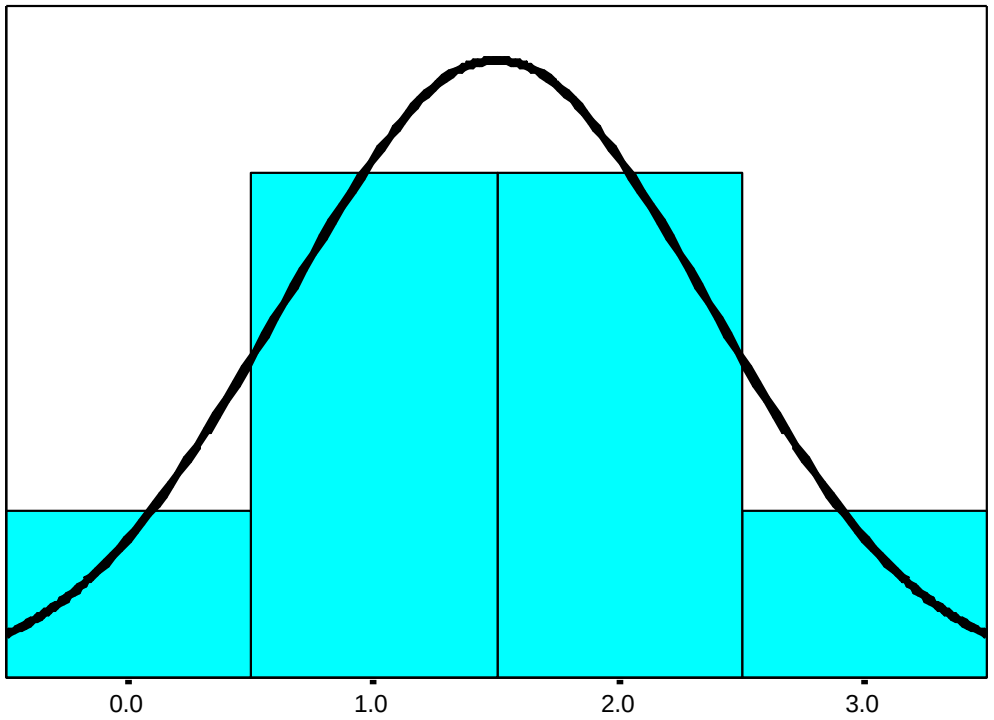
1 BG

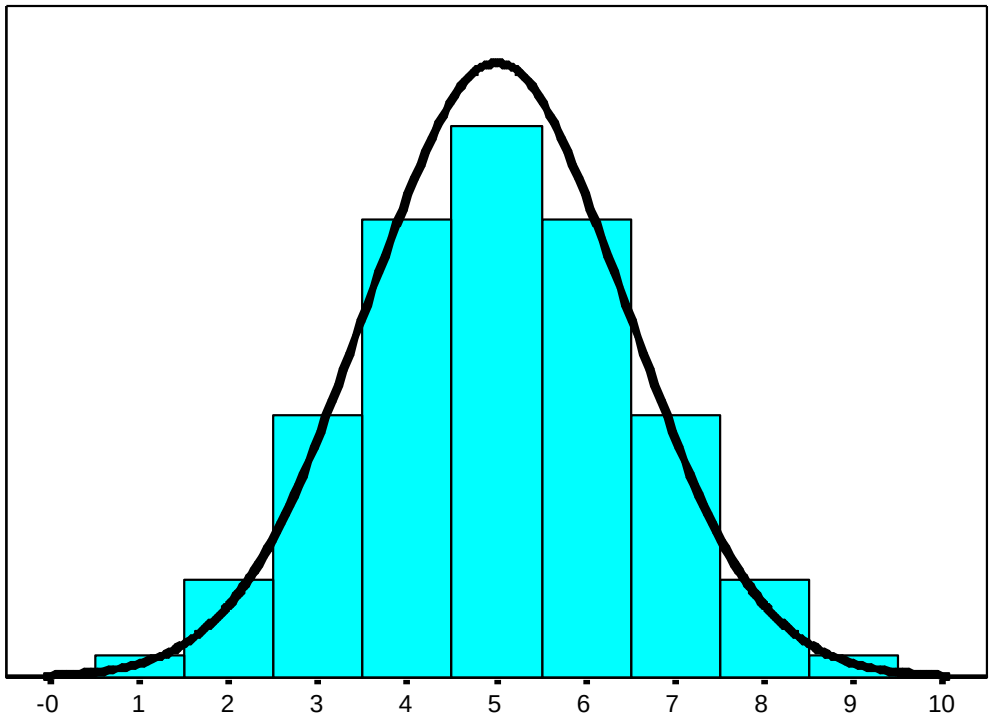
1 GB

0 BB



As family size increases, the distribution looks more and more normal.





Coin toss

- Toss a coin 30 times
- Tabulate results
- Think of the coin tosses as samples of all possible coin tosses

Sampling Distribution

- Imagine repeatedly taking samples of the same size from the large population and calculating a statistic (like the mean or variance) for each sample.
- The probability distribution of these calculated statistics is called the sampling distribution.
- Aim of sampling
  - Reduces cost of research (e.g. political polls)
  - Generalize about a larger population (e.g., benefits of sampling city r/t neighborhood)
  - In some cases (e.g. industrial production) analysis may be destructive, so sampling is needed

Central Limit Theorem

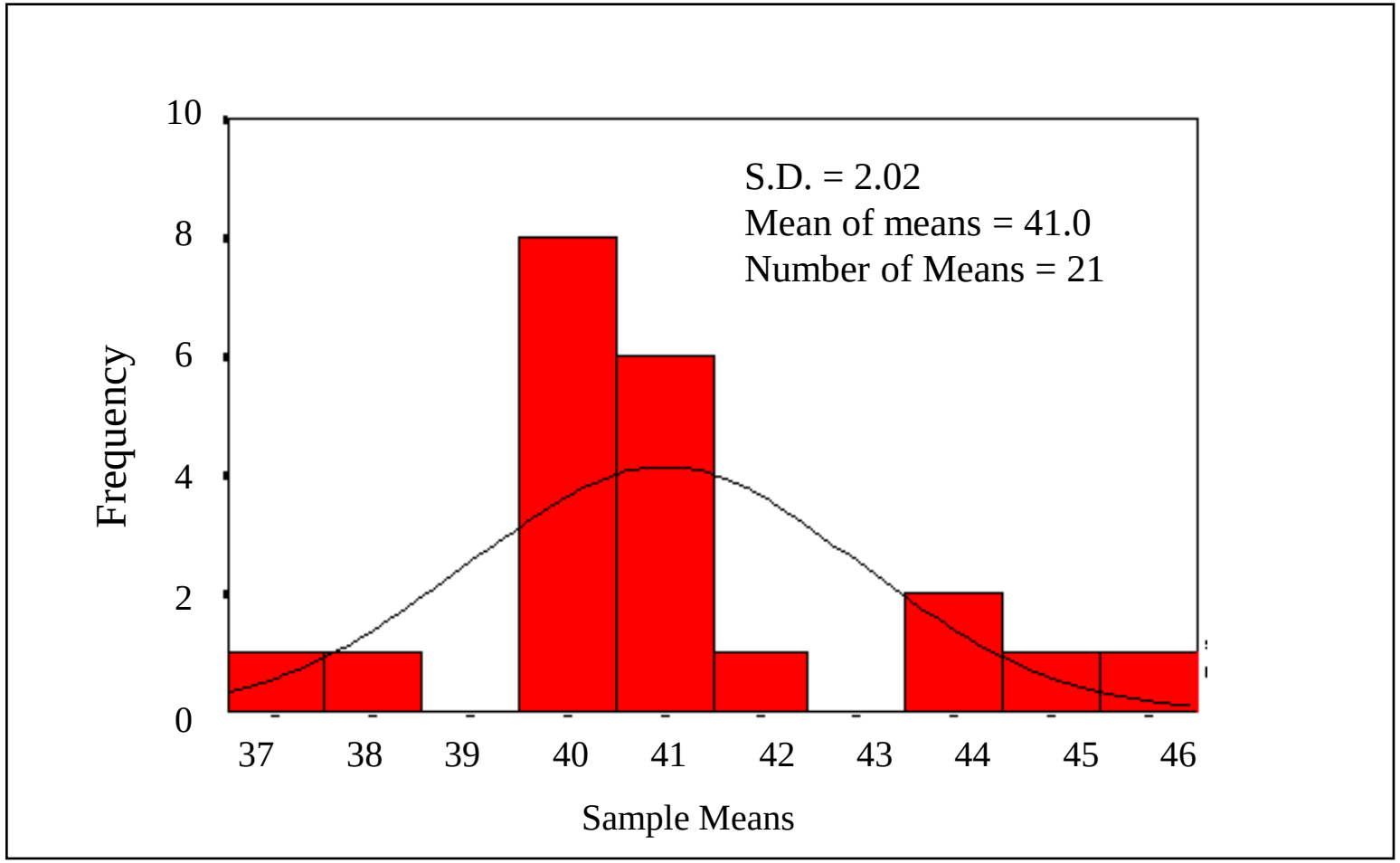
- No matter what we are measuring, the distribution of any measure across all possible samples we could take, approximates a normal distribution, as long as the number of cases in each sample is about 30 or larger.

If we repeatedly drew samples from a population and calculated the mean of a variable or a percentage, those sample means or percentages would be normally distributed.

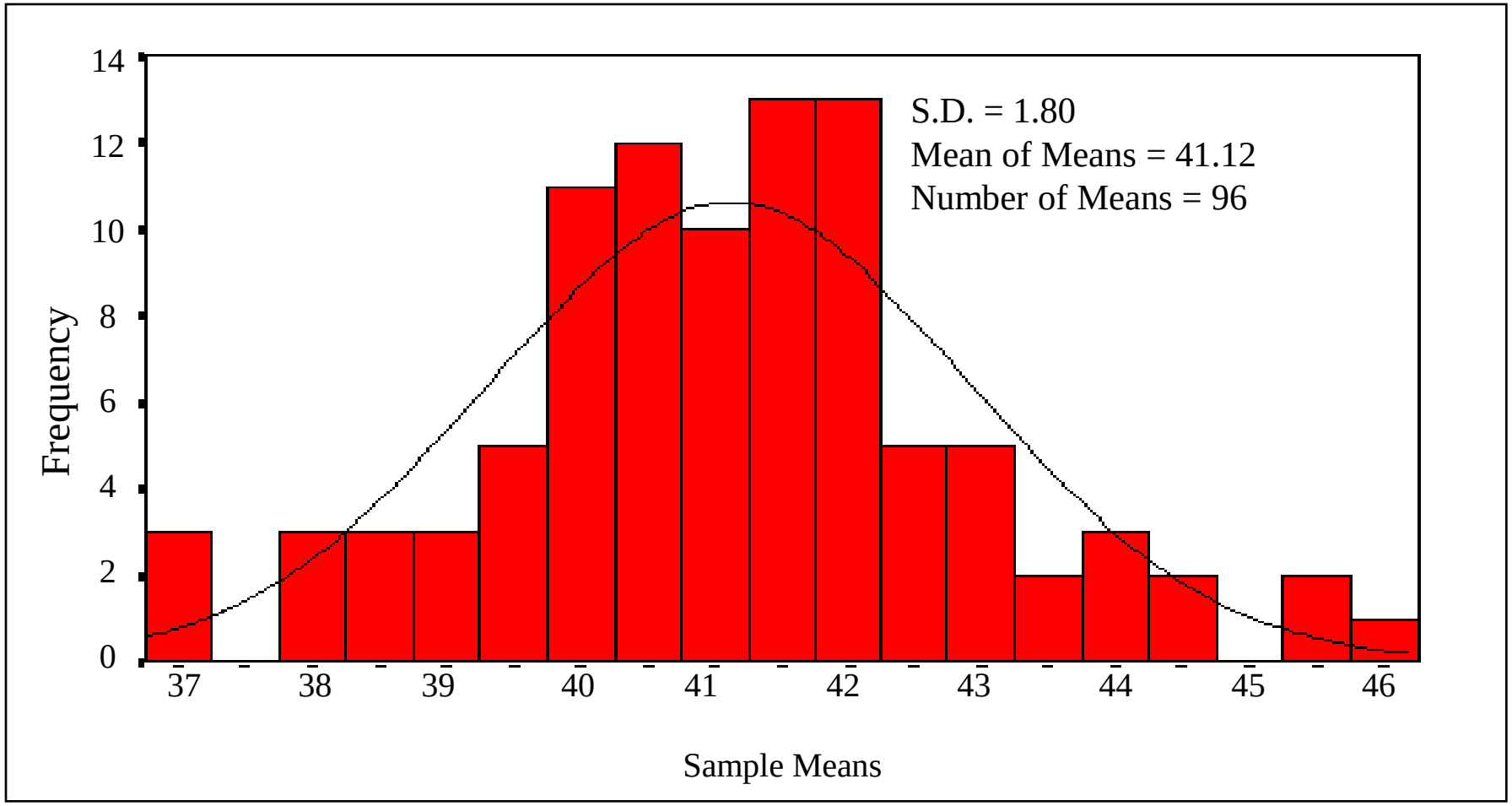
The Mean and Standard Deviation of the Sample Mean

- Suppose we wish to estimate the mean μ of a population. In actual practice we would typically take just one sample.
- Imagine however that we take sample after sample, all with same sample size n, and compute the sample mean $\bar{x}$ each time.
- The sample mean $\bar{x}$ is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty.
- Consider $\bar{X}$, as a random variable of the sample mean, and write x for the values that it takes.
- The random variable $\bar{X}$ has a mean, denoted $\mu_{\bar{x}}$, and a standard deviation, denoted $\sigma_{\bar{x}}$.

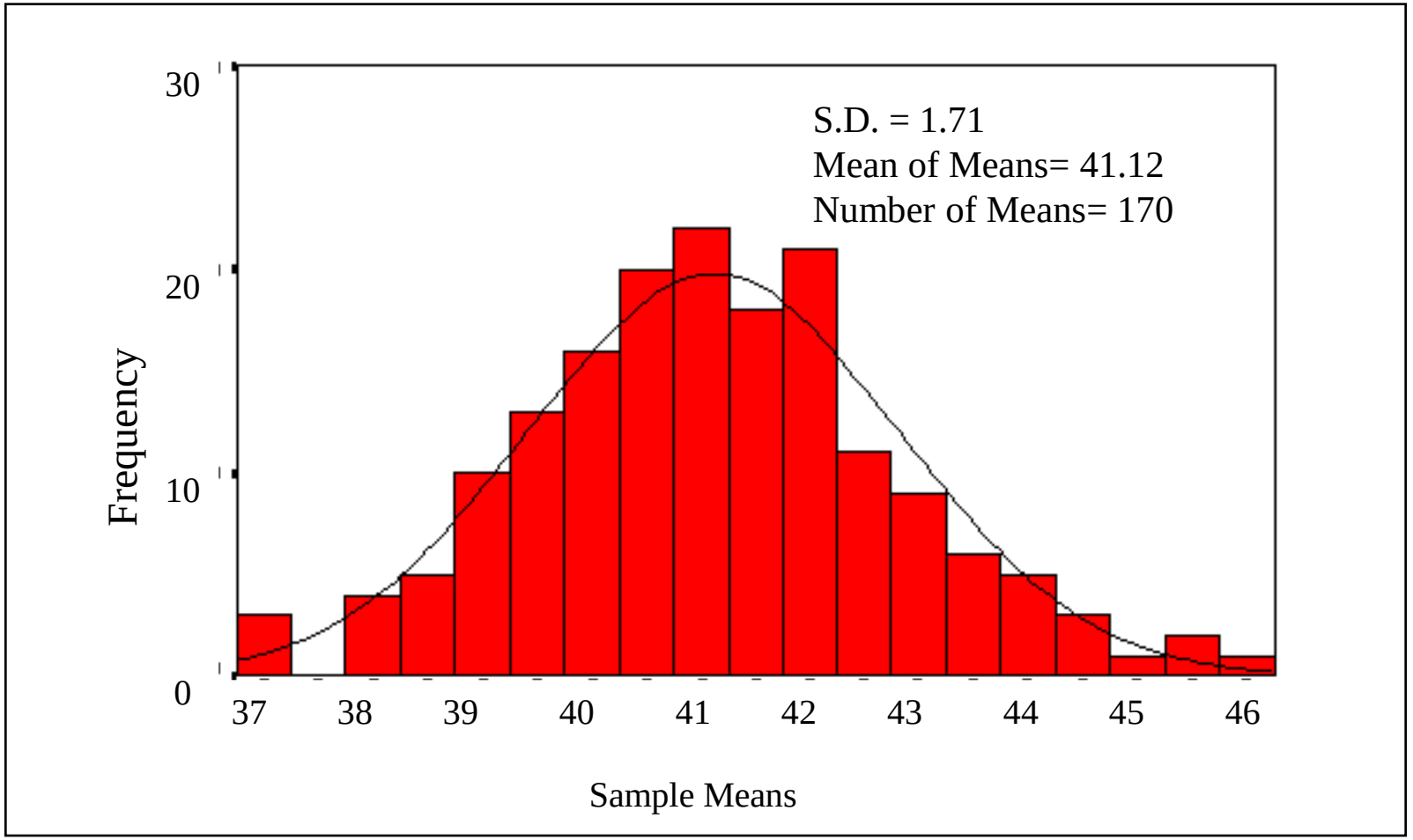
Distribution of Sample Means with 21 Samples



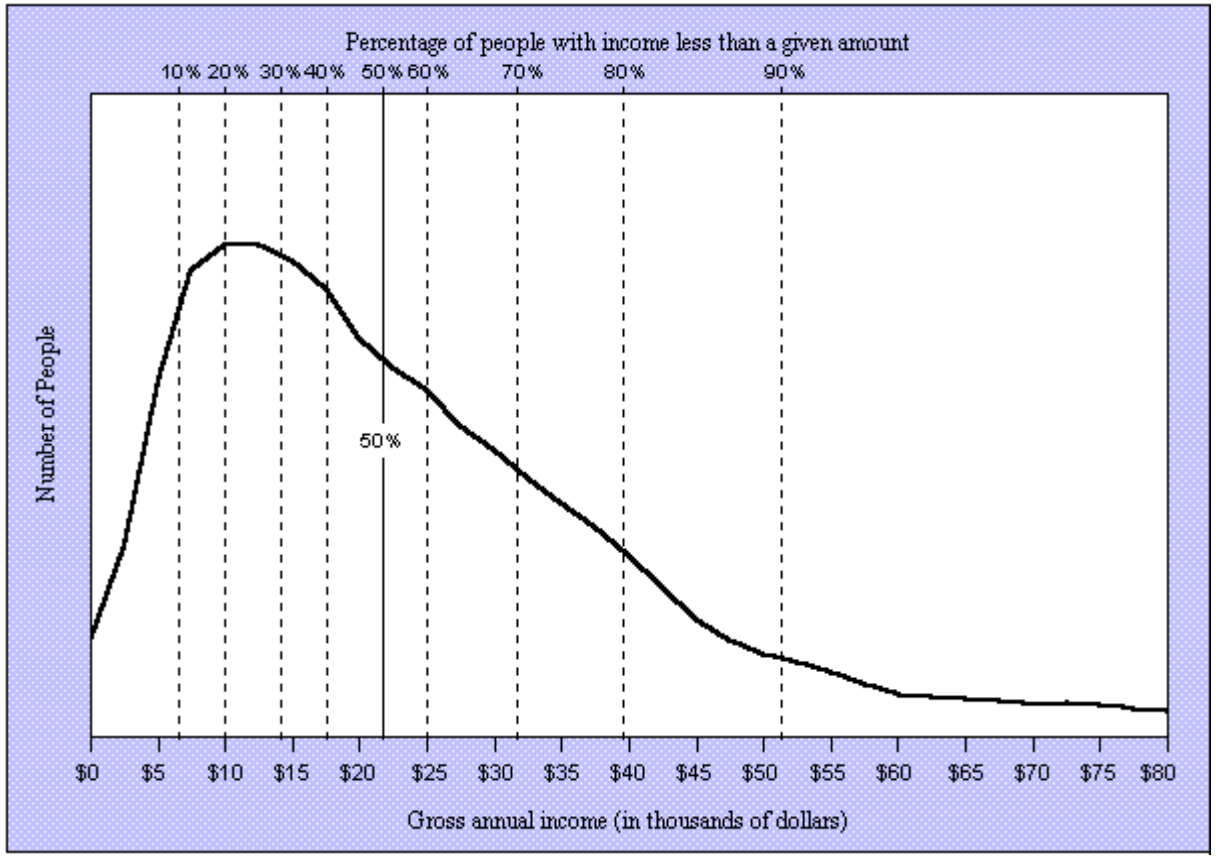
Distribution of Sample Means with 96 Samples



Distribution of Sample Means with 170 Samples

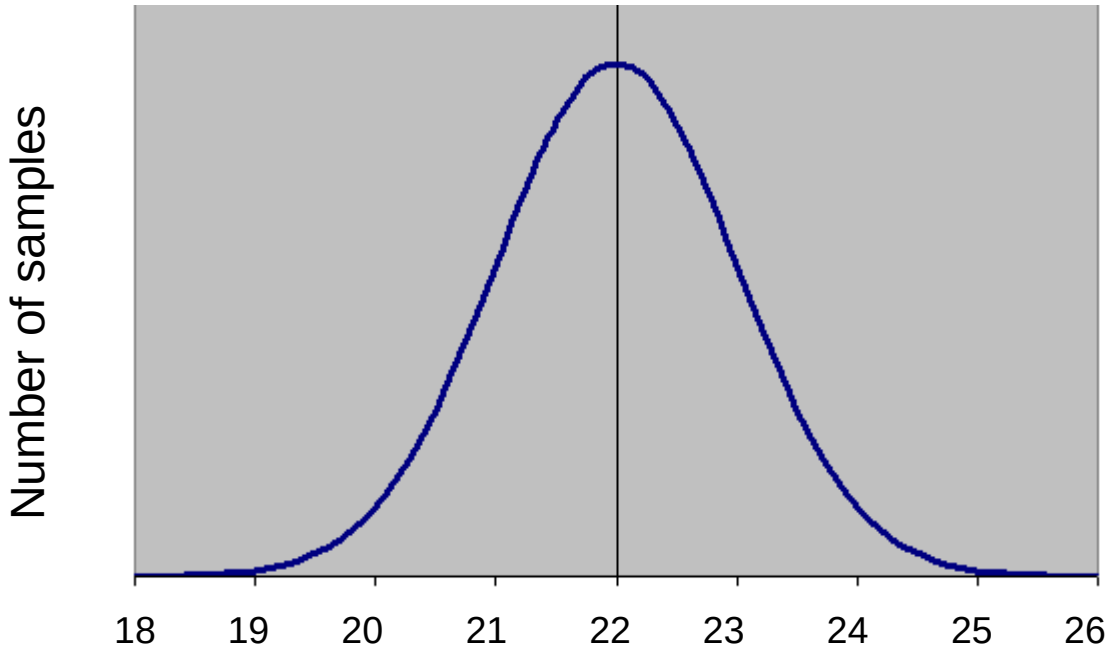


Most empirical distributions are not normal:



U.S. Income distribution 1992

But the sampling distribution of mean income over many samples *is* normal



Sampling Distribution of Income, 1992 (thousands)

Central Limit Theorem

When large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.

Let us assume we have a random variable X.

Let σ be its standard deviation and μ is the mean of the random variable.

Now as per the Central Limit Theorem, the sample mean $\bar{X}$ will approximate to the normal distribution which is given as $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.

Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

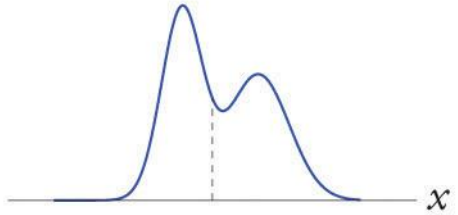
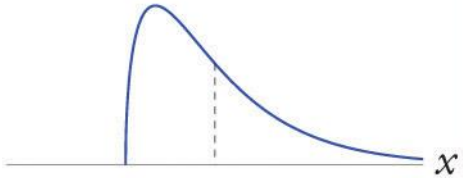
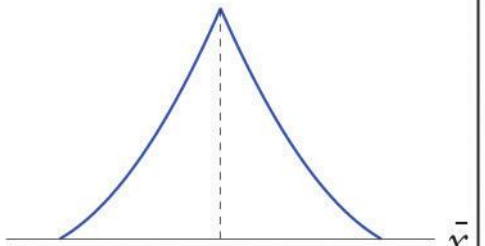
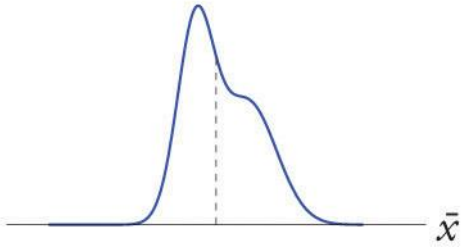
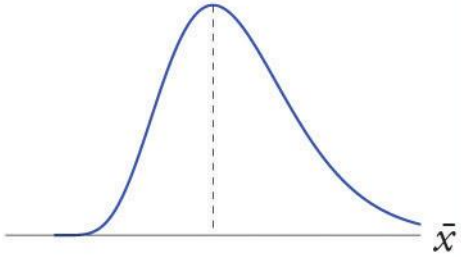
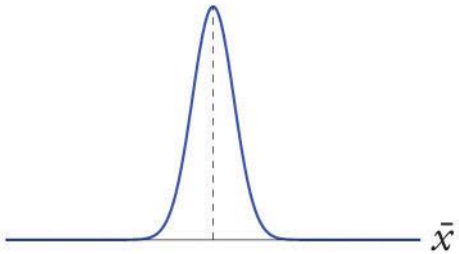
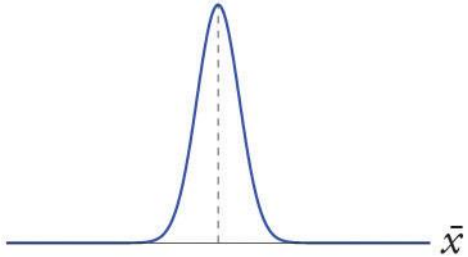
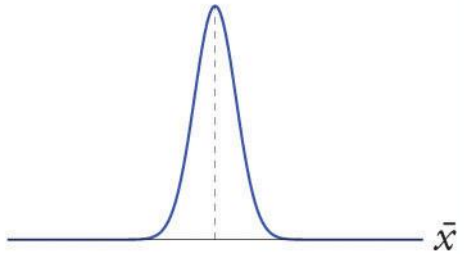
Sample Mean = Population Mean = μ

$$\text{Sample Standard Deviation} = \frac{\text{Standard Deviation}}{n}$$

OR

$$\text{Sample Standard Deviation} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

Population distribution			
Sampling distribution of $\overline{X}$ with $n = 5$			
Sampling distribution of $\overline{X}$ with $n = 30$			

Q1.

The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13,525$ and $\sigma = \$4,180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$?

Solution

Since $n = 100$, the formulas yield

$$\mu_{\bar{X}} = \mu = \$13,525$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

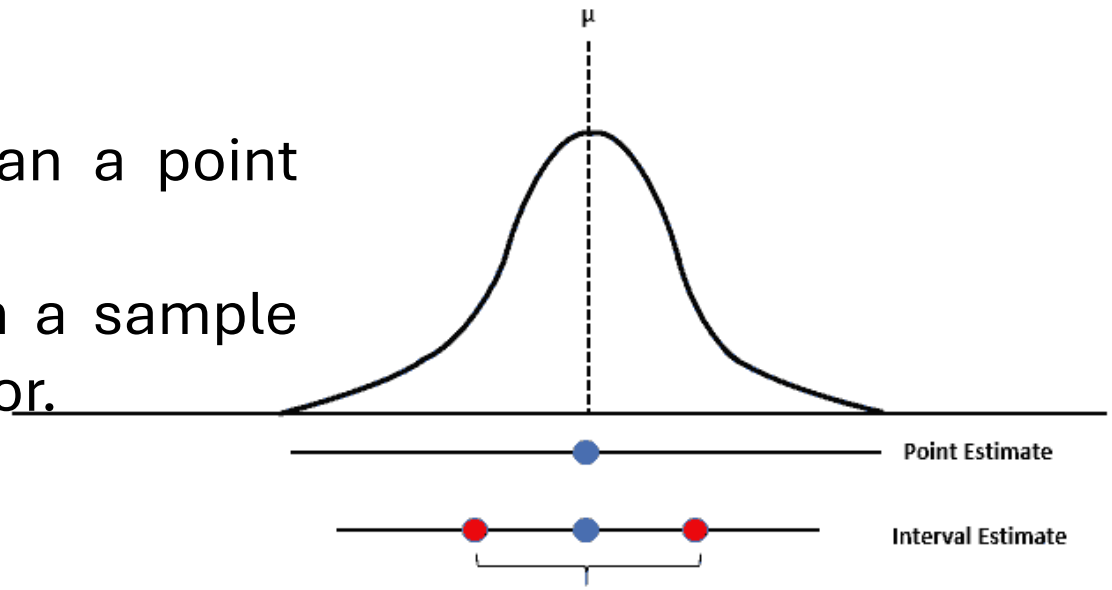
Point Estimate and Interval Estimate

- A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- A point estimate is a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter. In other words, we derive the point estimate from a single value in the sample and use it to estimate the population value.
- Take a sample, find $\bar{x}$. It is a close approximation of μ. But, depending on your sample size, that may not be a good point estimate.
- In fact, the probability that a single sample statistic is equal to the population parameter is very unlikely.

Point Estimate and Interval Estimate

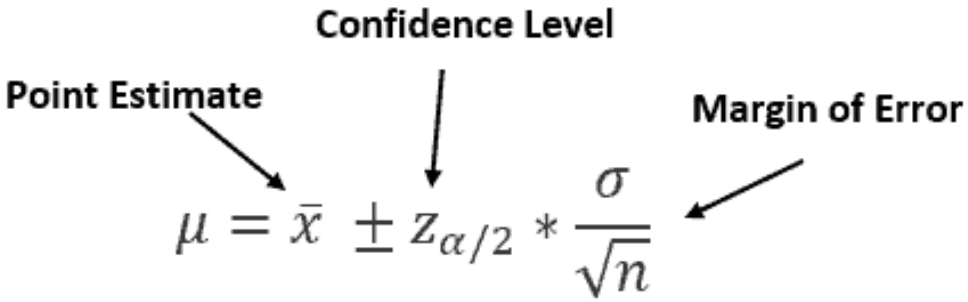
- An interval estimate gives you a range of values where the parameter is expected to lie.
- A confidence interval estimate is a range of values constructed from sample data so that the population parameter will likely occur within the range at a specified probability. Accordingly, the specified probability is the level of confidence.

- Broader and probably more accurate than a point estimate
- Any parameter estimate that is based on a sample statistic has some amount of sampling error.



Point Estimate and Interval Estimate

A Confidence interval is used to express the precision and ambiguity of a particular sampling method.



- A confidence *interval* is a range of values that probably contain the population mean.
- A Confidence *level* is a percentage of certainty that, in any given sample, that confidence interval will contain the population means.
- The Point estimate is a statistic (value from a sample) used to estimate a parameter (value from the population).
- The margin of error is the maximum expected difference between the actual population parameter and a sample estimate of the parameter. In other words, it is the range of values above and below sample statistics.

Statistics

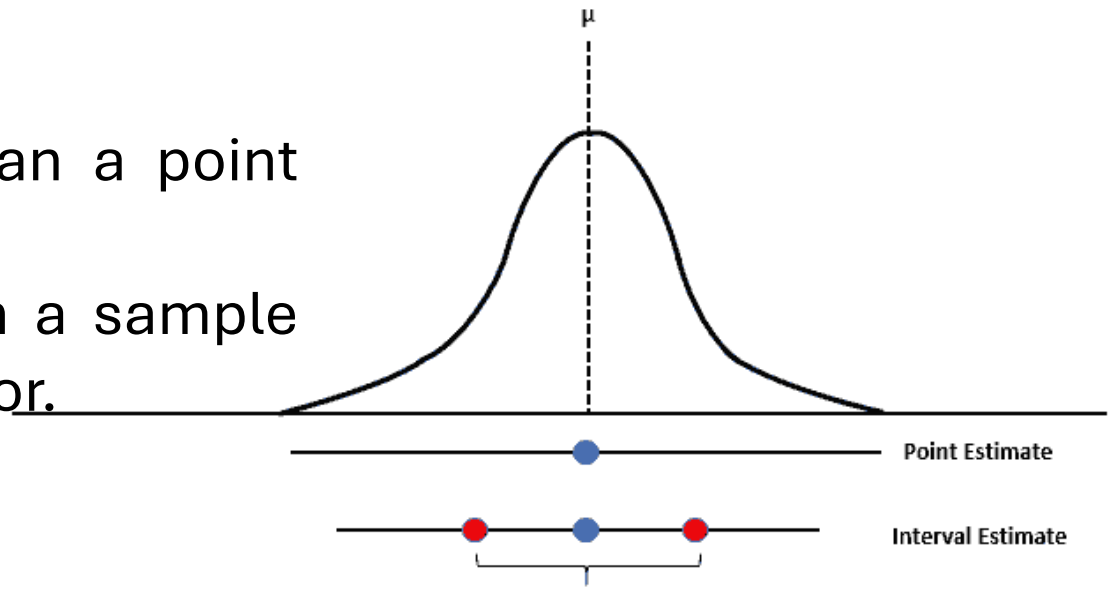
Point Estimate and Interval Estimate

- A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- A point estimate is a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter. In other words, we derive the point estimate from a single value in the sample and use it to estimate the population value.
- Take a sample, find $\bar{x}$. It is a close approximation of μ. But, depending on your sample size, that may not be a good point estimate.
- In fact, the probability that a single sample statistic is equal to the population parameter is very unlikely.

Point Estimate and Interval Estimate

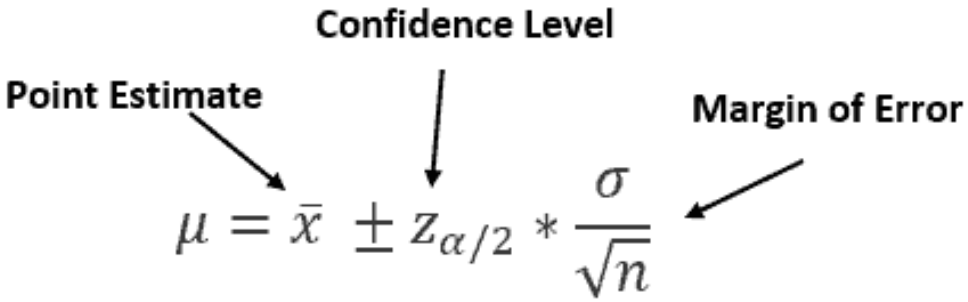
- An interval estimate gives you a range of values where the parameter is expected to lie.
- A confidence interval estimate is a range of values constructed from sample data so that the population parameter will likely occur within the range at a specified probability. Accordingly, the specified probability is the level of confidence.

- Broader and probably more accurate than a point estimate
- Any parameter estimate that is based on a sample statistic has some amount of sampling error.



Point Estimate and Interval Estimate

A Confidence interval is used to express the precision and ambiguity of a particular sampling method.



- A confidence *interval* is a range of values that probably contain the population mean.
- A Confidence *level* is a percentage of certainty that, in any given sample, that confidence interval will contain the population means.
- The Point estimate is a statistic (value from a sample) used to estimate a parameter (value from the population).
- The margin of error is the maximum expected difference between the actual population parameter and a sample estimate of the parameter. In other words, it is the range of values above and below sample statistics.

Point Estimation

- Here, we assume that θ is an unknown parameter to be estimated.
- For example, θ might be the expected value of a random variable, θ=EX. Θ is a fixed (non-random) quantity.
- To estimate θ, we define a point estimator $\hat{\Theta}$ that is a function of the random sample, i.e.,

$$\hat{\Theta} = h(X_1, X_2, \cdots, X_n).$$

For example, if $\theta = EX$, we may choose $\hat{\Theta}$ to be the sample mean

$$\hat{\Theta} = \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Point Estimation

Mean ($\bar{x}$) → Estimates Population Mean (μ)

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Variance (s^2) → Estimates Population Variance (σ^2)

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Standard Deviation (s) → Estimates Population Standard Deviation (σ)

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

Proportion (p-hat) → Estimates Population Proportion (p)

$$\hat{p} = \frac{x}{n}$$

Properties of Estimators

- Estimators should be **unbiased**.
  - expected value equals the true parameter value.
- The estimator should be **efficient.**
  - it has the **lowest variance** among all unbiased estimators of a parameter.
- An estimator should be **consistent.**
  - as the sample size increases, the estimated value gets closer to the true population parameter.
  - More data improves the accuracy of estimation.

Evaluating Estimators

- Three main desirable properties for point estimators

1. The **bias** of an estimator $\hat{\Theta}$ tells us on average how far $\hat{\Theta}$ is from the real value of θ.

i. Unbiasedness

- An estimator **θˆ** is **unbiased** if its expected value is equal to the true population parameter (θ):

$$E(\hat{\theta}) = \theta$$

- Example: The sample mean $\bar{x}$ is an unbiased estimator of the population mean **μ**:

$$E(\bar{X}) = \mu$$

ii. *consistency*

- An estimator **θˆ** is **consistent** if it gets **closer to the true parameter (θ) as the sample size (n) increases**.

Let $\hat{\Theta}_1, \hat{\Theta}_2, \cdots, \hat{\Theta}_n, \cdots$, be a sequence of point estimators of θ. We say that $\hat{\Theta}_n$ is a **consistent** estimator of θ, if

$$\lim_{n\to\infty} P\big(|\hat{\Theta}_n - \theta| \geq \epsilon\big) = 0, \text{ for all } \epsilon > 0.$$

- Example: The **sample mean ($\bar{x}$)** is a consistent estimator of **μ** because as we take larger samples, it converges to **μ**.

iii. Efficiency

- Among multiple unbiased estimators, the **most efficient** estimator has the **smallest variance**.

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2) \Rightarrow \hat{\theta}_1 \text{ is more efficient}$$

Example: If we have two estimators of **μ**, the one with the smaller variance is preferred.

Estimator 1: $Var(\hat{\theta}_1) = 5$

Estimator 2: $Var(\hat{\theta}_2) = 2$

Hypothesis Testing

It refers to

- Making an assumption, called hypothesis, about a population parameter.
- Collecting sample data and calculating sample statistic.
- Using the sample statistic to evaluate the hypothesis (how likely is it that our hypothesized parameter is correct).
- To test the validity of our assumption we determine the difference between the hypothesized parameter value and the sample value.

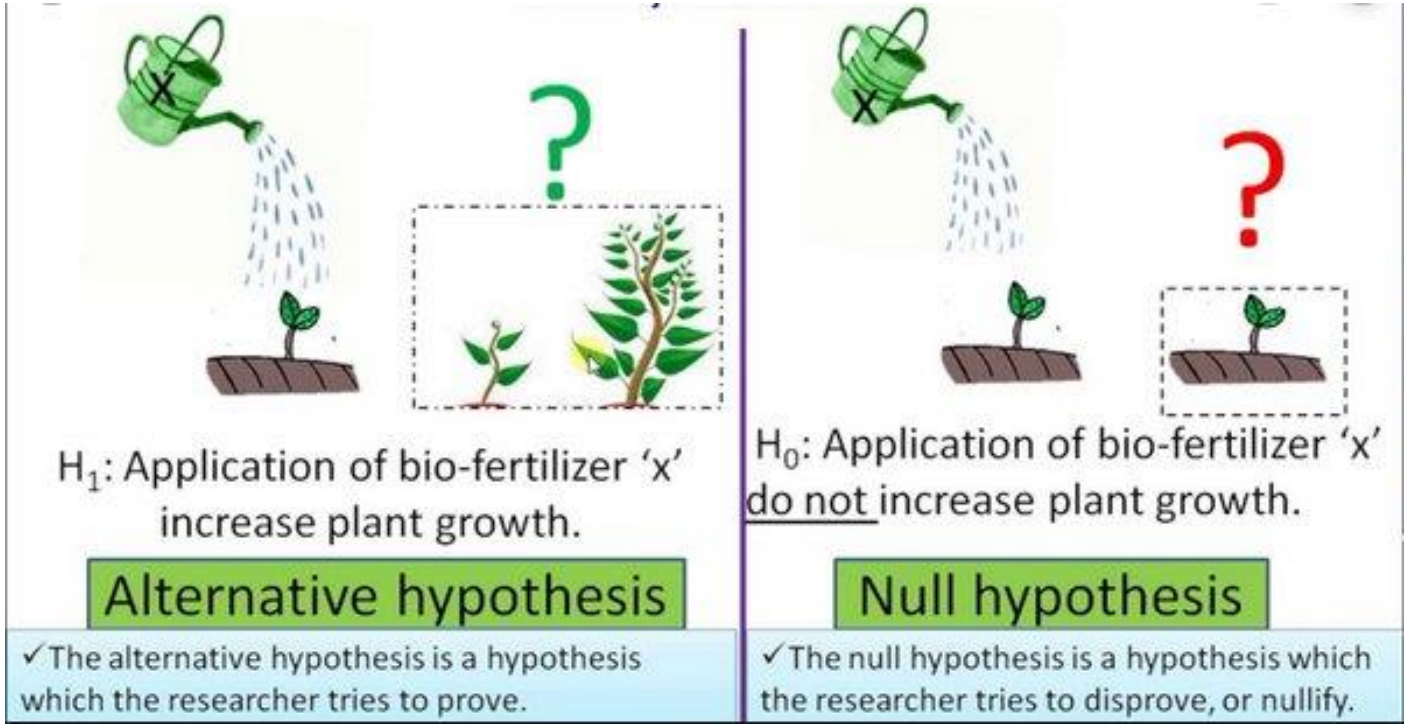
Hypothesis Testing

Example:

- A pharmaceutical company might be interested in knowing if a new drug is effective in treating a disease. Here, there are two hypotheses.
- The first one is that the drug is not effective (hypotheses H0), while the second hypothesis is that the drug is effective (hypotheses H1).
- The hypothesis H0 is called the **null hypothesis** and the hypothesis H1 is called the **alternative hypothesis**.

Hypothesis Testing

The null hypothesis represents the default assumption that no significant difference or relationship exists between the studied variables. In contrast, the alternative hypothesis represents the claim or hypothesis the researcher is testing.



Hypothesis Testing

You have a coin and you would like to check whether it is fair or not. More specifically, let θ be the probability of heads, $\theta = P(H)$. You have two hypotheses:

H_0 (the null hypothesis): The coin is fair, i.e. $\theta = \theta_0 = \frac{1}{2}$.

H_1 (the alternative hypothesis): The coin is not fair, i.e., $\theta \neq \frac{1}{2}$.

We need to design a test to either accept H_0 or H_1. To check whether the coin is fair or not, we perform the following experiment. We toss the coin 100 times and record the number of heads. Let X be the number of heads that we observe, so

$$X \sim Binomial(100, \theta).$$

Now, if H_0 is true, then $\theta = \theta_0 = \frac{1}{2}$, so we expect the number of heads to be close to 50. Thus, intuitively we can say that if we observe close to 50 heads we should accept H_0, otherwise we should reject it. More specifically, we suggest the following criteria: If $|X - 50|$ is less than or equal to some threshold, we accept H_0. On the other hand, if $|X - 50|$ is larger than the threshold we reject H_0 and accept H_1. Let's call that threshold t.

If $|X - 50| \leq t$, accept H_0.

If $|X - 50| > t$, accept H_1.

Hypothesis Testing

- **Level of significance**: It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so we select a level of significance. This is normally denoted with alpha(α) and generally, it is 0.05 or 5% which means your output should be 95% confident to give a similar kind of result in each sample.
- **Test Statistic:** Test statistic is the number that helps you decide whether your result is significant. It's calculated from the sample data you collect it could be used to test if a machine learning model performs better than a random guess.
- **Critical value:** Critical value is a boundary or threshold that helps you decide if your test statistic is enough to reject the null hypothesis

Hypothesis Testing

- **P-value**: The p-value is the probability of observing a test statistic given that the null hypothesis is true.
  - A **small p-value** usually less than 0.05 means the results are unlikely to be due to random chance so we reject the null hypothesis.
  - A **large p-value** means the results could easily happen by chance so we don't reject the null hypothesis.

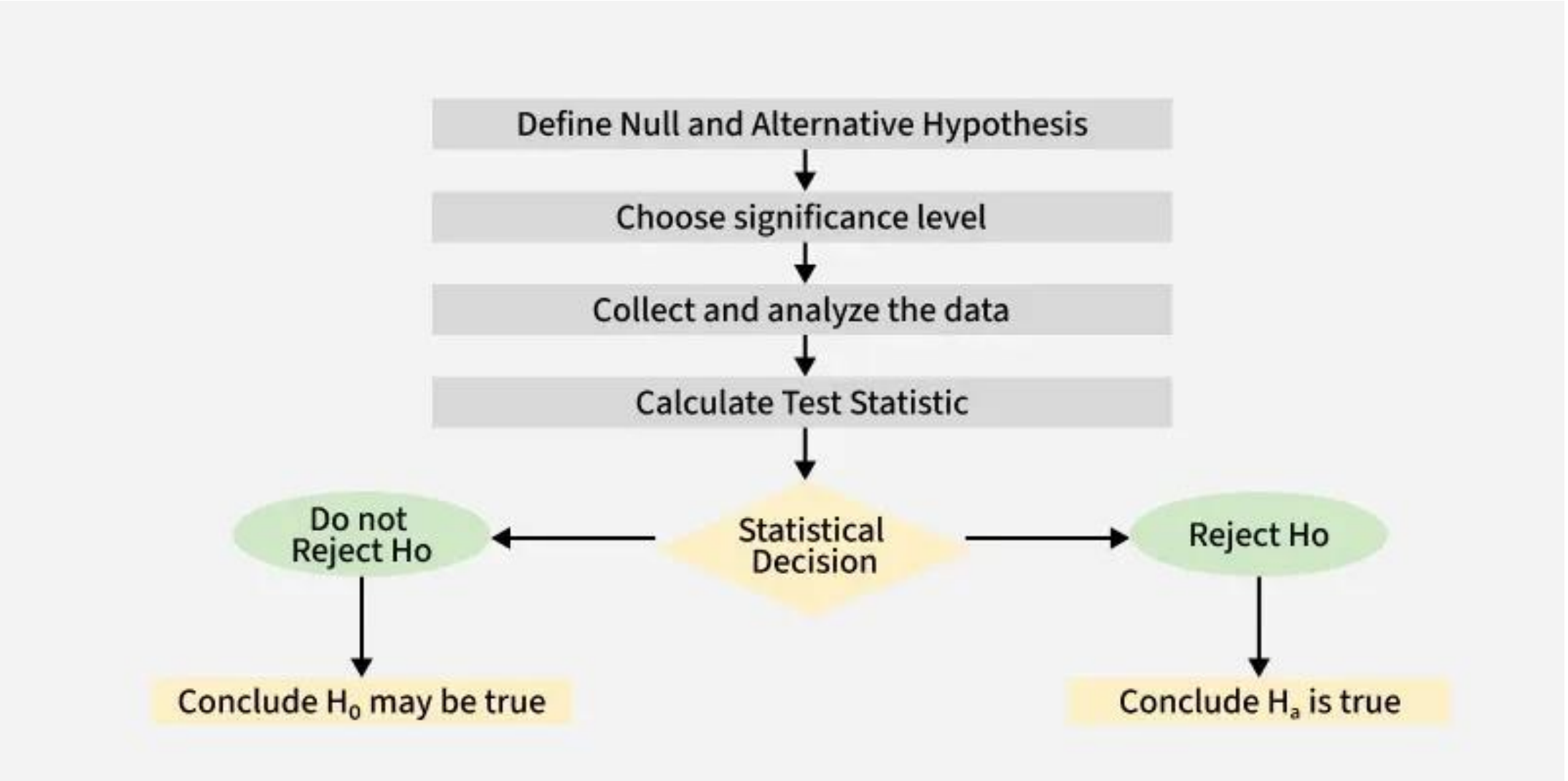
Hypothesis Testing

In hypothesis testing Type I and Type II errors are two possible errors that can happen when we are finding conclusions about a population based on a sample of data. These errors are associated with the decisions we made regarding the null hypothesis and the alternative hypothesis.

- **Type I error:** When we reject the null hypothesis although that hypothesis was true. Type I error is denoted by alpha(α).
- **Type II errors:** When we accept the null hypothesis, but it is false. Type II errors are denoted by beta(β).

	Null Hypothesis is True	**Null Hypothesis is False**
Null Hypothesis is True (Accept)	Correct Decision	Type II Error (False Negative)
Alternative Hypothesis is True (Reject)	Type I Error (False Positive)	Correct Decision

Hypothesis Testing



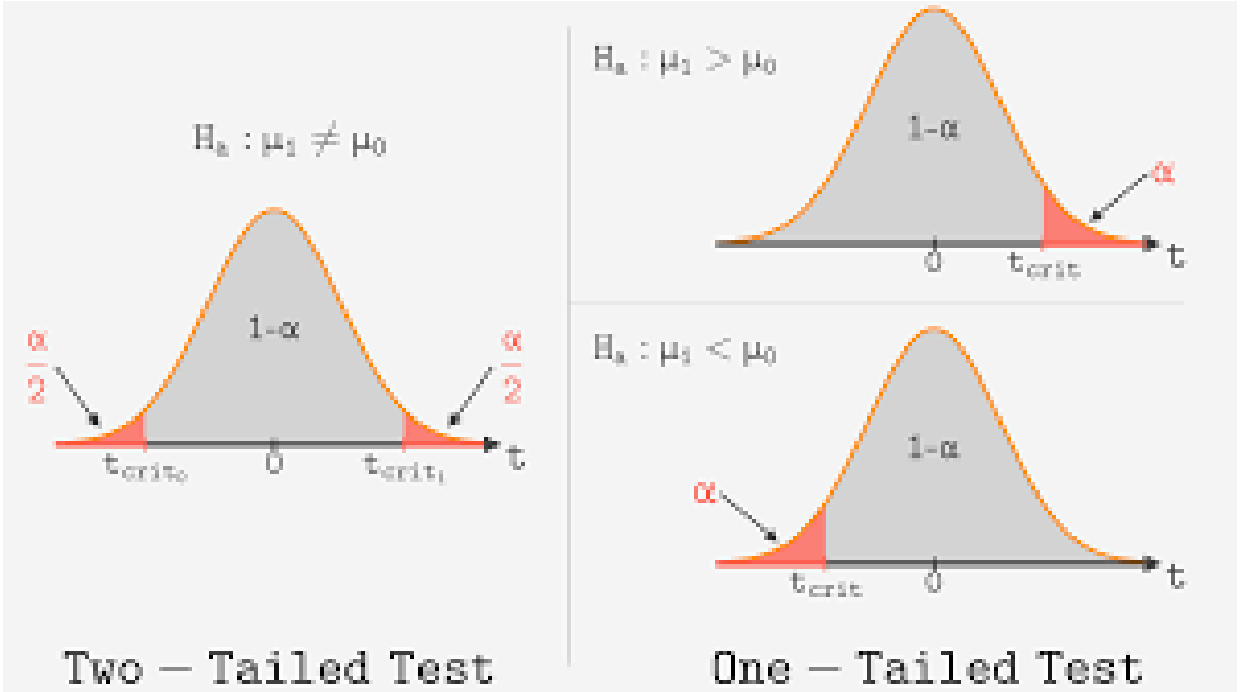
One-Sided (One-Tailed) Tests

- **Greater Than Test (Right-Tailed Test):** This type of one-sided test is used when you want to determine if a parameter or effect is greater than a specified value.
  - **Null Hypothesis (H0):** The parameter is less than or equal to the specified value.
  - **Alternative Hypothesis (H1 or Ha):** The parameter is greater than the specified value.
  - ***Example:*** Testing if a new drug improves patient recovery time **H0:** The drug does not improve recovery time. **Ha:** The drug improves recovery time.
- **Less Than Test (Left-Tailed Test):** This one-sided test is used when you want to determine if a parameter or effect is less than a specified value.
  - **Null Hypothesis (H0):** The parameter is greater than or equal to the specified value.
  - **Alternative Hypothesis (H1 or Ha):** The parameter is less than the specified value.
  - ***Example:*** Testing if a manufacturing process meets quality standards. **H0:** The process meets quality standards. **Ha:** The process does not meet quality standards.

Two-Sided (Two-Tailed) Tests

Two-Sided Test: This type of test is used when you want to determine if a parameter or effect is significantly different from a specified value, without specifying whether it's greater or less than that value.

- **Null Hypothesis (H0):** The parameter is equal to the specified value.
- **Alternative Hypothesis (H1 or Ha):** The parameter is not equal to the specified value.
- ***Example:*** Testing if a coin is fair (i.e., equally likely to land heads or tails). **H0:** The coin is fair. **Ha:** The coin is not fair.



Statistics

z-Statistics

It is used when population means and standard deviations are known. The formula of z-statistics is given by:

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

where

- $\bar{x}$ is the sample mean,
- μ represents the population mean,
- σ is the standard deviation
- and n is the size of the sample.

z-Statistics

Q1. A company manufactures light bulbs and claims that the average lifespan of their bulbs is 1000 hours. A consumer group wants to test this claim. They randomly sample 64 bulbs and find that the sample mean lifespan is 980 hours. Assume the population standard deviation is known to be 100 hours.

1. **State the hypotheses:**

- **Null hypothesis (H0):** The average lifespan of the bulbs is 1000 hours. ($\mu = 1000$)
- **Alternative hypothesis (Ha):** The average lifespan of the bulbs is not 1000 hours. ($\mu \neq 1000$)

2. **Determine the level of significance:**

- Let's choose a significance level of 0.05 (alpha = 0.05). This means we are willing to accept a 5% chance of rejecting the null hypothesis when it is actually true.

3. **Calculate the test statistic:**

- The formula for the z-test statistic is: $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$
  - Where:
    - $\bar{x}$ is the sample mean (980 hours)
    - μ is the population mean under the null hypothesis (1000 hours)
    - σ is the population standard deviation (100 hours)
    - n is the sample size (64)
- Plugging in the values: $z = (980 - 1000) / (100 / \sqrt{64}) = -1.6$

4.Find the critical value:

• This is a two-tailed test (Ha: $\mu \neq 1000$)

• For a two-tailed test with alpha = 0.05, the critical values are approximately ±1.96.

5.Make a decision:

• Compare test statistic (-1.6) to the critical values (±1.96).

•Since the absolute value of the test statistic (1.6) is less than the critical value (1.96), we fail to reject the null hypothesis.

Conclusion:

At a 0.05 significance level, there is not enough evidence to conclude that the average lifespan of the light bulbs is different from 1000 hours.

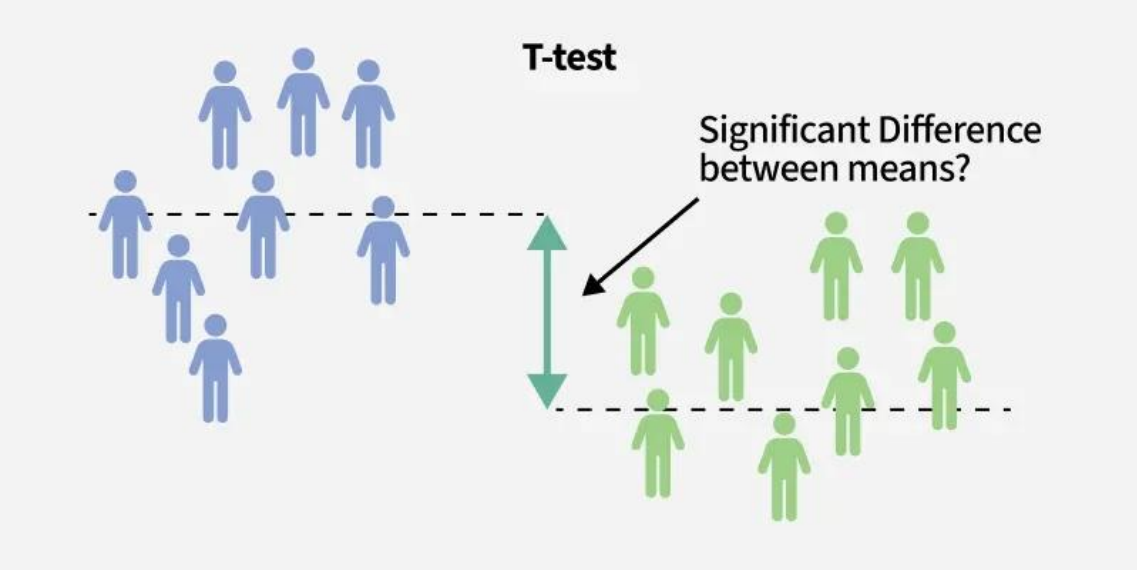
T-Statistics

T-test is used to compare the means of two datasets (e.g., experimental vs. control groups) to assess if the difference is statistically significant. It is used when n<30 t-statistic calculation is given by:

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

where

- t = t-score,
- $\bar{x}$ = sample mean
- μ = population mean,
- s = standard deviation of the sample,
- n = sample size



T-Statistics

Suppose You want to compare the test scores of two groups of students:

•Group 1: 30 students who studied with Method A.

•Group 2: 30 students who studied with Method B.

You use a **t-test** to check if there is a significant difference in the average test scores between the two.

The t-test is part of **hypothesis testing** where you start with an assumption the null hypothesis that the two-group means are the same. Then the test helps you decide if there's enough evidence to reject that assumption and conclude that the groups are different.

T-Statistics

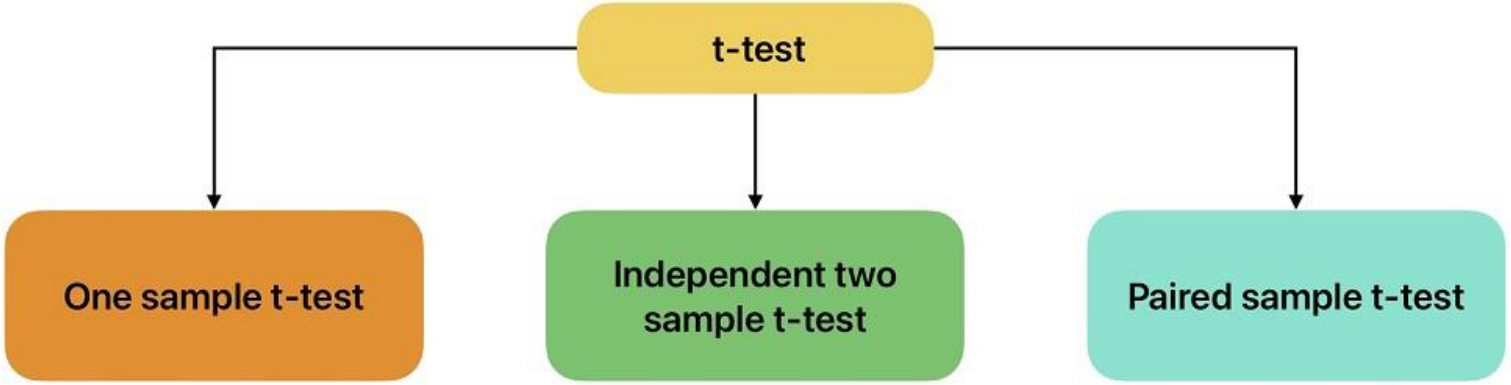
- **Degree of freedom (df):** The degree of freedom tells us the number of independent variables used for calculating the estimate between 2 sample groups.

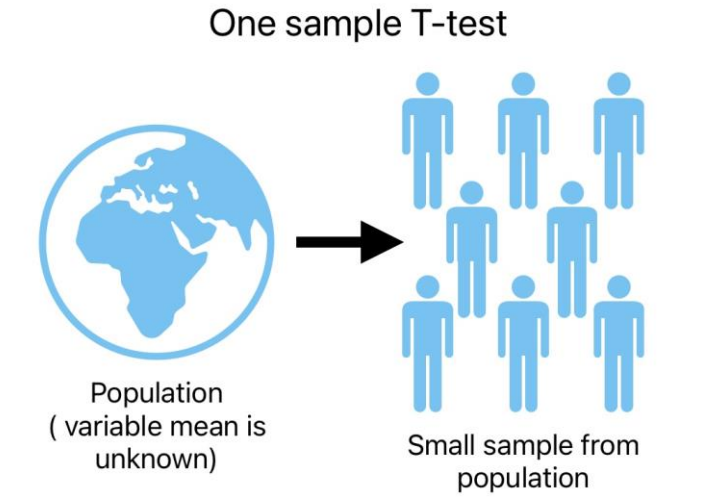
  In a t-test the degree of freedom is calculated as the total sample size minus 1 i.e. df=∑ns−1, where "n_s" is the number of observations in the sample. Suppose, we have 2 samples A and B. The df would be calculated as df=(nA−1)+(nB−1)

- **Significance Level:** The significance level is the predetermined threshold that is used to decide whether to reject the null hypothesis. Commonly used significance levels are 0.05, 0.01, or 0.10.

- **T-statistic**: The t-statistic is a measure of the difference between the means of two groups. It is calculated as the difference between the sample means divided by the standard error of the difference. It is also known as the t-value or t-score.
  - If the t-value is large => the two groups belong to different groups.
  - If the t-value is small => the two groups belong to the same group.

T-Statistics





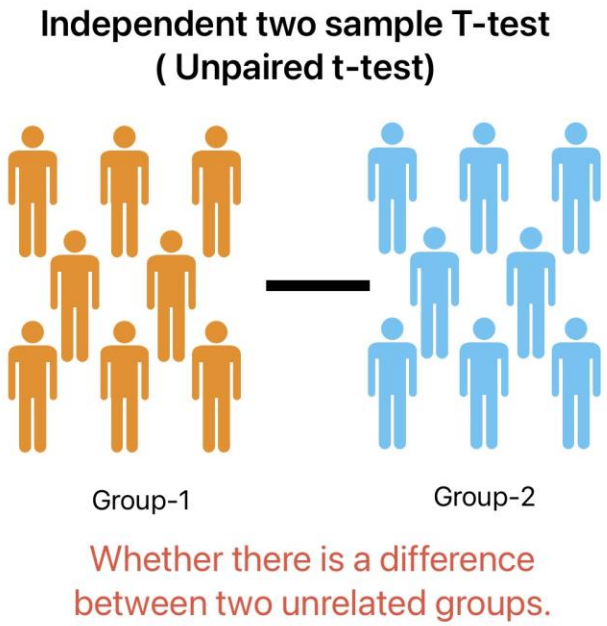
One Sample T-Test

The value against which we are comparing is a single value, i.e. we compare the mean of a sample with a single value to check how much the mean deviates from that single value.

Two Sample T-Test

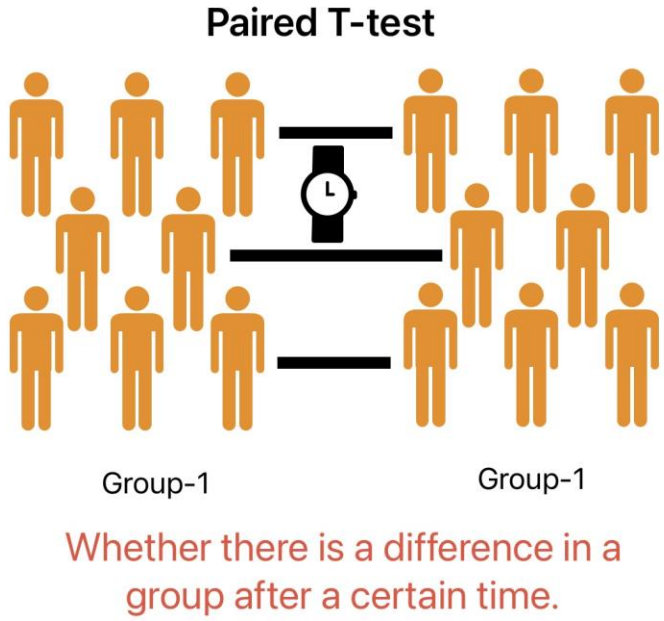
We compare the means and variances of two samples, we assess how much they differ.

T-Statistics



In an **independent two-sample t-test** (unpaired t-test), the samples in the two groups being compared are unrelated. The samples are drawn from two different populations or groups of subjects, and the difference between the means of the two groups is calculated using the means and variances of the two separate samples.

In a **dependent two-sample t-test** (also known as a **paired t-test**), the samples in the two groups being compared are related in some way. For example, the samples may be pairs of measurements taken on the same subjects. In this case, the difference between the means of the two groups is calculated by taking the differences between the pairs of measurements and treating these differences as a single sample.



T-Statistics

One-Sample T-Test

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

$\bar{x}$ = obersved mean of the sample
μ = assumed mean
s = standard deviation
n = sample size

Two-Sample T-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ = observed mean of 1st sample
$\bar{x}_2$ = observed mean of 2nd sample
s_1 = standard deviation of 1st sample
s_2 = standard deviation of 2nd sample
n_1 = sample size of 1st sample
n_2 = sample size of 2nd sample

Paired sample T-test

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where: $sd = \text{sqrt}[\ \Sigma(di - \bar{d})^2 / (n - 1)\ ]$

- $\bar{d}$ is the mean of the difference scores
- s_d is the standard deviation of the difference scores
- n is the number of pairs of observations

di: The difference between the paired measurements for the i-th participant (After - Before for the i-th participant)
$\bar{d}$: The mean difference (average of all the di's)

T-Statistics

Q1. A manufacturer claims that the average weight of their product is 50 grams. You want to test this claim. You randomly sample 25 products and find the following weights (in grams):

48, 52, 49, 51, 50, 47, 53, 50, 49, 52, 48, 51, 50, 49, 50, 51, 48, 52, 49, 50, 51, 47, 53, 50, 49

1. **State the hypotheses:**
   - Null hypothesis (H0): The population mean is equal to 50 grams ($\mu = 50$).
   - Alternative hypothesis (Ha): The population mean is not equal to 50 grams ($\mu \neq 50$).
2. **Calculate the sample mean ($\bar{x}$) and sample standard deviation (s):**
   - $\bar{x}$ = (sum of all weights) / (number of samples) = 1245 / 25 = 49.8 grams
   - s = (square root of [sum of (each weight - $\bar{x}$)^2] / (number of samples - 1)) ≈ 1.83 grams
3. **Calculate the t-statistic:**
   - $t = (\bar{x} - \mu) / (s / \sqrt{n}) = (49.8 - 50) / (1.83 / \sqrt{25}) \approx -0.55$
4. **Determine the degrees of freedom (df):**
   - df = n - 1 = 25 - 1 = 24
5. **Find the critical value:**
   - You need to choose a significance level (alpha). Let's say $\alpha = 0.05$.
   - Consult a t-distribution table or use a calculator to find the critical value for a two-tailed test with df = 24 and $\alpha = 0.05$. The critical value is approximately ±2.064.

Since our calculated t-statistic (-0.55) falls within the range of -2.064 to +2.064, we fail to reject the null hypothesis.

Conclusion: There is not enough evidence to conclude that the average weight of the products is different from 50 grams.

t-test table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

T-Statistics

Q2. A researcher wants to know whether there is a significant difference in the average test scores of students who are taught using two different methods. The researcher randomly assigns 20 students to one of two groups. Group A is taught using method A, and group B is taught using method B. After the students have completed the course, they are given a test. The test scores are shown below:

Group A: 85, 90, 92, 88, 89, 91, 93, 87, 86, 94

Group B: 78, 82, 80, 85, 83, 81, 79, 84, 77, 86

Solution:

- t-test is used to determine whether there is a significant difference in the average test scores of the two groups.
- The null hypothesis is that there is no significant difference in the average test scores of the two groups.
- The alternative hypothesis is that there is a significant difference in the average test scores of the two groups.
- Calculates the mean and standard deviation of the test scores for each group.

Group A: mean = 89.5, standard deviation = 2.87

Group B: mean = 81.5, standard deviation = 2.87

T-Statistics

- The null hypothesis is that there is no significant difference in the average test scores of the two groups.
- The alternative hypothesis is that there is a significant difference in the average test scores of the two groups.
- Calculates the mean and standard deviation of the test scores for each group.

Group A: mean = 89.5, standard deviation = 2.87 **Group B:** mean = 81.5, standard deviation = 2.87

- Calculate t-score using two sample t-test formula:

t = (89.5 - 81.5) / sqrt((2.87^2 / 10) + (2.87^2 / 10)) = 5.57

- Degree of freedom df = =(nA−1)+(nB−1) = 18

Let, the significance level is 0.05, the critical value is obtained from the t-table as 2.101.

Since the t-statistic (5.57) is greater than the critical value (2.101), the researcher rejects the null hypothesis. This means that there is a significant difference in the average test scores of the two groups.

t-test table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

T-Statistics

Q3. A researcher wants to study the effectiveness of a new memory-enhancing drug. They recruit 20 participants and administer a memory test before and after taking the drug for a month. Is there a statistically significant difference in memory test scores before and after taking the drug?

Solution: this is a paired t-test question, as we have two measurements (before and after) for each participant.

1.Calculate the Differences: Subtract the "Before Drug" score from the "After Drug" score for each participant.

2.Calculate the Mean Difference ($\bar{d}$): Sum the differences and divide by the number of participants (n = 20).

$\bar{d}$ = (7+7+5+6+5+2+7+5+5+5+7+5+6+6+4+4+7+5+5+6) / 20 = 5.35

3.Calculate the Standard Deviation of the Differences (sd):

sd = sqrt[$\Sigma(di - \bar{d})^2$ / (n - 1)] **= 1.35**

4. Calculate the t-statistic: t = ($\bar{d}$) / (sd / $\sqrt{n}$) = 5.35 / (1.35 / $\sqrt{20}$) ≈ 17.7

5. Determine the Degrees of Freedom: df = n - 1 = 20 - 1 = 19

6. Find the Critical Value: for alpha 0.05 and df =19, critical value is ±2.093.

Since the absolute value of our t-statistic is greater than the critical value, we reject the null hypothesis.

Participant	Before Drug	After Drug	Difference (After - Before)
1	75	82	7
2	68	75	7
3	80	85	5
4	72	78	6
5	85	90	5
6	70	72	2
7	78	85	7
8	65	70	5
9	90	95	5
10	75	80	5
11	72	79	7
12	68	73	5
13	82	88	6
14	77	83	6
15	88	92	4
16	71	75	4
17	79	86	7
18	66	71	5
19	84	89	5
20	73	79	6

Statistics

Chi-Square test

In recent years, the use of specialized statistical methods for categorical data has increased dramatically, particularly for applications in the biomedical and social sciences. Categorical scales occur frequently in the health sciences, for measuring responses. E.g.

- patient survives an operation (yes, no),
- severity of an injury (none, mild, moderate, severe), and
- stage of a disease (initial, advanced).

Studies often collect data on categorical variables that can be summarized as a series of counts and commonly arranged in a tabular format known as a **contingency table.**

Chi-Square test χ^2

The most obvious difference between the chi-square tests and the other hypothesis tests we have considered (T test) is the *nature of the data (categorical data).*

- For chi-square, the data are ***frequencies*** rather than numerical scores.
- Used for testing significance of patterns in qualitative data.
- Test statistic is based on counts (frequencies) that represent the number of items that fall in each category
- Test statistics measures the agreement between actual counts(observed) and expected counts assuming the null hypothesis

Chi-Square test χ^2

Chi-square Test – Distribution table and formulas

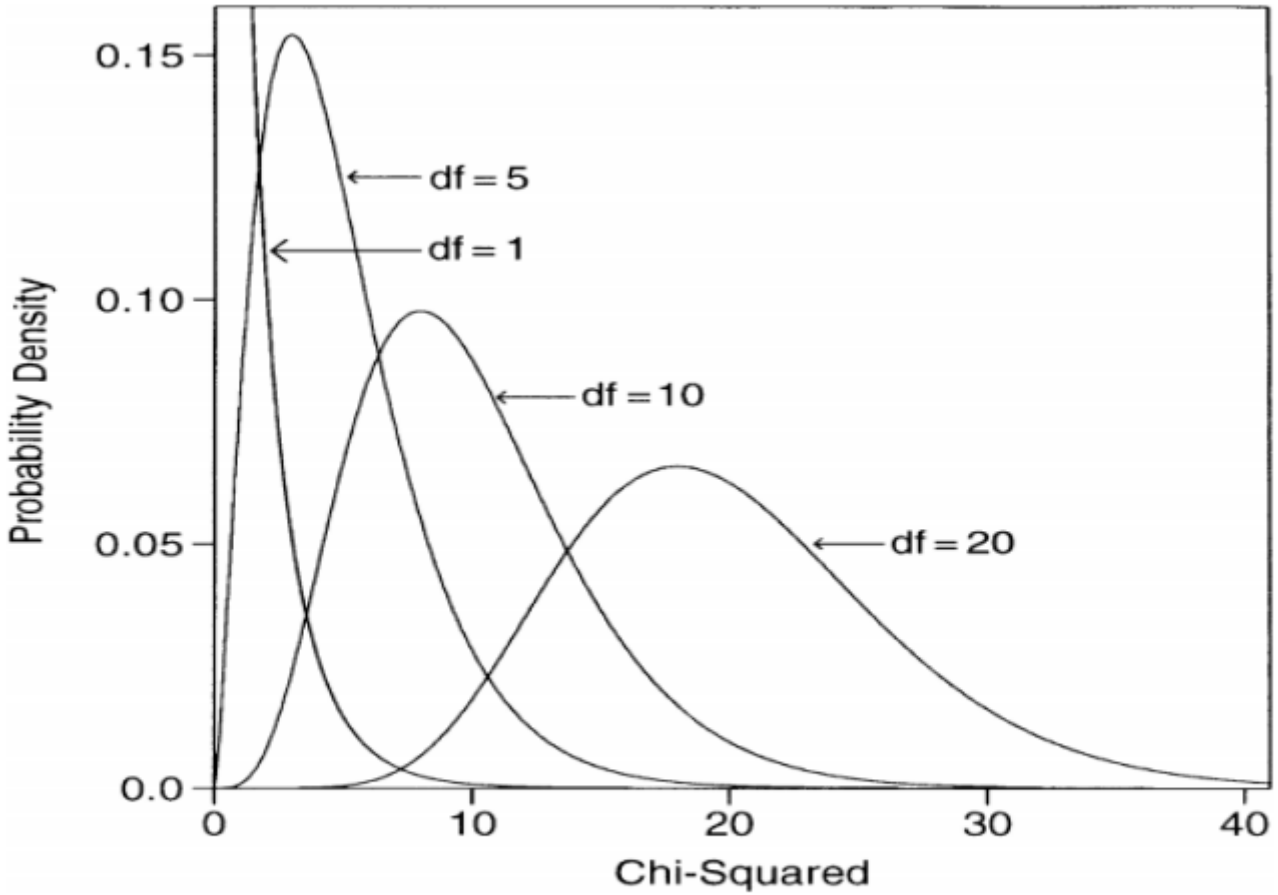
Degrees of Freedom (df) Significance Level (α)	0.01	0.05	0.10	0.25	0.50
1	6.635	3.841	2.706	1.323	0.454
2	9.210	5.991	4.605	2.773	1.386
3	11.345	7.815	6.251	3.930	2.366
4	13.277	9.488	7.779	5.178	3.357
5	15.086	11.070	9.236	6.571	4.351
6	16.812	12.592	10.645	7.962	5.348
7	18.475	14.067	12.017	9.364	6.346
8	20.090	15.507	13.362	10.773	7.344
9	21.666	16.919	14.684	12.189	8.343
10	23.209	18.307	15.987	13.603	9.342

$$\chi^2=\sum\frac{(O_i-E_i)^2}{E_i}$$

$$df=(r-1)\times(c-1)$$

$$E_i = \frac{Row\ Total\ \times Column\ Total}{Grand\ Total}$$

Chi-Square test χ^2



- The degrees of freedom for tests of hypothesis that involve an rxc contingency table is **equal to (r-1)x(c-1);**

Chi-Square test χ^2

Application of chi square test

1. **Goodness-of-fit:** uses frequency data from a sample to test hypotheses about the shape or proportions of a ***population***.
2. **Test for independence**:
    1. (2×2 chi-square test): Testing hypotheses about the relationship between two variables in a population,
    2. (a x b chi-square test) or (r x c chi-square test)

Chi-Square test χ^2

Q1. Given Eye colour in a sample of 40 people: Blue 12, brown 21, green 3, others 4
Given Eye colour in population: Brown 80%, Blue 10%, Green 2%, Others 8%
Is there any difference between proportion of sample to that of population (use alpha= 0.05)

Solution: Assume Sample is randomly selected from the population.
Null hypothesis: there is no significant difference in proportion of eye colour of sample to that of the population.
Alternative hypothesis: there is significant difference in proportion of eye colour of sample to that of the population.

$$\chi^2 = \frac{(12-4)^2}{4} + \frac{(21-32)^2}{32} + \frac{(3-0.8)^2}{0.8} + \frac{(4-3)^2}{3}$$

$$=(64/4) + (121/32)+(4.8/0.8)+(1/3)$$
$$=16+3.78+6+0.3$$
$$=26.08$$

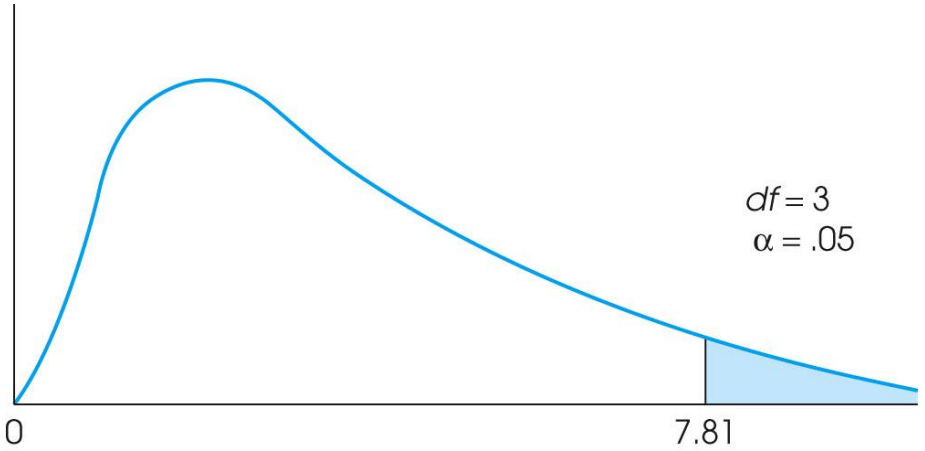
Color	Sample frequency	Expected frequency
Blue	12	40*10/100= 4
Brown	21	40*80/100=32
Green	3	40*2/100=0.8
Others	4	40*8/100=3

Chi-Square test χ^2

Q1. Given Eye colour in a sample of 40 people: Blue 12, brown 21, green 3, others 4
Given Eye colour in population: Brown 80%, Blue 10%, Green 2%, Others 8%
Is there any difference between proportion of sample to that of population (use alpha= 0.05)

Solution: Assume Sample is randomly selected from the population.
Null hypothesis: there is no significant difference in proportion of eye colour of sample to that of the population.
Alternative hypothesis: there is significant difference in proportion of eye colour of sample to that of the population.

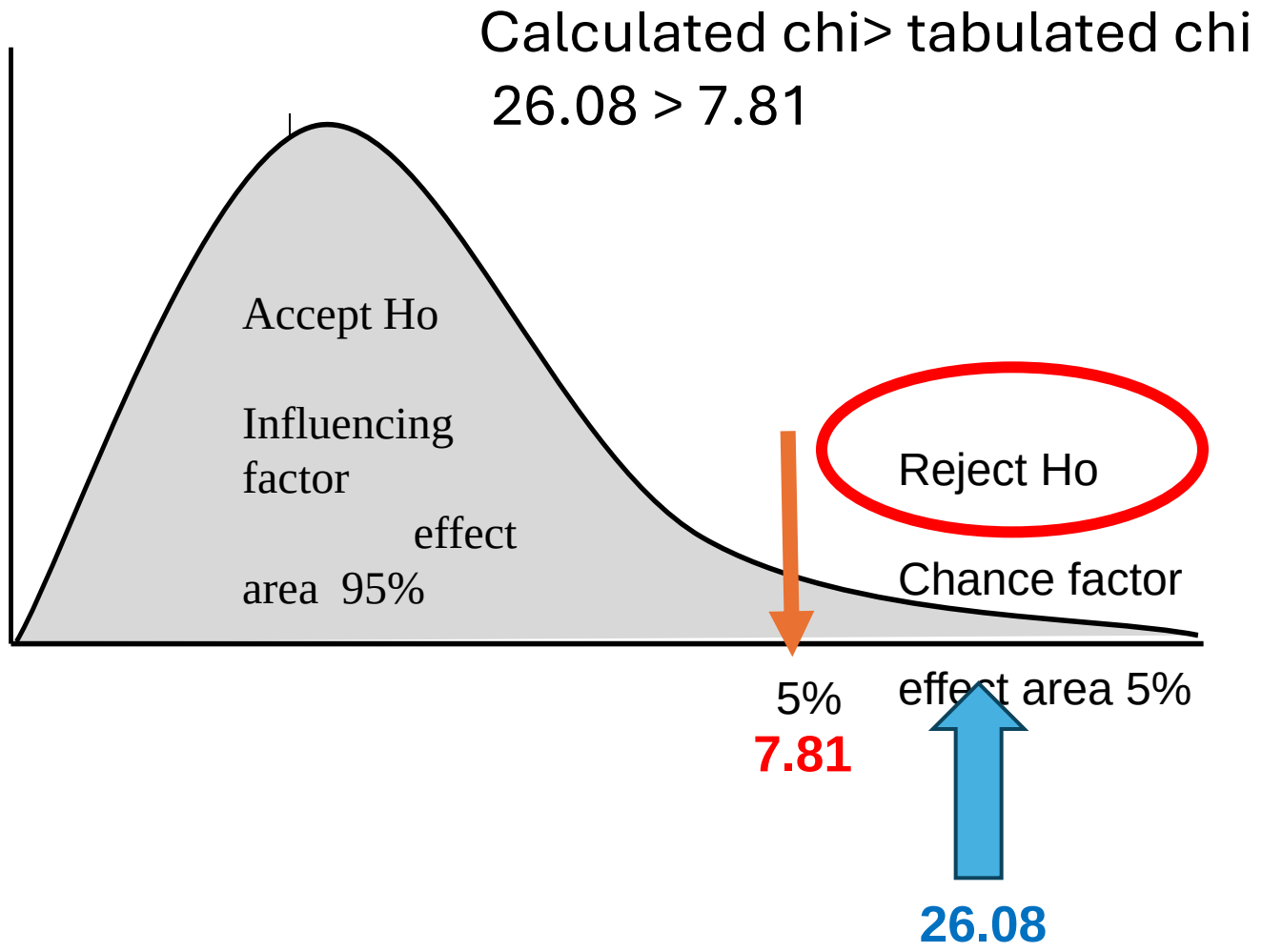
α =0.05
d.f.(degree of freedom)=K-1 = 4-1 =3
(K=Number of subgroups)
critical value for α =0.5 and df=3 => 7.81



Chi-Square test χ^2

Conclusion: We reject H0 &accept HA

There is significant difference in **<u>proportion</u>** of eye colour of sample to that of the population.



Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let α 0.05

HO: there is no significant **association** between type of operator & deafness.
HA:there is significant **association** between type of operator & deafness.

α =0.05
d.f.(degree of freedom)=(2-1)(2-1) = 1
critical value for α =0.5 and df=1 => 3.841

Operator	deaf	Not deaf.	total
A	100	900	1000
B	60	440	500
total	160	1340	1500

Total number of items=1500

Total number of defective items=160

Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let α 0.05

Expected deaf from Operator A = 1000 * 160/1500 = 106.7

(expected not deaf=1000-106.7=893.3)

Expected deaf from Operator B = 500 * 160/1500 = 53.3

$$\chi^2 = \frac{(100-106.7)^2}{106.7} + \frac{(900-893.3)^2}{893.3} + \frac{(60-53.3)^2}{53.3} + \frac{(440-446.7)^2}{446.7}$$

$$= 0.42+0.05+o.84+0.10 = 1.41$$

Operator	deaf	Not deaf.	total
A	100	900	1000
B	60	440	500
total	160	1340	1500

Total number of items=1500

Total number of defective items=160

Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let α 0.05

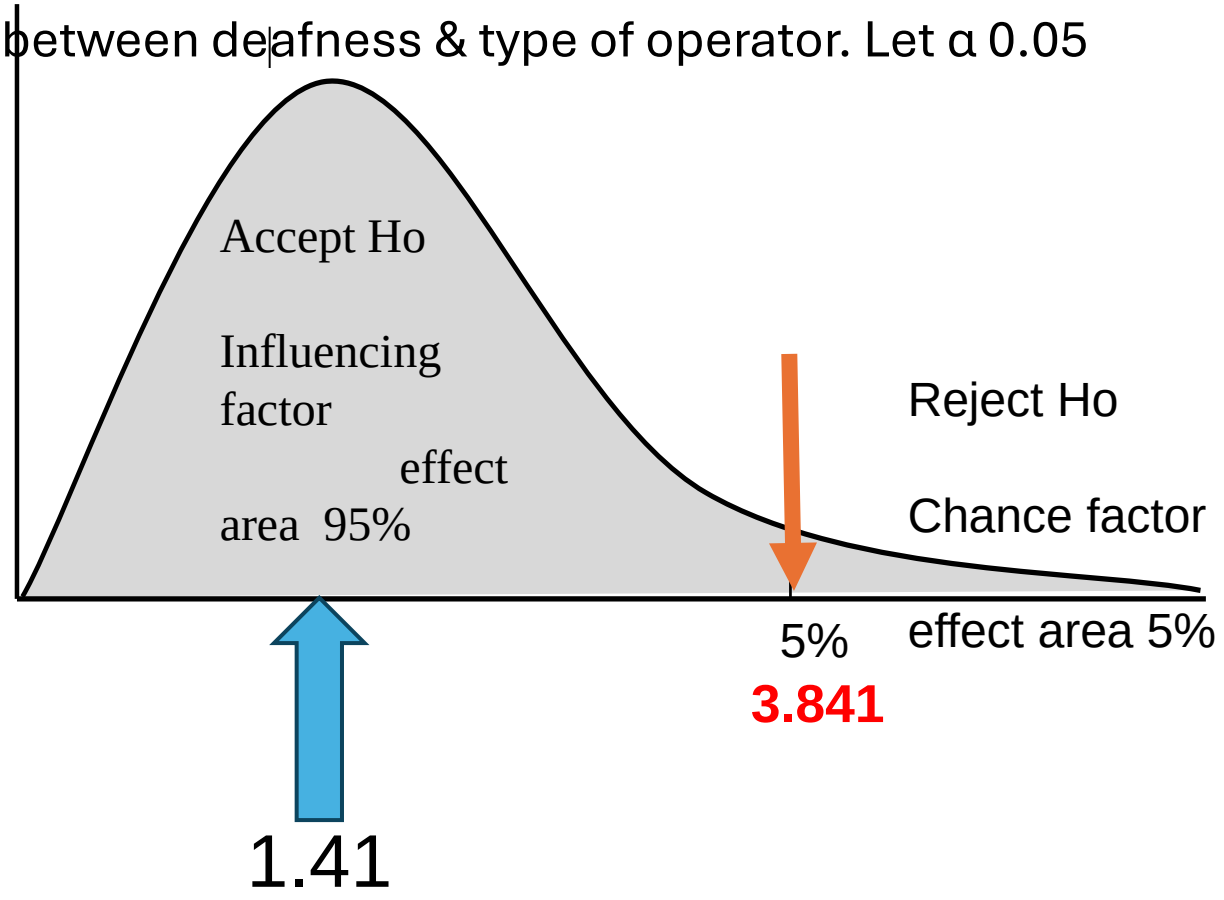
Calculated chi< tabulated chi

1.41 < 3.841

Conclusion: We accept H0

HO may be true

There is no significant **<u>association</u>** between type of operator & deafness.



Chi-Square test

Test for Independence using
(a x b chi-square test) or
(r x c chi-square test)

Calculation of expected frequencies: For r × c contingency table, the expected frequencies are as follow:

$$e_i = \frac{Row\ total(rt_i) \times Column\ total(ct_i)}{Grand\ total(n)}$$

Where e_i= expected frequency of cells and is $e_1, e_2, ..., e_k$ where k is the number of cells in the body of the table.

Consider the following 3 by 2 contingency table

Classification criteria 2	*Classification criteria 1*		
	Class 1	*Class 2*	*Total*
Category 1	a	b	$a+b$
Category 2	c	d	$c+d$
Category 3	e	f	$e+f$
Total	$a+c+e$	$b+d+f$	n

The expected value for the first cell (a), $e_1 = \frac{(a+b)(a+c+e)}{n}$

The expected value for the first cell (b), $e_2 = \frac{(a+b)(b+d+f)}{n}$

.....:

The expected value for the first cell (f), $e_6 = \frac{(e+f)(b+d+f)}{n}$

Chi-Square test χ^2

Q3. Perform a Chi-Square test to analyze the relationship between alcohol consumption (number of beers per day) and liver disease. The contingency table is given below.

Alcohol Drinking	*Liver Disease*		
(No. of bottle beers/day)	*Yes*	*No*	*Total*
≤2	20	80	100
3-5	90	30	120
≥6	240	40	280
Total	350	150	500

1. State the Hypotheses:

•**Null Hypothesis (H0):** There is no association between the number of beers consumed per day and the presence of liver disease. The two variables are independent.

•**Alternative Hypothesis (H1):** There is an association between the number of beers consumed per day and the presence of liver disease. The two variables are not independent.

2. The expected frequency for each cell is calculated as:

(Row Total * Column Total) / Grand Total

Expected values are shown in brackets with each cell.

Beers/Day	Liver Disease (Yes)	Liver Disease (No)	Total
≤2	20 (70)	80 (30)	100
3-5	90 (84)	30 (36)	120
≥6	240 (196)	40 (84)	280
Total	350	150	500

3. Calculate the Chi-Square Statistic (χ^2):

$\chi^2 = \Sigma [(\text{Observed} - \text{Expected})^2 / \text{Expected}]$

$= 35.71 + 83.33 + 0.43 + 1.00 + 2.21 + 7.74 = 153.4$

4. Find critical value for df= (3-1)(2-1) = 2 and alpha = 0.05

Critical value = 5.991

5. Compare: Our calculated χ^2 (130.42) is much greater than the critical value (5.991). Therefore, we reject the null hypothesis.

Conclusion: There is a statistically significant association between the number of beers consumed per day and the presence of liver disease. The variables are not independent.