Population and Sample

Population and Sample

Population:

- A **population** is any large collection of objects or individuals, such as Americans, students, or trees about which information is desired.

  Example:

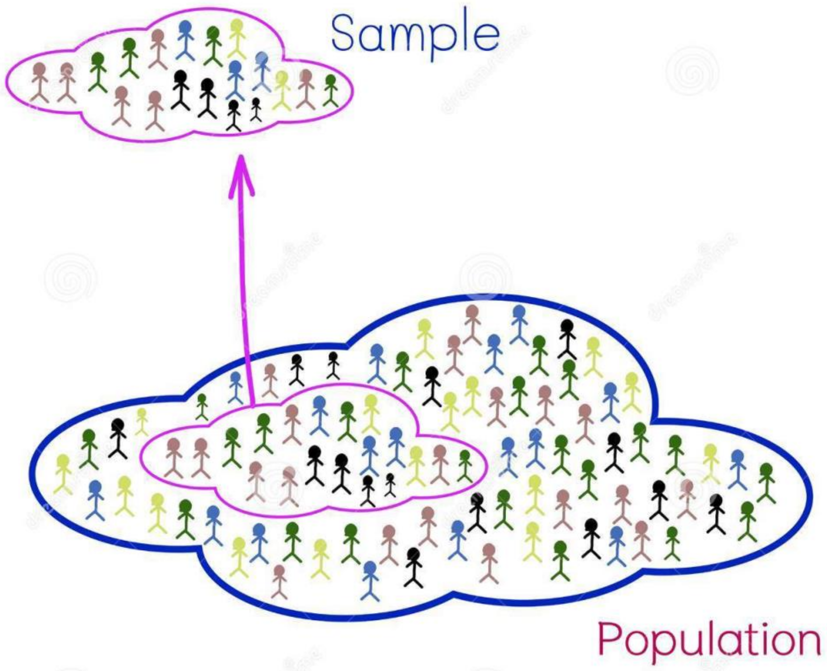
  - ✓ Packet of Food grains
  - ✓ A group of people suffering from a particular disease,
  - ✓ Collection of books,

Sample:

- Sample is the representative unit of the target population, which is worked upon by the researchers

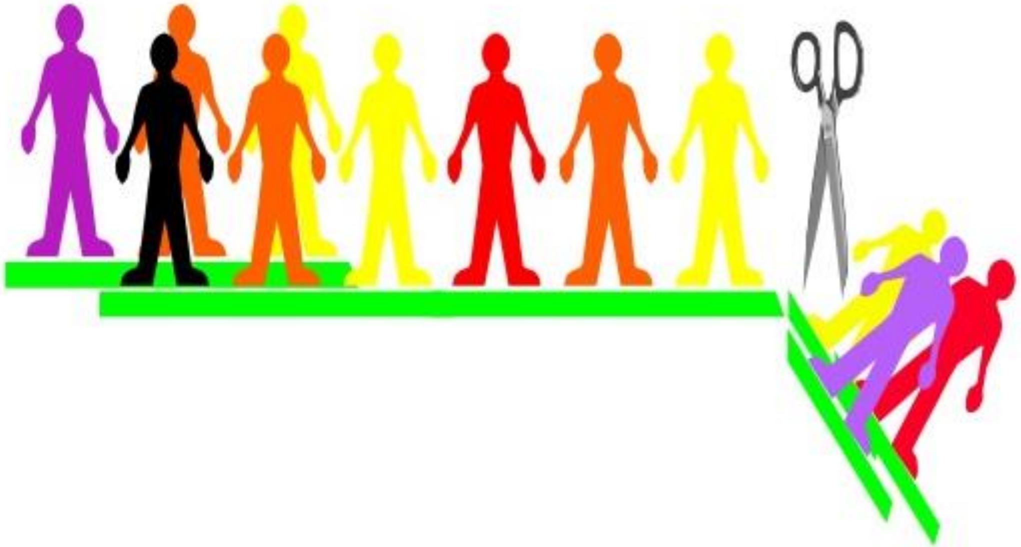
- While purchasing food grains, we inspect only a handful of grains and draw conclusions about the quality of the whole lot.

Note: In this case handful of grains is a sample and the whole lot is a population.



Introduction to Sampling

- Sampling is the method of selecting the number of individuals or objects in such a way that it represents the whole population.
- A sample is used to find out the characteristics of the population.
- The purpose of the sampling is to gather data in order to make inferences and make decisions about the population.

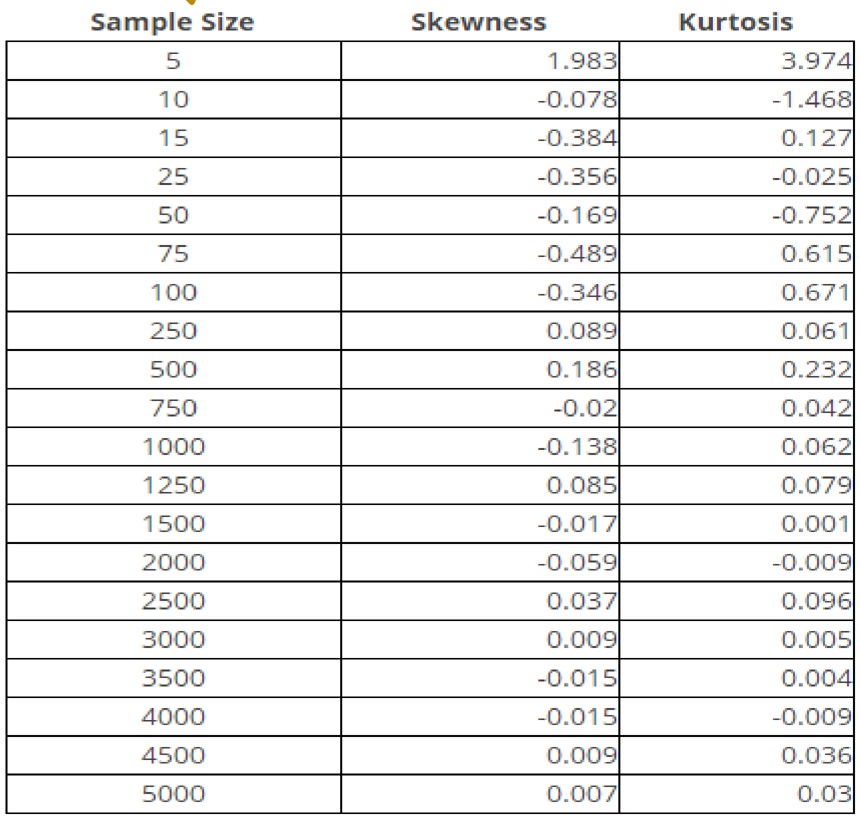


Sampling

proschool
An IMS Initiative

- Sampling considerations
  - ✓ Larger sample sizes are more accurate representations of the whole population.
  - ✓ The sample size chosen is a balance between obtaining a statistically valid representation, and the time, energy, money, labor, equipment and access available
  - ✓ A sampling strategy made with the minimum of bias is the most statistically valid

Larger the sample size, it would be closer to normal distribution.



Sample Size	Skewness	Kurtosis
5	1.983	3.974
10	-0.078	-1.468
15	-0.384	0.127
25	-0.356	-0.025
50	-0.169	-0.752
75	-0.489	0.615
100	-0.346	0.671
250	0.089	0.061
500	0.186	0.232
750	-0.02	0.042
1000	-0.138	0.062
1250	0.085	0.079
1500	-0.017	0.001
2000	-0.059	-0.009
2500	0.037	0.096
3000	0.009	0.005
3500	-0.015	0.004
4000	-0.015	-0.009
4500	0.009	0.036
5000	0.007	0.03

Advantages and Disadvantages of sampling

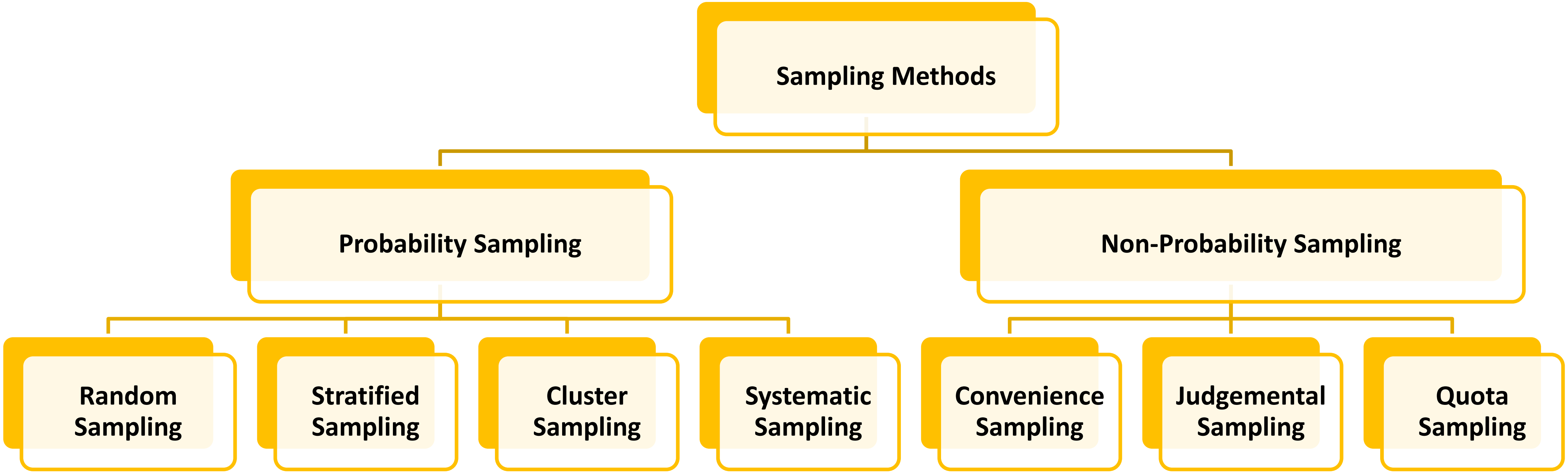
Advantages of Sampling

- ✓ Low cost
- ✓ Less time consuming
- ✓ Suitable in limited resources

Disadvantages of Sampling

- ✓ Difficult to select a truly representative sample
- ✓ It is important to have subject specific knowledge
- ✓ Chances of bias
- ✓ Sampling is impossible when population is too small and heterogeneous

Sampling Method



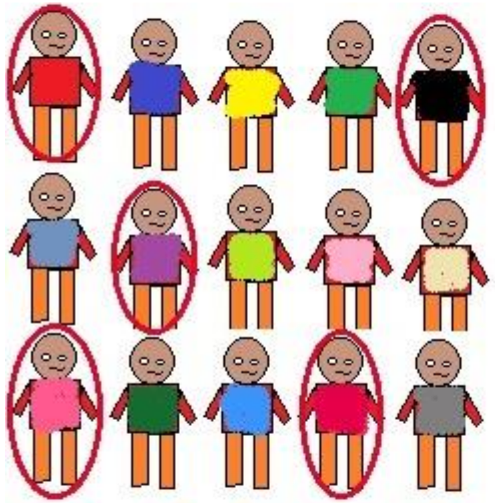
What is Probability Sampling?

- **Probability sampling** is a **sampling** technique where in the samples are gathered in a process that gives all the individuals in the population equal chances of being selected this is known as an 'equal probability of selection' (EPS) design.
- EPS designs are also referred to as 'self-weighting' because all sampled units are given the same weight.
- A researcher must identify specific sampling elements (e.g. persons) to include in the sample
- For example: A company HR wants to conduct a survey among the employees regarding company facilities and he selects employee who have specifically 2 years of company experience.

- Different types of Probability Sampling methods are:
  - ✓ Simple Random Sampling
  - ✓ Stratified Sampling
  - ✓ Cluster Sampling
  - ✓ Systematic Sampling

Simple Random Sampling

- It is applicable when population is small, homogeneous & readily available
- All subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection.
- It provides for greatest number of possible samples. This is done by assigning a number to each unit in the sampling frame.
- A table of random number or lottery system is used to determine which units are to be selected.
- The key to random selection is that there is no bias involved in the selection of the sample.
- Any variation between the sample characteristics and the population characteristics is only a matter of chance.



Simple Random Sampling

Simple Random Sampling Example

- A government income tax auditor must choose a sample of 5 of 11 returns to audit…[Can do many different ways]

Person	Generate Random #
baker	0.87487
george	0.89068
ralph	0.11597
mary	0.58635
sally	0.34346
joe	0.24662
andrea	0.47609
mark	0.08350
greg	0.53542
aaron	0.37239
kim	0.73809

	Person	Sorted Random #
1	mark	0.08350
2	ralph	0.11597
3	joe	0.24662
4	sally	0.34346
5	aaron	0.37239
	andrea	0.47609
	greg	0.53542
	mary	0.58635
	kim	0.73809
	baker	0.87487
	george	0.89068

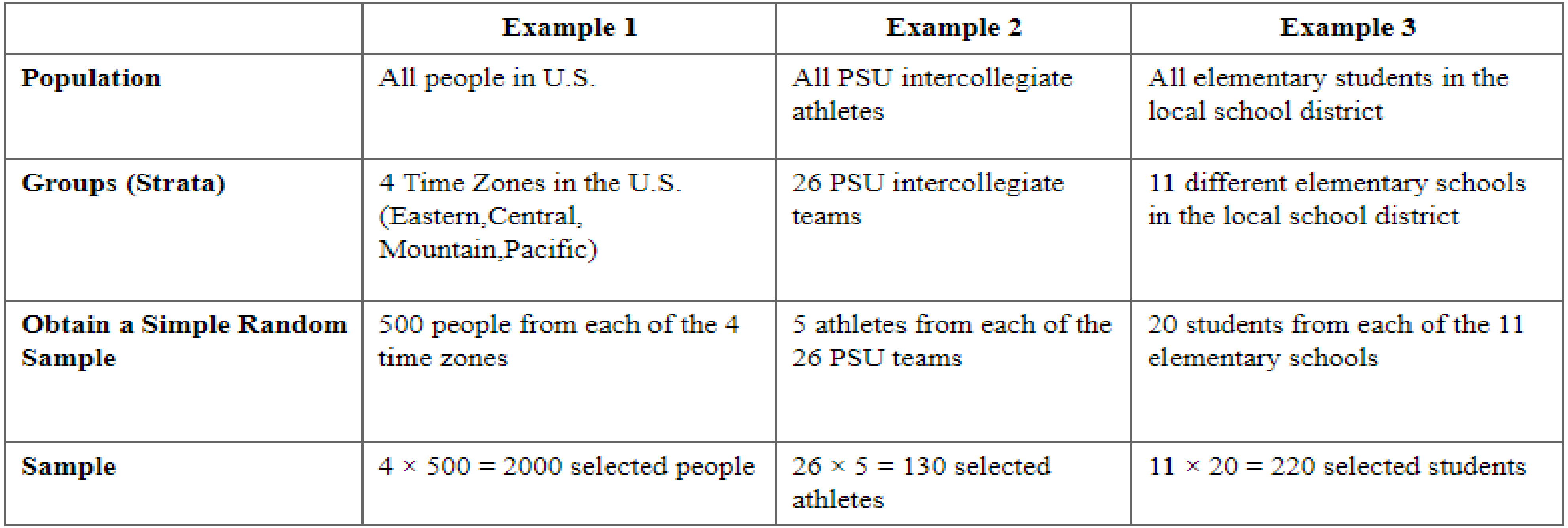
Stratified Sampling

- In stratified sampling, the researcher divides the population into separate groups, called strata for example we can divide the population in two strata: male and female
- The population is randomly sampled *within* each category or stratum.

- Steps for Stratified Sampling technique:
  - ✓ Partition the Population into groups(Strata)
  - ✓ Obtain a simple random sample from each group(Stratum)
  - ✓ Collect Data on each sampling unit that was randomly sampled from each group(Stratum)

Stratified Sampling Example

- For marketing Analysis scenario, The ABC company wants to find the different age group men and woman for launching some age group specific products in a given geography



	Example 1	Example 2	Example 3
Population	All people in U.S.	All PSU intercollegiate athletes	All elementary students in the local school district
Groups (Strata)	4 Time Zones in the U.S. (Eastern,Central, Mountain,Pacific)	26 PSU intercollegiate teams	11 different elementary schools in the local school district
Obtain a Simple Random Sample	500 people from each of the 4 time zones	5 athletes from each of the 26 PSU teams	20 students from each of the 11 elementary schools
Sample	4 × 500 = 2000 selected people	26 × 5 = 130 selected athletes	11 × 20 = 220 selected students

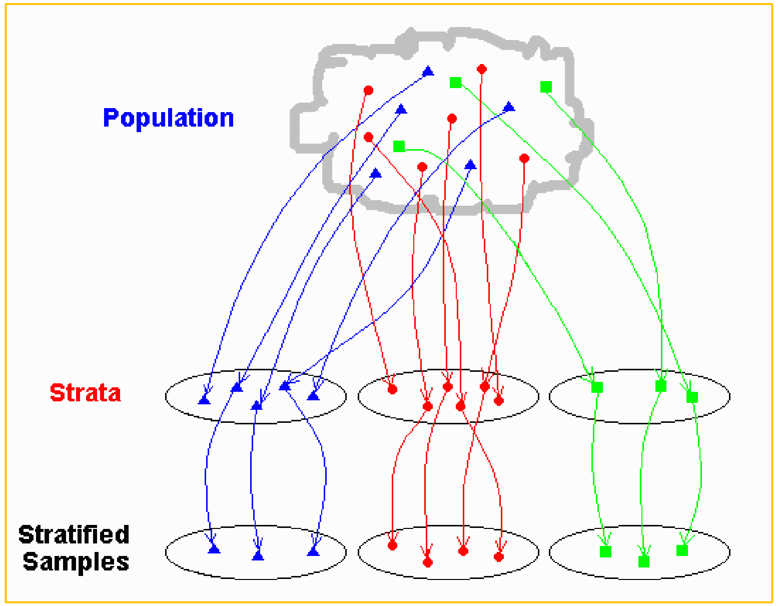
Cluster Sampling

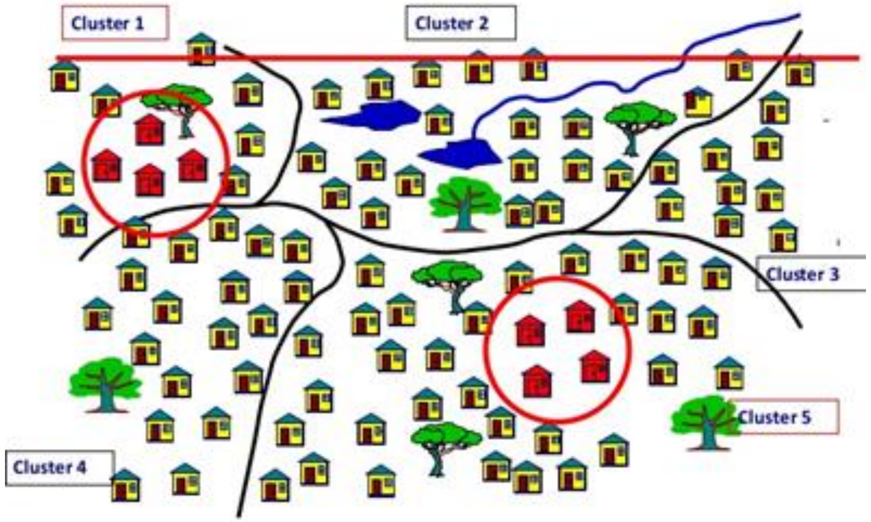
- **Cluster sampling** is a sampling technique used when "natural" but relatively homogeneous groupings are evident in a statistical population.
- Cluster Sampling is an example of 'two-stage sampling' .
    1. A sample of areas is chosen;
    2. Sample of respondents *within* those areas is selected.
- Population divided into clusters of homogeneous units, usually based on geographical contiguity.
- Sampling units are groups rather than individuals.
- A sample of such clusters is then selected.
- All units from the selected clusters are studied
- It is often used in marketing research.
- In this technique, the total population is divided into these groups (or clusters) and a simple random sample of the groups is selected.

CLUSTER SAMPLING
1st Ave
2nd Ave
3rd Ave
Clusters
People on block
CHOOSE

Difference between Strata and Cluster

- All strata are represented in the sample; but only a subset of clusters are in the sample.
- With stratified sampling, the best survey results occur when elements within strata are internally **homogeneous**.
- However, with cluster sampling, the best results occur when elements within clusters are internally **heterogeneous**.



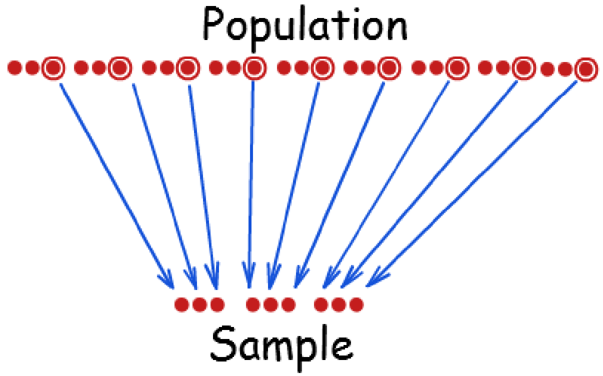


Systematic Sampling

- Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list.
- Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards.

  In this case,

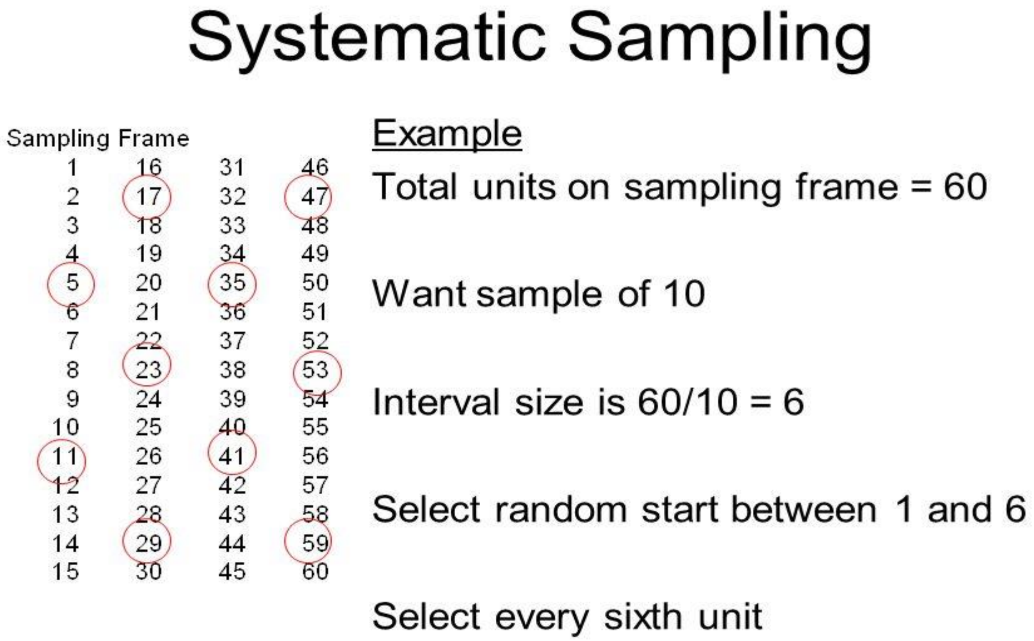
  k = (population size/sample size).
- It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.
- Samples are chosen in a systematic, or regular way.
- They are evenly/regularly distributed in a spatial context, for example every two meters along a transect line.
- They can be at equal/regular intervals in a temporal context, for example every half hour or at set times of the day.
- They can be regularly numbered, for example every 10th house or person
- A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10’).



Illustrative examples

Michael, I want to send my employees to a weeklong training session that is out of town. Due to limited funding, I cannot send all of them; I can afford only 10 employees out of total 60 employees. But I want to be unbiased during selection. Please help me out

Alex, I would suggest you to go for systematic sampling. Create a list of employees from 1 to 60 arranging them according to their work experience. Now, just select any random number between 1 and 6. Let's take it as 5, then add 6 to it to select the next employee .Therefore, we will send employees who have codes – 5,11,17,23,29,35,41,47,53 and 59.



Systematic Sampling

Sampling Frame

1	16	31	46
2	17	32	47
3	18	33	48
4	19	34	49
5	20	35	50
6	21	36	51
7	22	37	52
8	23	38	53
9	24	39	54
10	25	40	55
11	26	41	56
12	27	42	57
13	28	43	58
14	29	44	59
15	30	45	60

Example

Total units on sampling frame = 60

Want sample of 10

Interval size is 60/10 = 6

Select random start between 1 and 6

Select every sixth unit

Non Probability Sampling Method

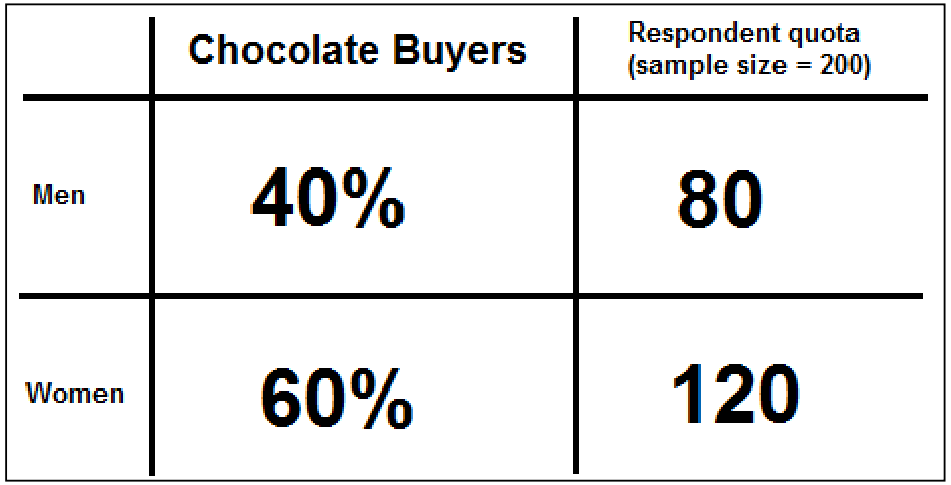
- Unequal chance of being included in the sample (non-random)
- Non random or non - probability sampling refers to the sampling process in which, the samples are selected for a specific purpose with a predetermined basis of selection.
- The sample is not a proportion of the population and there is no system in selecting the sample. The selection depends upon the situation.
- No assurance is given that each item has a chance of being included as a sample
- There is an assumption that there is an even distribution of characteristics within the population, believing that any sample would be representative.
- Non Probability sampling includes

  - ✓ Quota Sampling
  - ✓ Convenience Sampling
  - ✓ Judgement Sampling

Quota Sampling

- The defining characteristic of a quota sample is that the researcher deliberately sets the proportions of levels or strata within the sample. This is generally done to insure the inclusion of a particular segment of the population.

- The proportions may or may not differ dramatically from the actual proportion in the population. The researcher sets a quota, independent of population characteristics.



	Chocolate Buyers	Respondent quota (sample size = 200)
Men	40%	80
Women	60%	120

Quota Sampling

- *Example:* A researcher is interested in the attitudes of members of different religions towards the death penalty.

- In a school a random sample might miss females (because there are not many in that school). To be sure of their inclusion, a researcher could set a quota of 3% females for the sample.

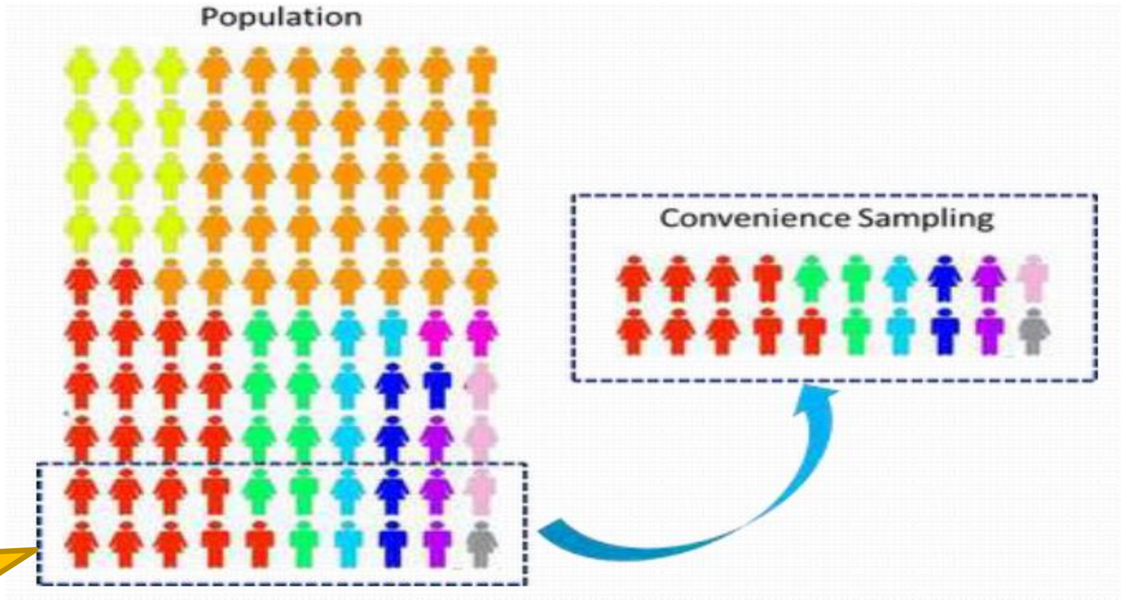
- However, the sample will no longer be representative of the actual proportions in the population. This may limit generalizing to the school population. But the quota will guarantee that the views of females are represented in the survey.

Convenience Sampling

- **Convenience sampling** is a sample taken from a group you have easy access to. The idea is that anything learned from this study will be applicable to the larger population.
- By using a large, convenient size, you are able to more confidently say the sample represents the population.
- Furthermore, the convenient group you are testing should not be fundamentally different than if you had taken a sample from another area. If you are trying to say something about women,
  For example, then your convenient sample cannot be men.
- Involves collecting information from members of a population who are conveniently available to provide this information
- Example : 'Pepsi Challenge' contest with the purpose of determining whether people prefer one product over another, might be set up at a shopping mall visited by many shoppers
- Example : Suppose 100 car owners are to be selected. Then we may collect from the RTO's office the list of car owners and then make a selection of 100 from that to form the sample.

Convenience Sampling

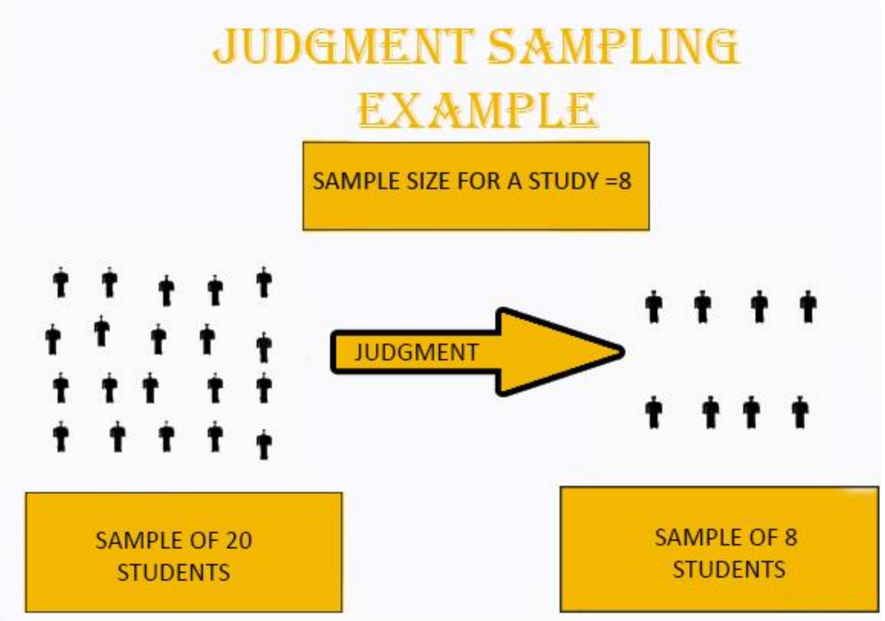
- You are interested in the effects of caffeine on study habits of college students. To test the whole population you would need all current college students and a whole lot of time and soda.
- A sample would be a test of a few college students from all of the colleges in the India, requiring you to fly them in for the testing.
- A convenience sample would be a large group of college students from your local college or colleges. They are close by, are in college, and are not different than other college students.



Easily accessible population area- convenient to sample

Judgement Sampling

- **Judgment sample** is a type of nonrandom sample that is selected based on the opinion of an expert.
- Results obtained from a judgment sample are subject to some degree of bias, due to the frame and population not being identical.
- The frame is a list of all the units, items, people, etc., that define the population to be studied.
- This is used primarily when there is a limited number of people that have expertise in the area being researched
- Example : A TV researcher wants a quick sample of opinions about a political announcement. Taking views of people in the street.



Probability

Importance of Probability

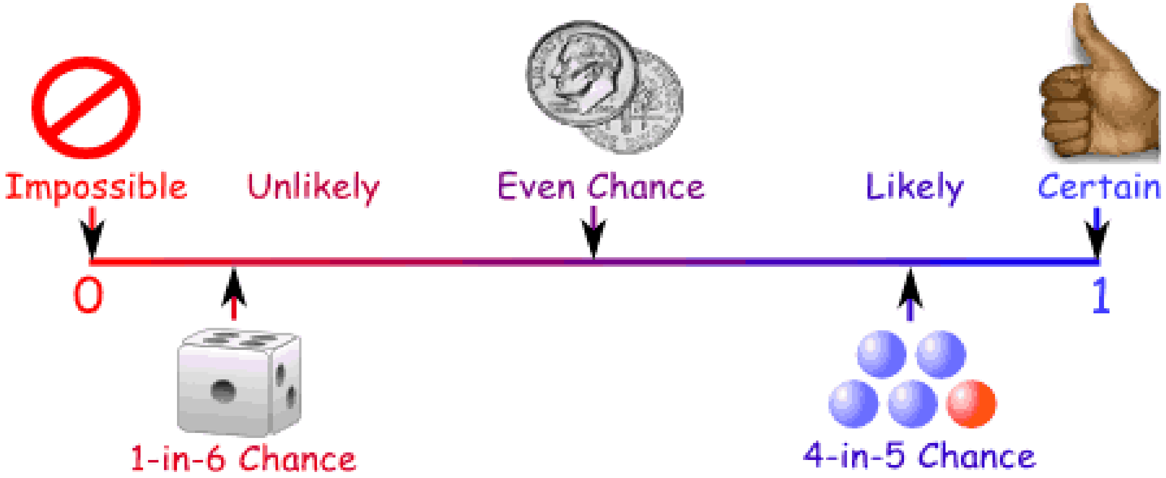
- There are branches of maths that equip you to make decisions in situations in which you have perfect information.
- Probability trains you to make decisions in situations which there are observable patterns, but a degree of uncertainty. Uncertainty and randomness occur in just about every field of application and in daily life for example probably the price of X share will go up, probably 'X' team will win the match , so it is extremely useful and interesting to understand probability.

Probability

- Probability is a branch of mathematics that deals with calculating the likelihood of a given event's occurrence.
- Probability is the measure of the likeliness that an event will occur.
- Probability is quantified as a number between 0 and 1

(where 0 indicates impossibility and 1 indicates certainty).

- The higher the probability of an event, the more certain that the event will occur.



Probability

- Probability does not tell us exactly what will happen, it is just a guide.

  Example: toss a coin 100 times, how many heads will come up?
- Probability says that heads have a ½ chance, so we can expect 50 Heads.
- But when we actually try it we might get 48 heads, or 55 heads ... or anything really, but in most cases it will be a number near 50.

Probability

- Probability is written as,

$$\textbf{Probabilit (Event happening)} = \frac{\textbf{Number of ways it can happen}}{\textbf{Total Number of Outcomes}}$$

- What is the probability that an even number would come if we throw a dice

Ans. Probability of the event would be -> number of ways it can be(3)/total number of outcome(6)
= 0.50

Example of Probability

- Example : The chances of rolling a 4 with a die

  Number of ways it can happen : 1(There is only one face with a "4" on it)

  Total number of outcomes : There are 6 faces altogether

  So, the probability = $\frac{1}{6}$

- Example : There are 5 marbles in a bag : 4 are blue and 1 is red. What is the probability that the blue marble gets picked?

  Number of ways it can happen : 4(There are "4" blue)

  Total number of outcomes : There are 5 marbles in a bag

  So, the probability = $\frac{4}{5}$ = 0.8

Key Terms in Probability

- **Statistic:** A statistic is a number that represents a property of the sample.

For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic.

- **Parameter:** A parameter is a number that is a property of the population.

For example, Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

- **Experiment or Trial:** An action where the result is uncertain.

For example: Tossing a coin, throwing dice, seeing what pizza people choose.

- **Sample Space:** The set of all possible outcomes of an experiment.

For Example: Selecting a card from a deck. There are 52 cards in a deck (excluding Jokers) Hence, the Sample Space is all 52 possible cards:

{Ace of Hearts, 2 of Hearts, etc... }

Key Terms in Probability

- **Sample Points:** The Sample Space is made up of Sample Points. The elements of sample space are sample points.
  For example: In the deck of cards- the 5 of Clubs is a sample point , the King of Hearts is a sample point "King" is not a sample point. As there are 4 Kings that is 4 different sample points.

- **Event:**

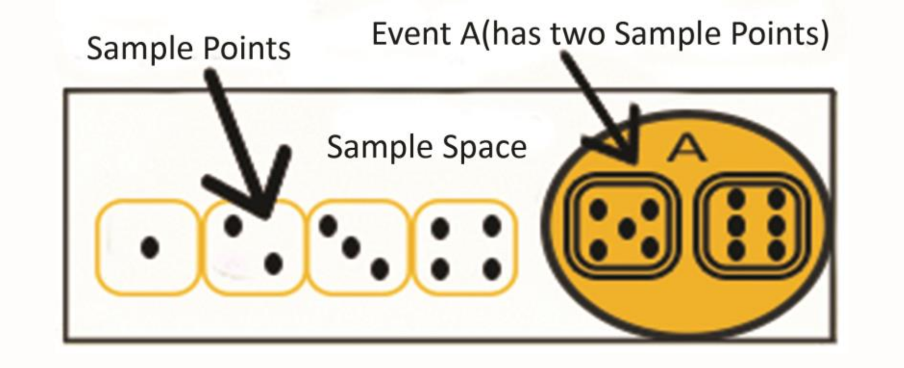
  A single result of an experiment

  Example :

  ✓ Getting a Tail when tossing a coin is an event, getting a "5" when a die is rolled is an event.

  ✓ An event can include one or more possible outcomes: Choosing a "King" from a deck of cards (any of the 4 Kings) is an even

  ✓ Rolling an "even number" (2, 4 or 6) is also an event



Types of Events

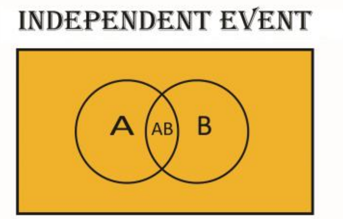
- **Dependent Event:** Dependent event also called "Conditional", where one event is affected by other events
- **Example**:
- Drawing 2 Cards from a Deck. After taking one card from the deck there is one card less than the previous, so the probability changes!

  Let's look at the chances of getting a King.
  For the 1st card, the chance of drawing a King is 4 out of 52
  But for the 2nd card, If the 1st card was a King, then the 2nd card is less likely to be a King, as only 3 of the 51 cards left are Kings.
  If the 1st card was not a King, then the 2nd card is slightly more likely to be a King, as 4 of the 51 cards left are King.
  This is because we are removing cards from the deck.

Types of Events

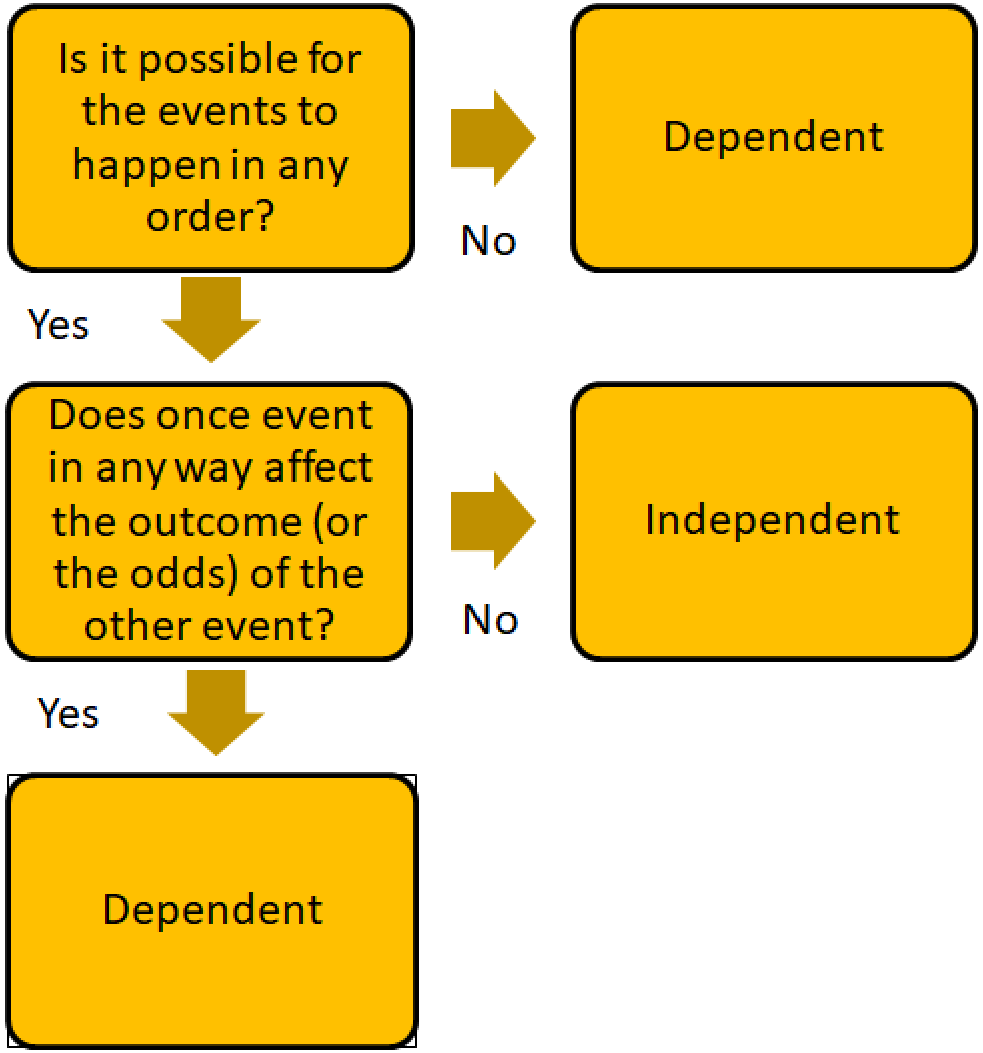
- **Independent Event:** One event is not affected by any other events.

- For example:
  - ✓ You toss a coin three times and it comes up "Heads" each time. what is the chance that the next toss will also be a "Head"?
    The chance is simply 1/2, or 50%, just like ANY OTHER toss of the coin. What it did in the past will not affect the current toss.
  - ✓ Radhika draws a pair of socks with each of the following color, Blue, Red, black and White. Each pair is folded together in a matching set. She reach to the drawer and choose a pair of socks without looking. She replace this pair and then choose another pair of socks. What is the probability that she will choose the red pair of socks both times?



Dependent or Independent

Dependent or Independent?



Is it possible for the events to happen in any order?

No

Dependent

Yes

Does once event in any way affect the outcome (or the odds) of the other event?

No

Independent

Yes

Dependent

Conditional Probability

- A **Conditional probability** is a probability whose sample space has been limited to only those outcomes that fulfill a certain condition.
- A rule that can be used to determine a conditional probability from unconditional probabilities is:

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

where:

- P(A | B) = Conditional probability that event A will occur given that event B has occurred already
- $P(A \cap B)$ = Unconditional probability that event A and event B both occur
- P(B) = Probability that event B occurs
- The usual notation for "event A occurs given that event B has occurred" is "A | B" (A given B).
- The symbol | is a vertical line and does not imply division.

Conditional Probability

- In many situations, once more information becomes available, we are able to revise our estimates for the probability of further outcomes or events happening.
- For example, suppose you go out for lunch at the same place and time every Friday and you are served lunch within 15 minutes with probability 0.9.
- However, given that you notice that the restaurant is exceptionally busy, the probability of being served lunch within 15 minutes may reduce to 0.7.
- This is the conditional probability of being served lunch within 15 minutes given that the restaurant is exceptionally busy.

Conditional Probability

Example 1: A math teacher gave her class two tests. 25% of the class passed both tests and 42% of the class passed the first test. What percent of those who passed the first test also passed the second test?

Solution:

$$P(\text{Second}|\text{First}) = \frac{\underline{P(\text{First and Second})}}{P(\text{First})} = \frac{\underline{0.25}}{0.42} = 0.60 = 60\%$$

- This problem describes a conditional probability since it asks us to find the probability that the second test was passed given that the first test was passed.

Example 2: The probability that it is Friday and that a student is absent is 0.03. Since there are 5 school days in a week, the probability that it is Friday is 0.2. What is the probability that a student is absent given that today is Friday?

Solution:

$$P(\text{Absent}|\text{Friday}) = \frac{\underline{P(\text{Friday and Absent})}}{P(\text{Friday})} = \frac{\underline{0.03}}{0.2} = 0.15 = 15\%$$

Bayes theorem

- Bayes theorem shows the relationship between one conditional probability and its inverse
- It provides a mathematical rule for revising an estimate of experience and observation

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

- Related terms in Bays theorem
  - ✓ P(A):The probability of event A not concerning its associated event B. This is also called as Prior probability of A
  - ✓ P(B):The probability of event B not concerning its associated event A. This is also called as Prior probability of A
  - ✓ P(B|A) : Conditional probability of B given A. This is also called as likelihood
  - ✓ P(A|B) : Conditional probability of A given B. This is also called as posterior probability

Bayes Example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on the day of Marie's wedding?

Solution: The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. Notation for these events appears below.

Event A_1. It rains on Marie's wedding.
Event A_2. It does not rain on Marie's wedding.
Event B. The weatherman predicts rain.

In terms of probabilities, we know the following
$(A_1) = 5/365 = 0.0136985$ [It rains 5 days out of the year.]
$P(A_2) = 360/365 = 0.9863014$ [It does not rain 360 days out of the year.]
$P(B \mid A_1) = 0.9$ [When it rains, the weatherman predicts rain 90% of the time.]
$P(B \mid A_2) = 0.1$ [When it does not rain, the weatherman predicts rain 10% of the time.]

Bayes Example Continued

- We want to know P(A_1 | B), the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes' theorem, as shown below.

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)}$$

$$P(A_1|B) = \frac{0.014 \ * 0.9}{0.014 * 0.9 + 0.986 * 0.1} = 0.111$$

- Note the somewhat unintuitive result. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, there is a good chance that Marie will not get rained on at her wedding.

Random Variable

- A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.
- A random variable can be classified as discrete or continuous depending upon the numerical value it assumes
- For Example: Let's look at a very commonly used example of a coin toss
- Suppose you flip a coin 100 times
  - ✓ How many times you will get a head?
  - ✓ Will you always get consecutive 5 heads?

Types of Random Variable

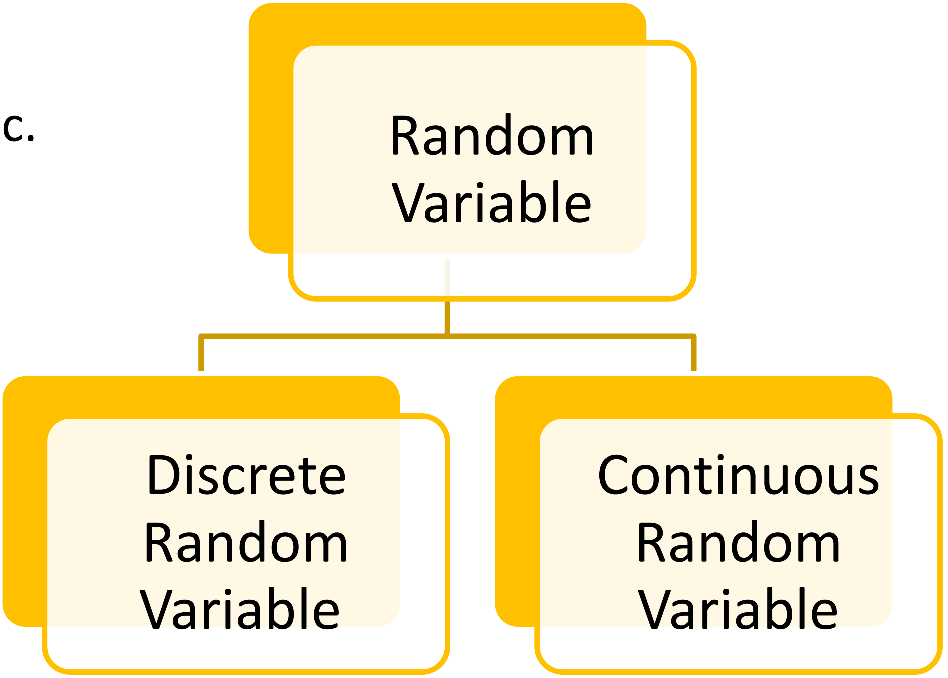
- Types of Random Variable:

✓ **Discrete Random Variable:** Variables assume only a countable number of values such as 1,2,3 , 4 ,5

Example: Attendance in a class, No of attendees in the seminar

✓ **Continuous Random Variable:** Any of the countless number of values in an interval such as: 1.1, 1.2, 1.3……….

Example: Height , Weight, Amount of Sugar in fruit etc.



Random Variable

Discrete Random Variable

Continuous Random Variable

Probability Distribution

A probability distribution is chart that links each outcome of a statistical experiment with its probability of occurrence

- Probability distribution for a random variable describes how probabilities are distributed over the values of the random variable
- Example: Probability distribution for Number of heads in 4 flips of a coin…or say probability of getting 0 head, 1 head, 2 head ……..
- Probability distribution chart is a visual representation
  - ✓ X-axis: All possible outcome of random variable
  - ✓ Y- axis: Probability of each outcome

Types of Probability distribution

- Discrete Probability Distribution: The list of each possible random variable can assume together with its probability

Example:

- ✓ Number of complaints per day
- ✓ Number of rings before the phone has answered
- ✓ Whether the piece is defective or not

- Continuous Probability Distribution : The probability of the random variable assuming a value within some given interval is defined to be the area under the graph of the probability density function between that interval

Example:

- ✓ Time required to complete a task
- ✓ Temperature of a solution
- ✓ Height in inches
- ✓ Weight in Kg

Discrete Probability Distribution

- If the random variable can have only discrete outcomes such 1,2,3,4 5 etc, we have to use a discrete probability distribution.

- Types of Discrete Distribution
  - ✓Binomial or Bernoulli distribution
  - ✓Negative Binomial distribution
  - ✓Geometric Distribution
  - ✓Poisson Distribution

Binomial Distribution

- The number of successes *x* in *n* repeated trials of a binomial experiment is called binomial random variable
  - ✓ Toss a coin it has only two outcome i.e. Head or Tail
  - ✓ Gender of Babies delivered in a hospital

- The probability distribution of a binomial random variable is called a **binomial distribution**.
- Properties of Binomial Distribution
  - ✓ The experiment consists of *n* repeated trials.
  - ✓ Each trial can result in just two possible outcomes – Success or Failure
  - ✓ The probability of success, denoted by *P*, is the same on every trial.
  - ✓ Independent trials i.e. the outcome on one trial does not affect the outcome on other trials

Binomial Distribution

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

- The experiment consists of repeated trials. We flip a coin 2 times.
- Each trial can result in just two possible outcomes - heads or tails.
- The probability of success is constant - 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Formula for Binomial Distribution

- The mathematical formula to calculate these probabilities is called **probability distribution function.**

- **For Binomial Distribution:**

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

Where,

x = Outcomes

n = Trials

p = probability of success on each trials

Binomial Distribution Example

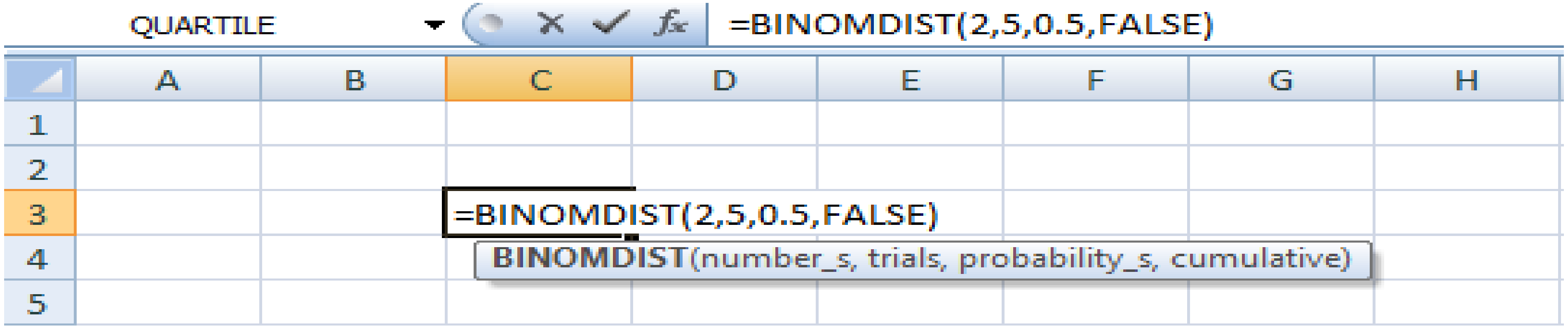
- If you Toss a coin 5 times, find the probability of getting exactly 2 heads

Solution:

Number_s	2
trails	5
Probability_s	0.5
Cumulative	FALSE

The probability is 31%

In Excel: =BINOM.DIST(2,5,0.50,FALSE)



```
=BINOMDIST(2,5,0.5,FALSE)
```

Binomial Distribution Example

proschool
An ims Initiative

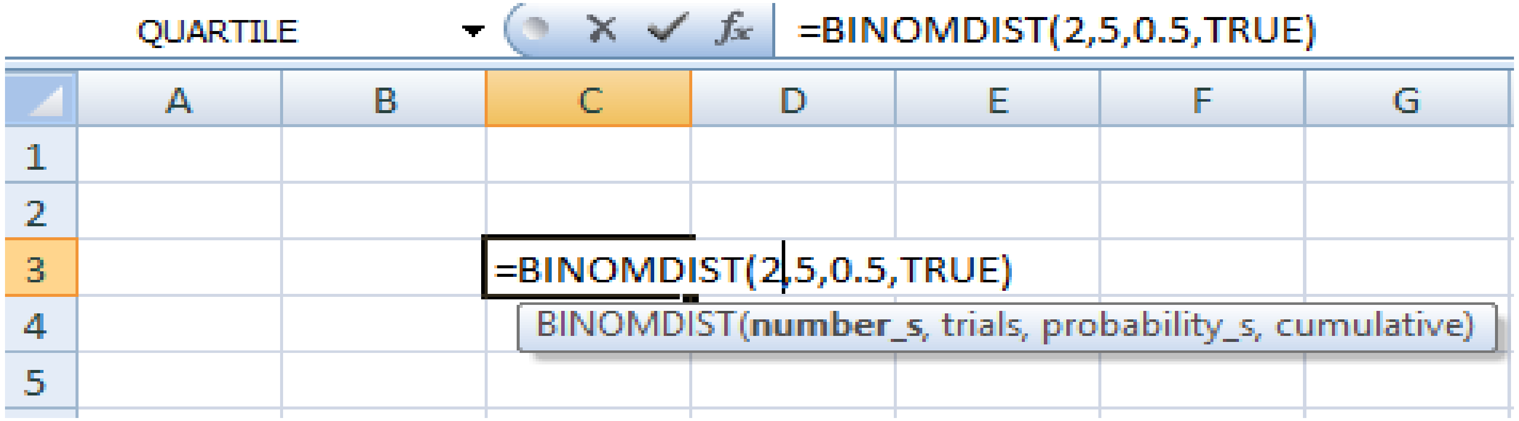
- If you toss the same coin 5 times, find the probability of getting upto 2 heads.

- When you have to find the values upto two heads which means: probability of getting 1 head & 2 head, you will write "True" in cumulative in excel instead of "false"

Solution:

Number_s	2
trails	5
Probability_s	0.5
Cumulative	TRUE

The probability is 50%
In Excel: =BINOM.DIST(2,5,0.50,True)



=BINOMDIST(2,5,0.5,TRUE)

=BINOMDIST(2,5,0.5,TRUE)

BINOMDIST(**number_s**, trials, probability_s, cumulative)

Binomial Distribution Example

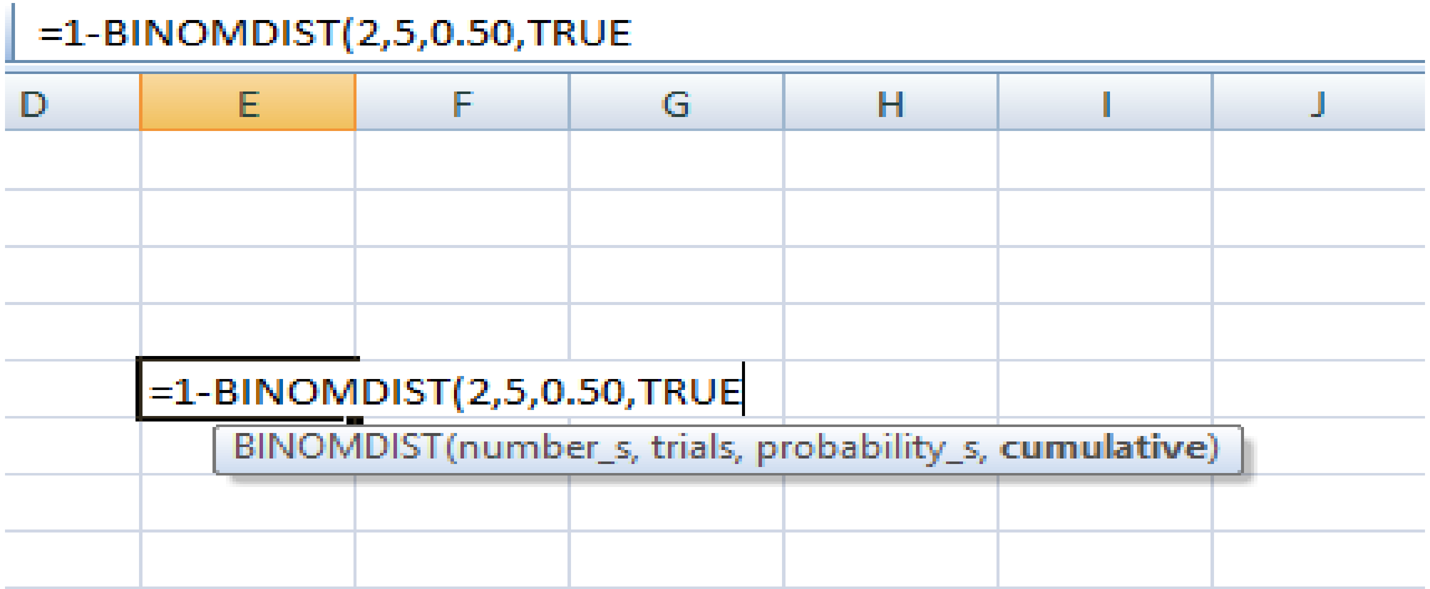
- If you toss the coin 5 times, find the probability of getting more than 2 heads.

Solution:

Number_s	2
trails	5
Probability_s	0.5
Cumulative	TRUE

The probability is 50%

In Excel: =1-BINOM.DIST(2,5,0.50,true)



=1-BINOMDIST(2,5,0.50,TRUE

=1-BINOMDIST(2,5,0.50,TRUE

BINOMDIST(number_s, trials, probability_s, cumulative)

Binomial Distribution Example

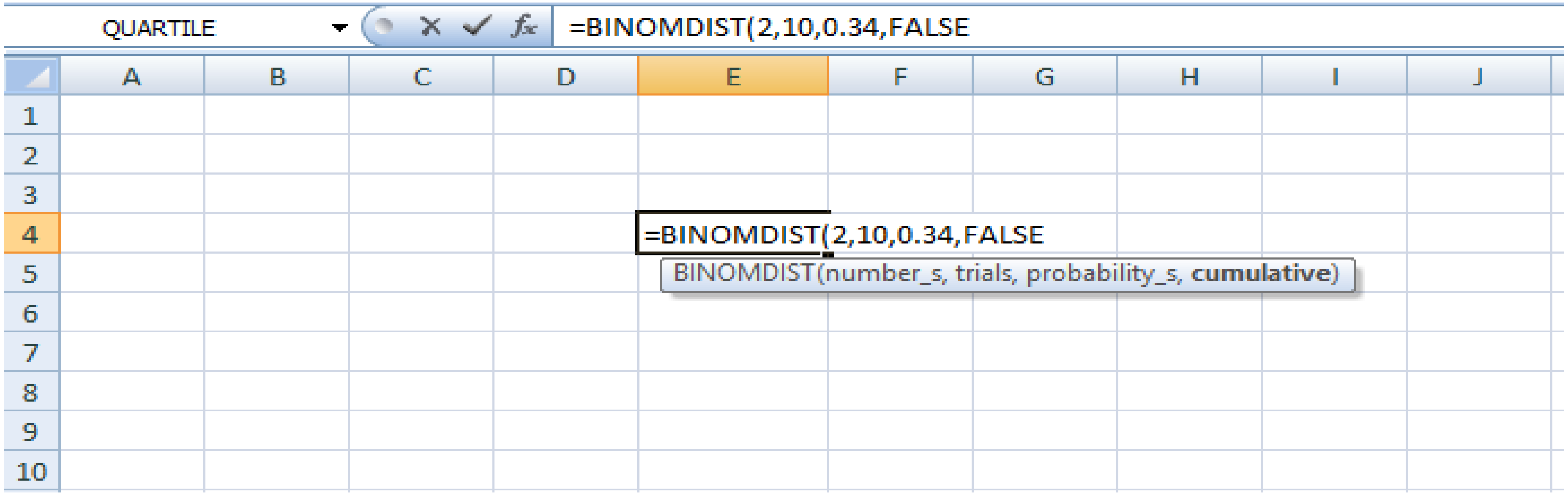
- An E-commerce company delivers 3.4% defective goods to its company . A sample of 10 deliveries is taken , What is the probability that the sample contains exactly 2 defective parts?

Solution:

Number_s	2
trails	10
Probability_s	0.034
Cumulative	FALSE

The probability that it examines 10 samples and finding 2 defects is 4%

In Excel: =BINOM.DIST(2,10,0.034,FALSE)



QUARTILE

=BINOMDIST(2,10,0.34,FALSE

=BINOMDIST(2,10,0.34,FALSE

BINOMDIST(number_s, trials, probability_s, **cumulative**)

Binomial Distribution Cumulative Value

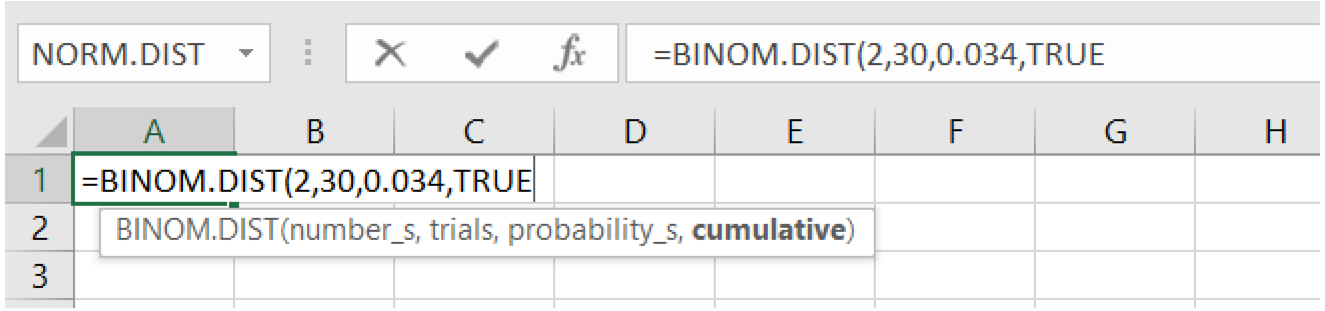
- An E-commerce company delivers 3.4% defective goods to its company . A sample of 30 deliveries is taken , What is the probability that the sample contains upto 2 defective parts?

Solution:

Number_s	2
trails	30
Probability_s	0.034
Cumulative	TRUE

The probability that it examines 30 samples and finding upto 2 defects is 91.92%

In Excel: =BINOM.DIST(2,30,0.034,True)



=BINOM.DIST(2,30,0.034,TRUE

BINOM.DIST(number_s, trials, probability_s, **cumulative**)

Hypergeometric Distribution

- The number of successes that result from a hypergeometric experiment is called a **Hypergeometric Random Variable**
- The probability distribution of a hypergeometric random variable is called a **Hypergeometric Random Distribution**.

- **Parameters of Hypergeometric distribution are as follows**

  - ✓ A Sample of size *n* is randomly selected without replacement from a population of *N* items.
  - ✓ In population,
    - *k* items can be classified as successes
    - *N* - *k* items can be classified as failures.

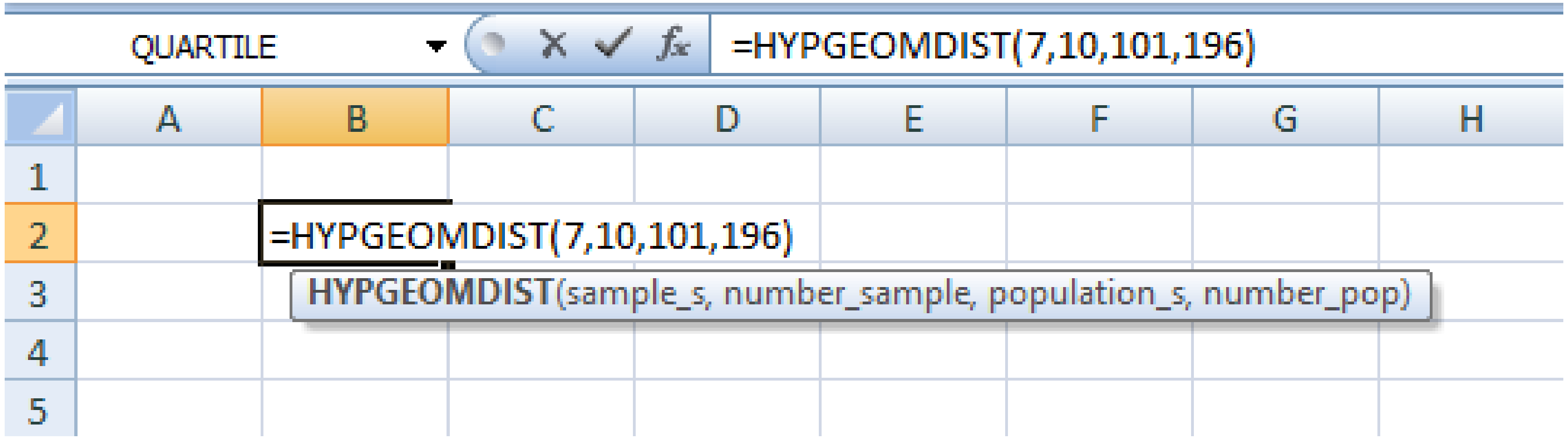
Hypergeometric Distribution Example

- A small voting district has 101 female voters and 95 male voters. A random sample of 10 voters is drawn. What is the probability exactly 7 of the voters will be female?

Solution:

Sample_s	7
number_sample	10
population_s	101
number_pop	196

The probability that exactly 7 voters will be female is 13%.



=HYPGEOMDIST(7,10,101,196)

=HYPGEOMDIST(7,10,101,196)

HYPGEOMDIST(sample_s, number_sample, population_s, number_pop)

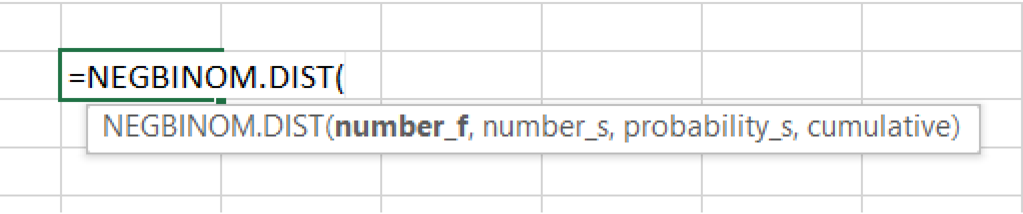
Negative Binomial Distribution Example

- This type of distribution concerns the number of trials that must occur in order to have a predetermined number of successes or in other words it is concerns with the number of trials *X* that must occur until we have *r* successes.
- For example: "What is the probability that we get three heads in the first *X* coin flips?

- An Indian oil company conducts a geological study that indicates that an exploratory oil well should have a 20% chance of striking oil. What is the probability that the first strike comes on the third well drilled?

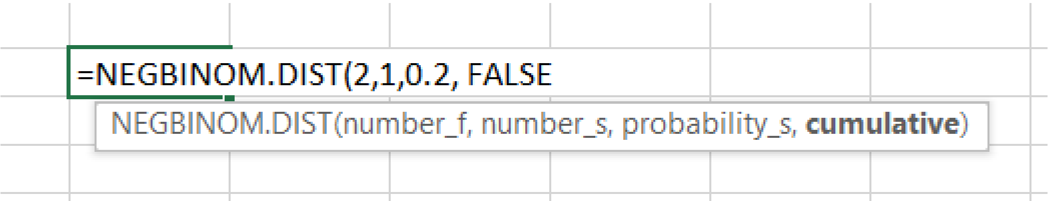
In our Example:

number_f(no. of failure)	2
number_s(no. of successes)	1
probability_s	0.2
Cumulative	FALSE



=NEGBINOM.DIST(

NEGBINOM.DIST(**number_f**, number_s, probability_s, cumulative)



=NEGBINOM.DIST(2,1,0.2, FALSE

NEGBINOM.DIST(number_f, number_s, probability_s, **cumulative**)

In Excel: NEGBINOM.DIST(2,1,0.2,false) = 0.128

Geometric Distribution

- Geometric distribution is a special case of negative binomial distribution where number of successes(r) is equal to 1
- The experiment consists of a sequence of trials with the following conditions:
  - ✓ The trials are independent.
  - ✓ Each trial can result in one of two possible outcomes, success and failure.
  - ✓ The probability of success is the same for all trials.

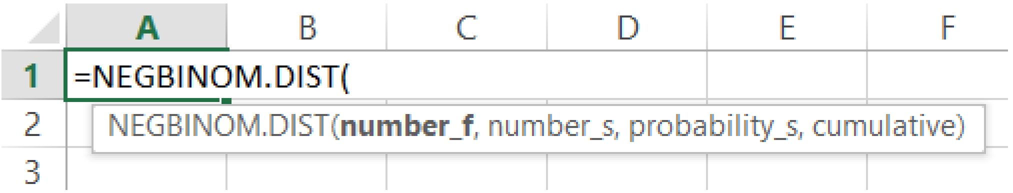
Geometric Distribution

- Example – In a country 10% of the people evade legitimate taxes. What is the probability that a tax official will need to raid at most 20 people, before finding a tax evader (first tax evader)

Solution:

number_f	19
number_s	1
probability_s	0.1
Cumulative	FALSE

The probability that the tax official will need to raid at most 20 people is 1%.



=NEGBINOM.DIST(

NEGBINOM.DIST(**number_f**, number_s, probability_s, cumulative)

Poisson Distribution

- A statistical distribution showing the frequency probability of specific events when the average probability of a single occurrence is known. The Poisson distribution is a discrete function.

- Example: On an average, 12 people visit a restaurant in one hour, what is the probability that 15 people may visit in next one hour.

- **Properties of Poisson Experiments :**
  - ✓ The experiment results classified as successes or failures.
  - ✓ Average number of successes that occurs in a specified region is known.
  - ✓ Probability that a success will occur is proportional to the size of the region.
  - ✓ Probability that a success will occur in an extremely small region is virtually zero.
  - ✓ Events have to be counted as a whole number

Poisson Distribution

- Poisson Probability can be calculated as :

$$P(X = x) = \frac{\lambda^x * e^{-\lambda}}{x!}$$

Where,
Lamda is a mean number of occurrences in a given interval of time

Applications of Poisson Distribution

- Manufacturing
- Operations and Supply chain
- Insurance

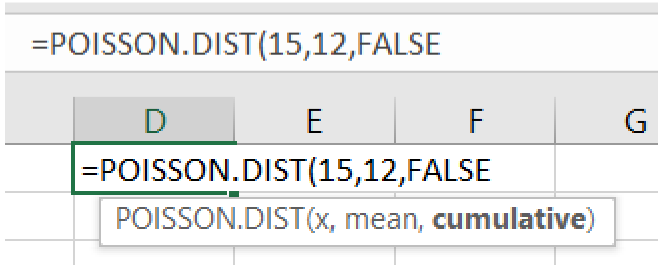
Poisson Distribution Example

- On an average, 12 people visit a restaurant in one hour, what is the probability that 15 people may visit in next one hour.

  Solution:

x	15
mean	12
Cumulative	FALSE

- The probability that 15 people may visit in next one hour is 7%.



=POISSON.DIST(15,12,FALSE

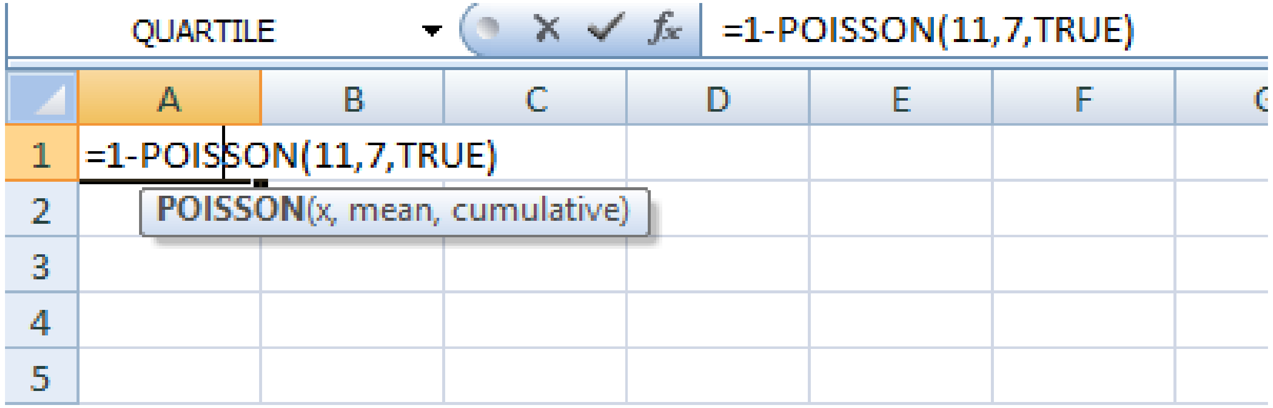
=POISSON.DIST(15,12,FALSE

POISSON.DIST(x, mean, **cumulative**)

Poisson Distribution Example

- A Pizza shop has a staff of 25 workers, which deliver 175 pizzas a day. A long weekend is coming up and 5 of the workers have asked for a holiday. You estimate remaining 20 workers can manage 15% greater delivery but want to plan for the chance of greater than 25% increase of delivery.
  Solution: 175/25 = 7 delivery a day
  If 15% greater delivery is received with 5 less resources= (175*1.15)/20 = 10.6 = 11

- We need a probability that if there is a requirement of 11 or more pizzas delivered in a day when the average is 7.



QUARTILE

=1-POISSON(11,7,TRUE)

POISSON(x, mean, cumulative)

Poisson or Binomial Distribution?

Poisson Distribution

- A Poisson Distribution is used, If a mean / average probability of an event happening per unit time/ per page/per mile cycled etc., is given, and you are asked to calculate a probability of n events happening in a given time / number of pages / number of miles cycled.
- It describes the distribution of binary data from a infinite sample. Thus, it gives the probability of getting r events in a population

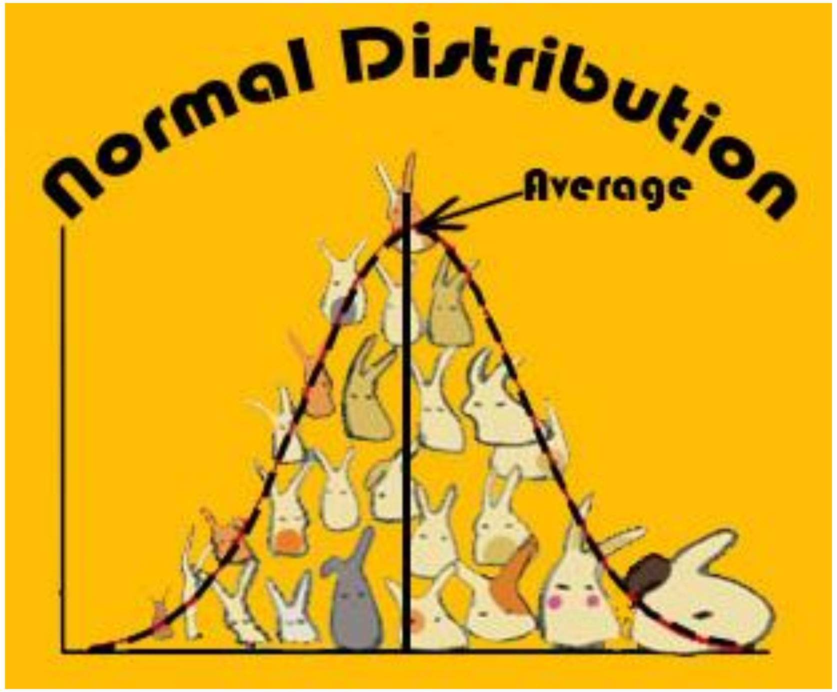
Binomial Distribution

- The Binomial Distribution is used when an exact probability of an event happening is given or implied, in the question and you are asked to calculate the probability of this event happening k times out of n.
- It describes the distribution of binary data from a finite sample. Thus, it gives the probability of getting r events out of n trials

Continuous Probability Distribution

- The probabilities of the possible values of a continuous random variable is a continuous distribution.
- A continuous random variable is a random variable with a set of possible values i.e. infinite and uncountable. For example: Height of women in Pune : 60 inch, 60.5 inch, 70.1 inch and so on.
- Normal Distribution is the most common kind of a continuous probability distribution due to its applications in statistics.

- Types of Continuous Probability Distribution
  - ✓ Normal Probability Distribution
  - ✓ Standard Normal Probability Distribution



Normal Probability Distribution

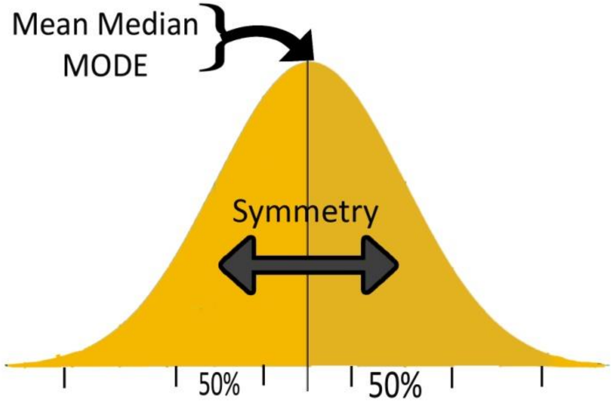
- There are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" this is also known as Bell Curve.
- A normal distribution is a very important statistical data distribution pattern occurring in many natural phenomena.
- For example, the bell curve is seen in Exam Results. The bulk of students will score the average(C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell (hence the nickname).
- In Corporate also, many of the times, HR uses Bell curve to evaluate the performance of the candidates.

- **Examples of Normal Distribution**
  - ✓ Heights of people
  - ✓ Size of things produced by machines
  - ✓ Errors in measurements
  - ✓ Blood pressure
  - ✓ Marks on a test

- Note: You can refer Normal Distribution, Standard Deviation and Empirical rule from chapter ……

Normal Probability Distribution

- This is important to understand if a distribution is normal, there are certain qualities that are consistent and help in quickly understanding the scores within the distribution.
- The Normal Distribution has:
  - ✓ Mean = Median = Mode
  - ✓ Symmetric about the center
  - ✓ 50% of values less than mean and 50% greater than mean



Normal Probability Distribution

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

Where,

P(x) = Normal Probability distribution function

x = Normal Random Variable

σ = Standard deviation

μ = Mean

e = Exponential constant = 2.71828

π = pi = 3.14 or 22/7

Example

- A company has 500 employees, salary of whom is normally distributed, with an average of Rs.40,000 and Standard deviation of Rs.6000. Suppose you pick a random employee from the 500 employees, what are chances he/she earns less than Rs.30,000

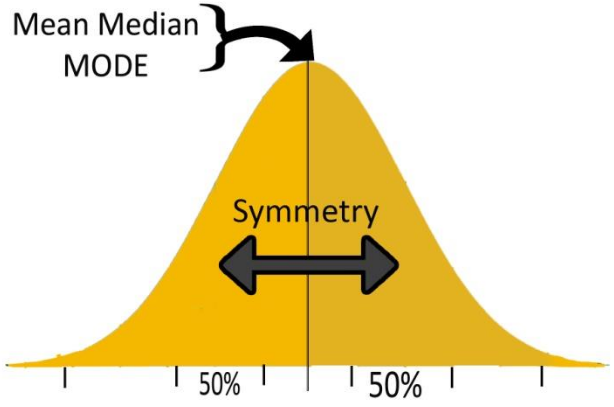
- The following information is available:

  Distribution: Normally distributed

  Mean: 40,000

  SD: Rs.6000

Make a bell curve showing mean and …..



Standard Scores or Z score

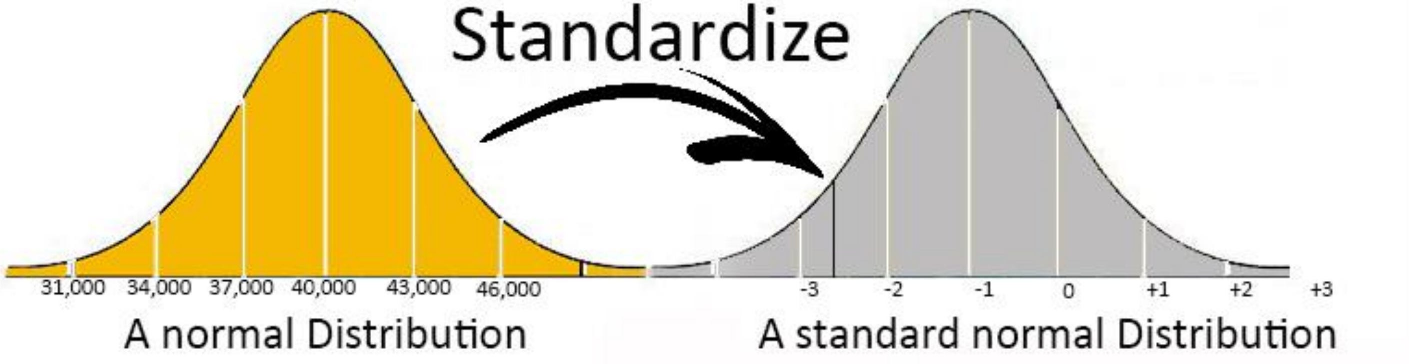
To find out the answer of previous questions, first of all we need to understand the standard scores or Z score

- The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score". Get used to those words!

$$z = \frac{x - \mu}{\sigma}$$

So to convert a value to a Standard Score ("z-score"):

- first subtract the observation from the mean: 30,000 – 40,000 = -10,000
- then divide by the Standard Deviation: -10,000/6000 = - 1.66
- And doing that is called "Standardizing":
- We can take any Normal Distribution and convert it to The Standard Normal Distribution.



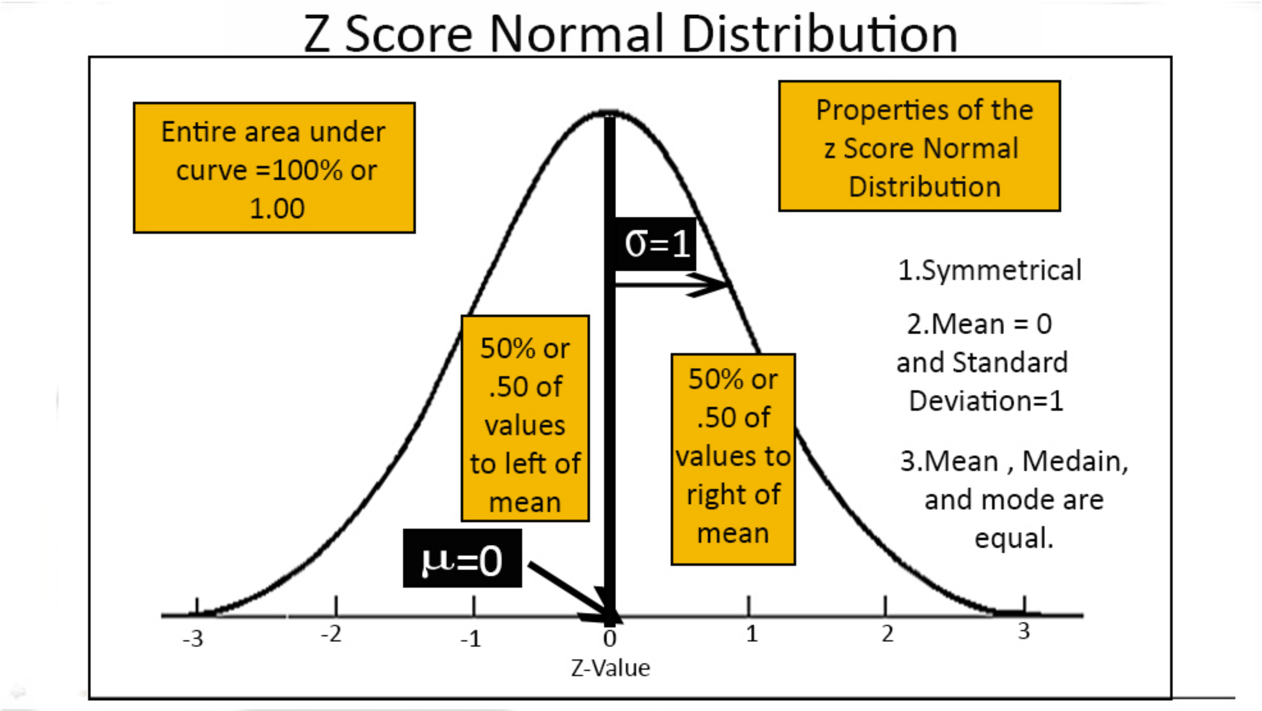
A normal Distribution

A standard normal Distribution

Z test

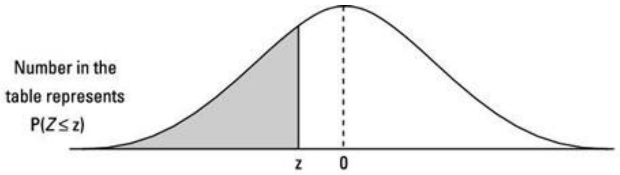
The area under the whole of a normal distribution curve is 1, or 100 percent. The z-table helps by telling us what percentage is under the curve at any particular point.

Now we need to use Z table to find the what percentage of is under the curve at any particular time

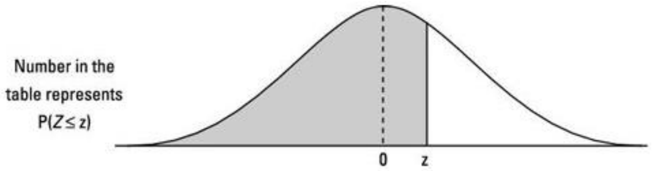


Z Score Normal Distribution
Entire area under curve =100% or 1.00
Properties of the z Score Normal Distribution
σ=1
1.Symmetrical
2.Mean = 0 and Standard Deviation=1
3.Mean , Medain, and mode are equal.
50% or .50 of values to left of mean
50% or .50 of values to right of mean
μ=0
-3
-2
-1
0
1
2
3
Z-Value

Z table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

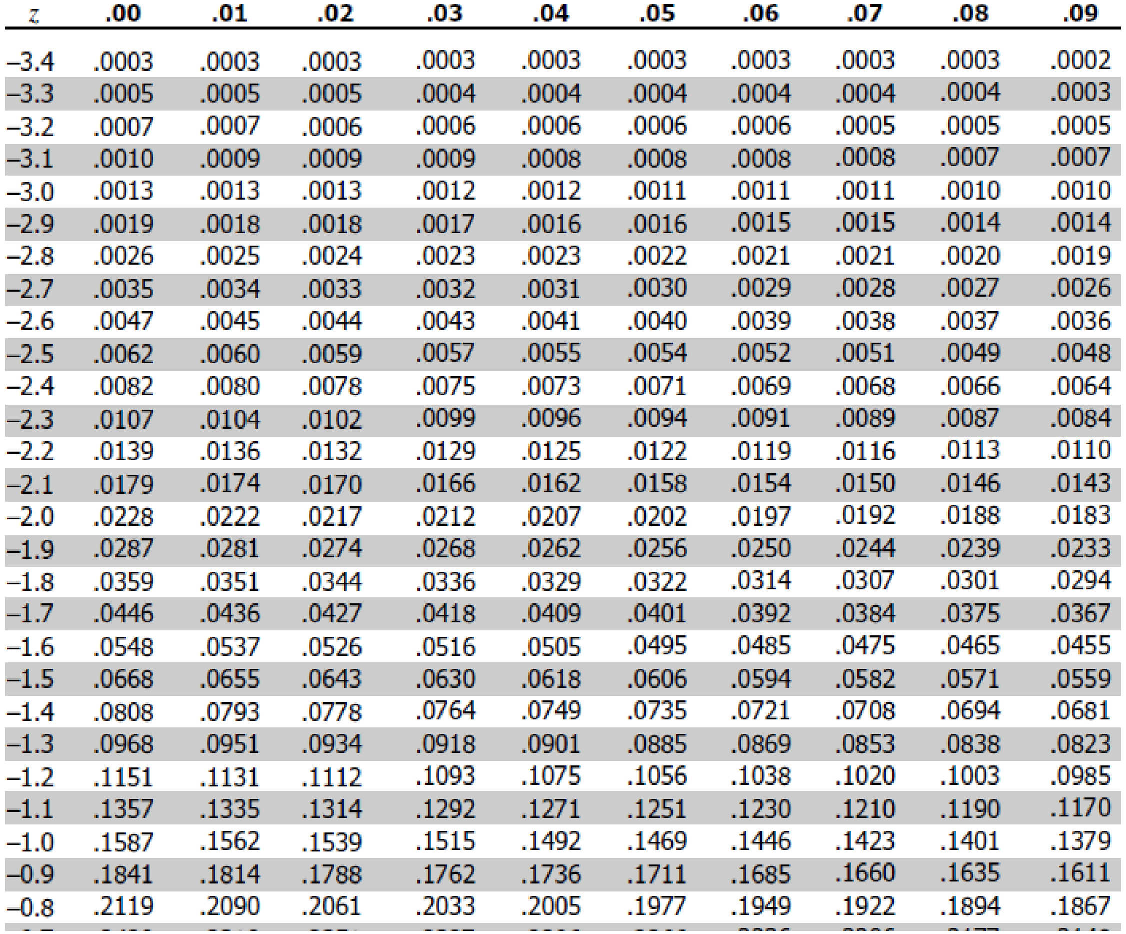
Z test

Since the z score was negative (-1.66), we need to use the negative Z- Scores

From the table, we can find out the probabilities of z score at -1.66

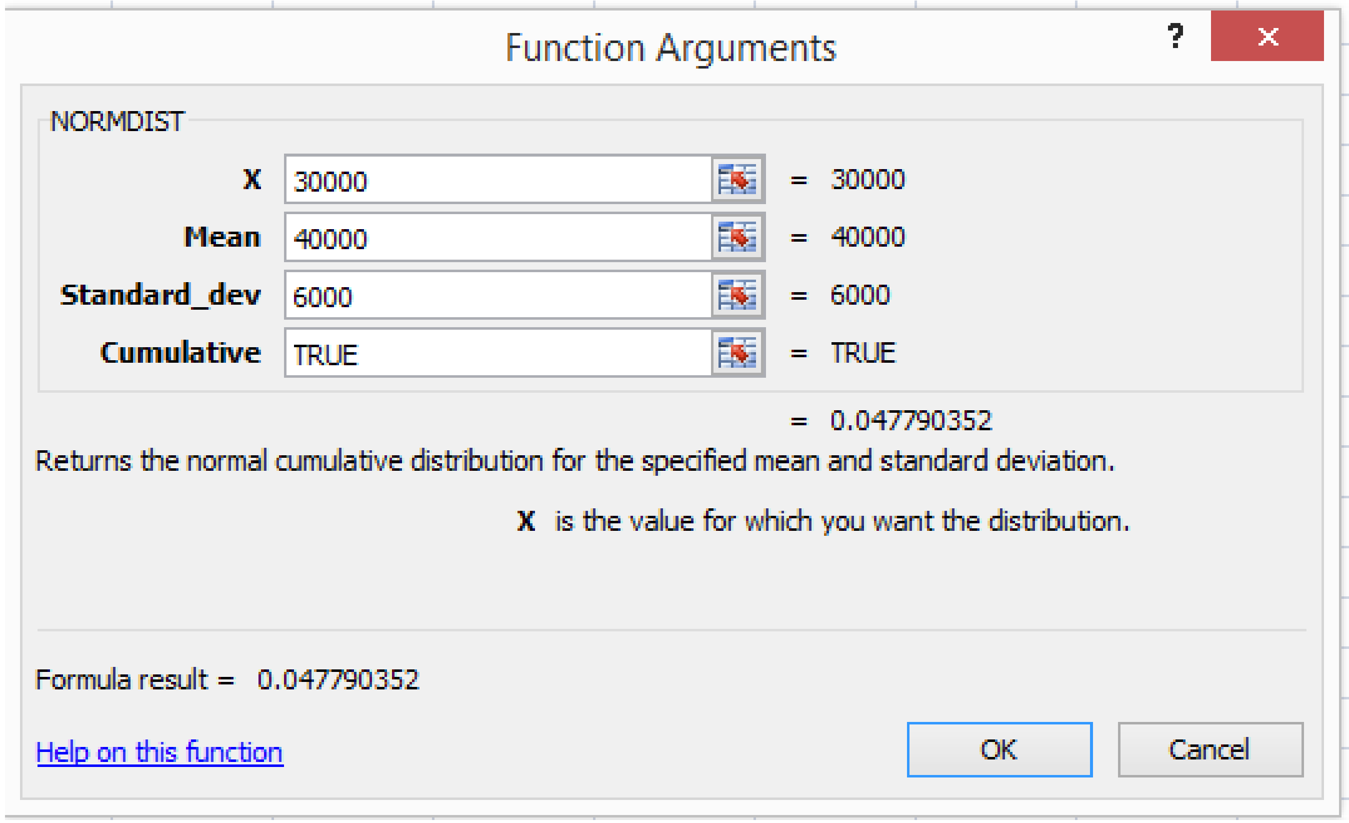
= 0.0485 0r 4.85%

It means that when we pick a random employee from the 500 employees, the chances he/she earns less than Rs.30,000 is 4.85%



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867

- Instead of using Z test table , we can also find the answer using excel function “normdist”



NORMDIST

X 30000 = 30000

Mean 40000 = 40000

Standard_dev 6000 = 6000

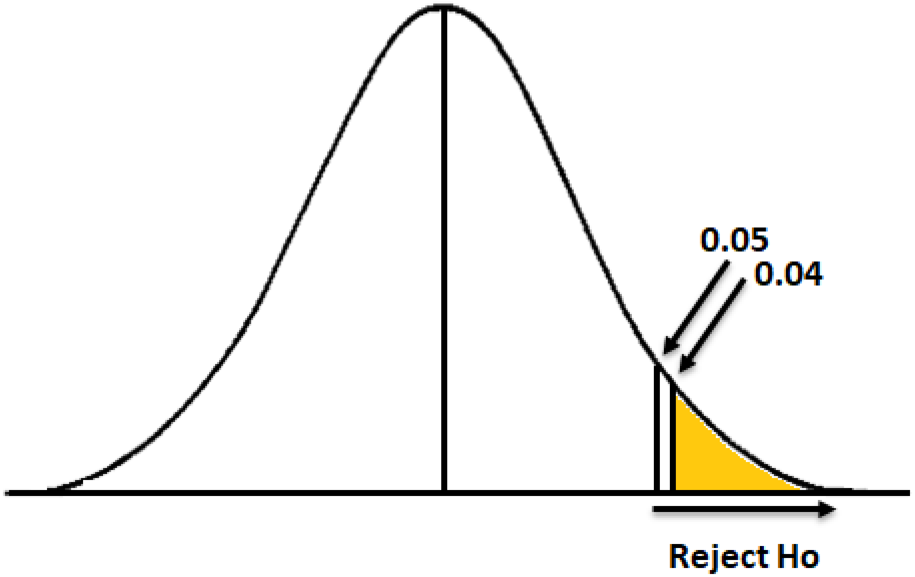
Cumulative TRUE = TRUE

= 0.047790352

Returns the normal cumulative distribution for the specified mean and standard deviation.

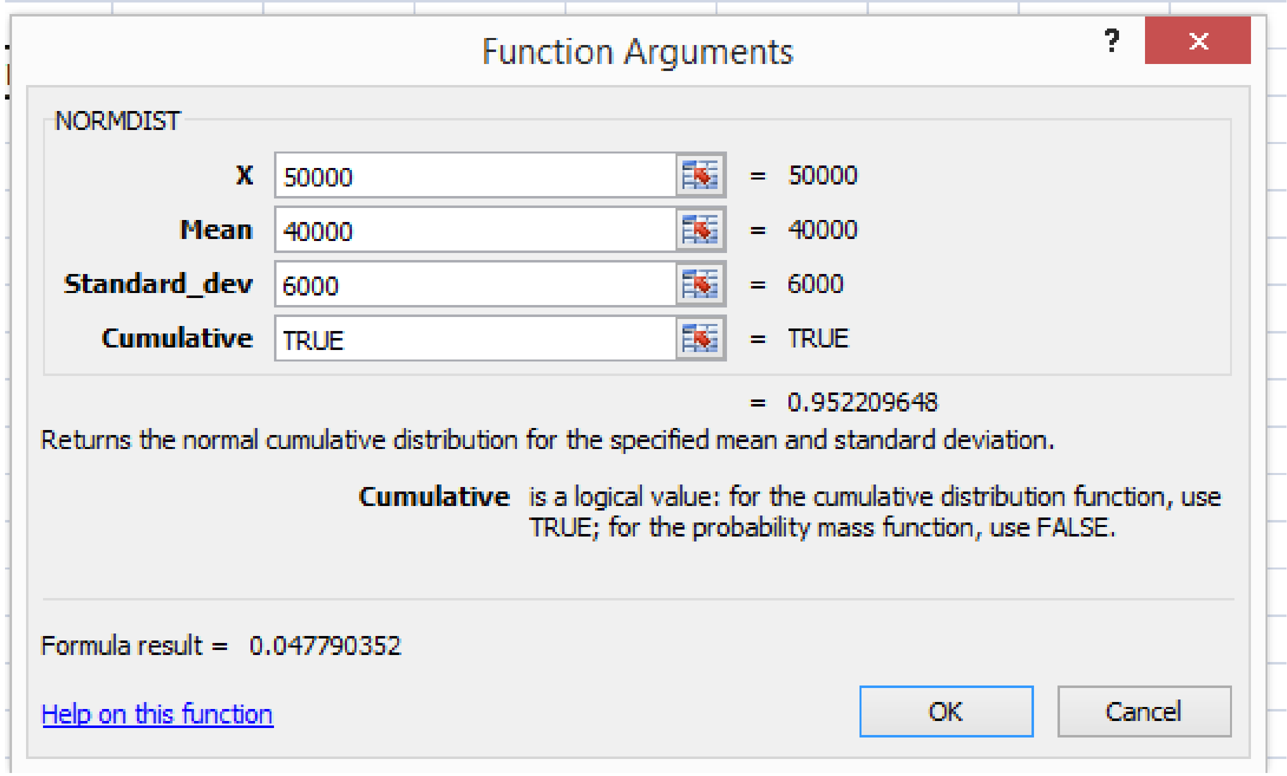
X is the value for which you want the distribution.

Formula result = 0.047790352



Note: You may get some variation in P value from Z table and P value from Excel

- What are chances he/she earns **more than Rs.50,000**



Function Arguments

NORMDIST

X 50000 = 50000

Mean 40000 = 40000

Standard_dev 6000 = 6000

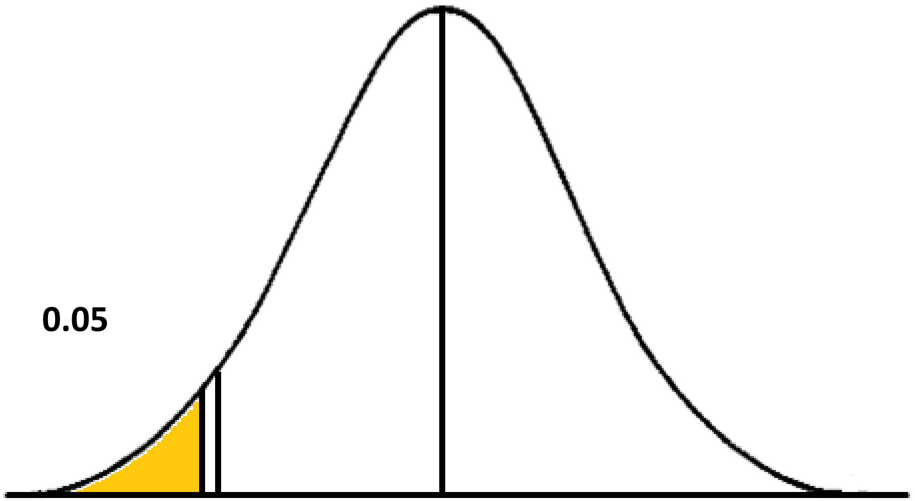
Cumulative TRUE = TRUE

= 0.952209648

Returns the normal cumulative distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.047790352

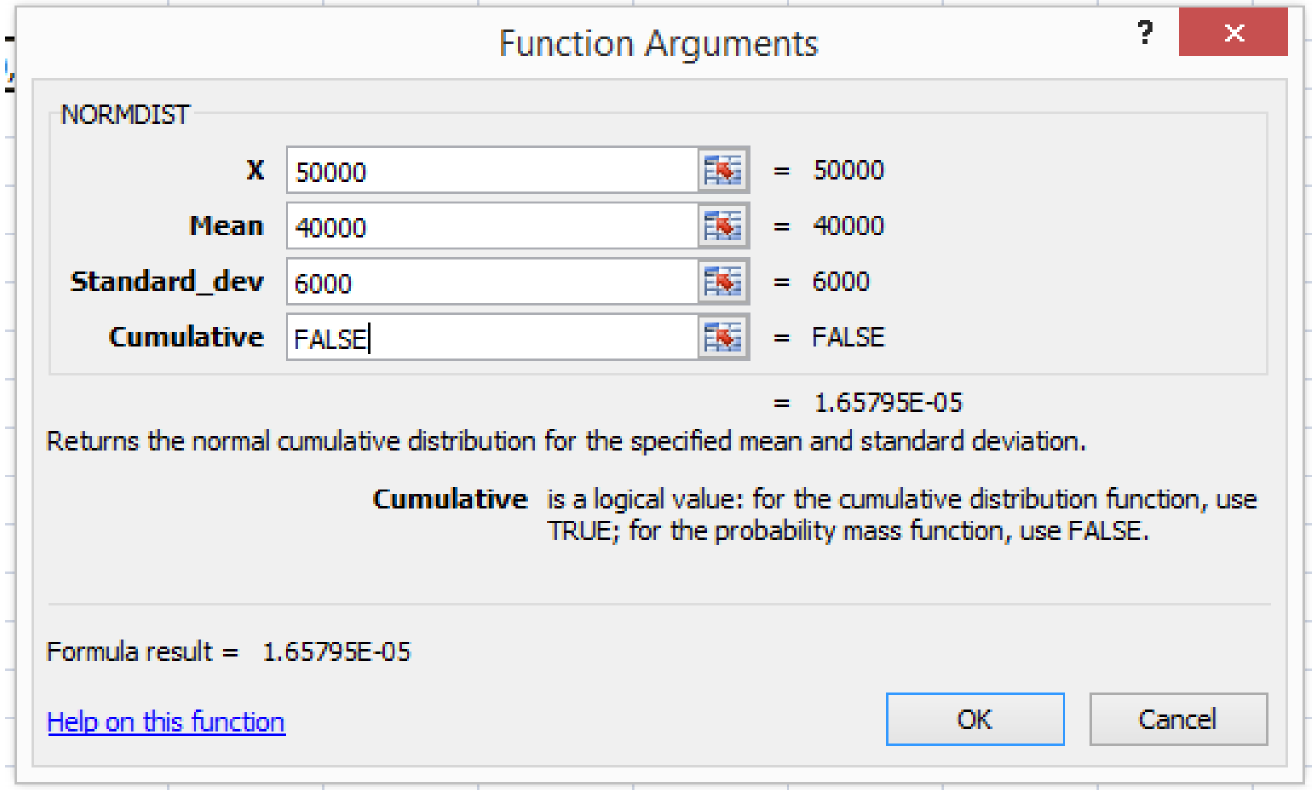


= 1-0.95= .05 or 5% or in the excel , you can directly use= 1-normdist(x, mean, standard_dev, cumulative)

- Even you can find out the chances he/she earns **exact Rs.50,000.**

- In this example, you have to use “false” in cumulative instead of “true”.

- Since the probability of finding exact Rs.50,000 is very low, the answer is around 0%



Function Arguments

NORMDIST

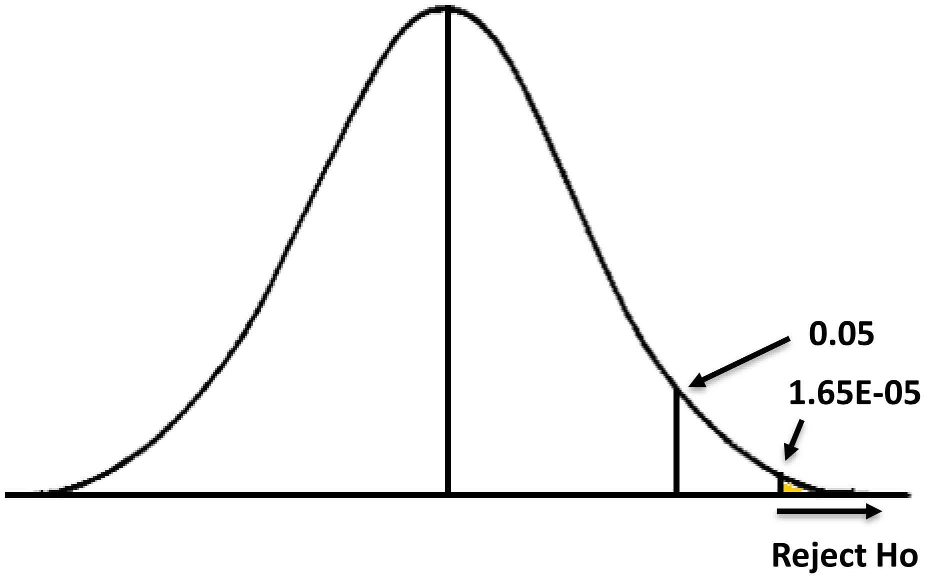
X	50000	= 50000
Mean	40000	= 40000
Standard_dev	6000	= 6000
Cumulative	FALSE	= FALSE

= 1.65795E-05

Returns the normal cumulative distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

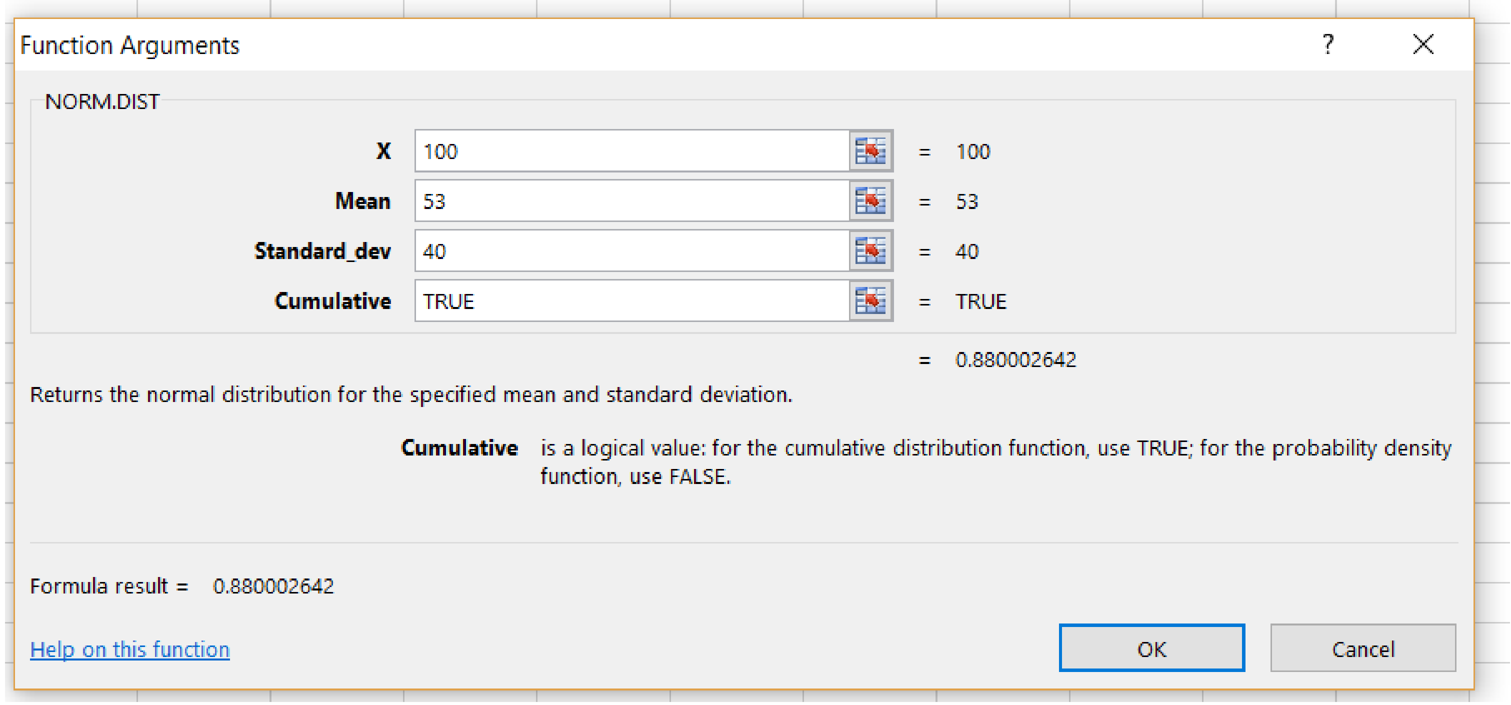
Formula result = 1.65795E-05



- Using the normdist you can also find the salary of employees between 50,000 and 70,000.

Example

The average score of Virat Kohli is 53 and the Standard Deviation is 40, assume the run scored by Kohli is normally distributed . Find out the probabilities of Kohli hitting 100 or more than 100 run in the next inning.
Solution:



Function Arguments

NORM.DIST

X 100 = 100

Mean 53 = 53

Standard_dev 40 = 40

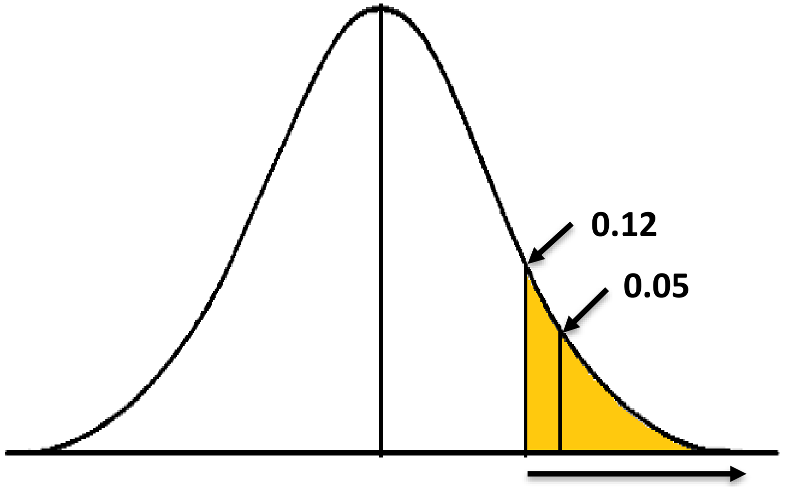
Cumulative TRUE = TRUE

= 0.880002642

Returns the normal distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability density function, use FALSE.

Formula result = 0.880002642



Reject Ho

The probabilities of kohli hitting more than 100 is 1 - 0.88 = 0.12

Exercise for Practice

Exercise for Basic
Stat 2

Thank You.