z-score

in below question if its asked about 70,80,90 etc means exact multiple of std we could use empirical formula but here 76 is asked so we have to use z-score method.

Q1. For a recent final written statistics exam for a "Data scientist" job selection process, the mean was 70 with a standard deviation of 10. If you scored 76 marks. What is your percentile or (area in the Normal distribution)?.

z=(x-x')/std

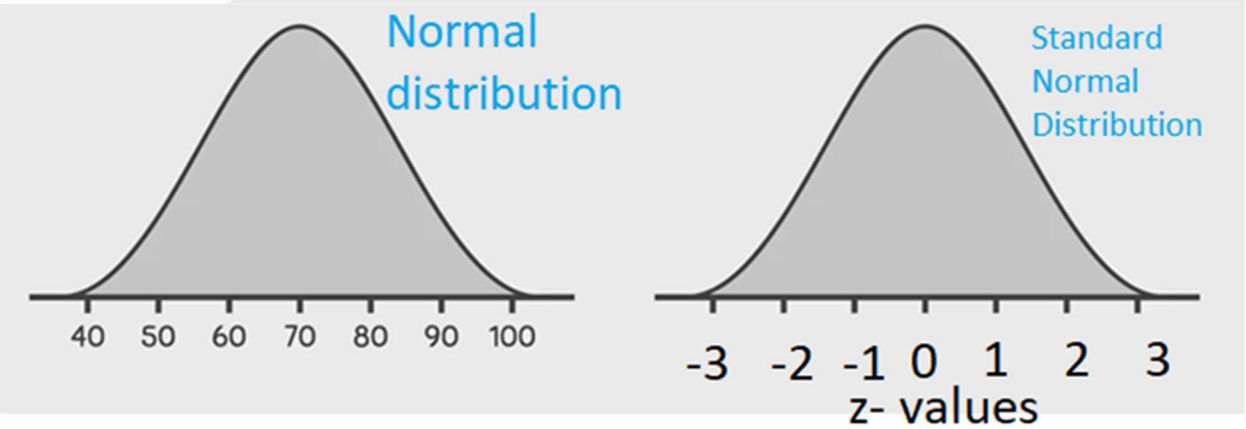
z= (76-70)/10
=.60

from z table .7257 is area under curve means prob means 72.57 %

Z-Score method basically convert data to the scale of std. its also a scaling method.

Normal distribution to standard Normal Distribution Conversion

$$Z = \frac{(observed\ value\ - mean)}{SD}$$



Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean.

- If the number of elements in the set is large, about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2 and about 99% have a z-score between -3 and 3.

Q1. For a recent final written statistics exam for a "Data scientist" job selection process, the mean was 70 with a standard deviation of 10. If you scored 76 marks. What is your percentile or (area in the Normal distribution)?.

Ans:-
Here, mean=70
SD=10.

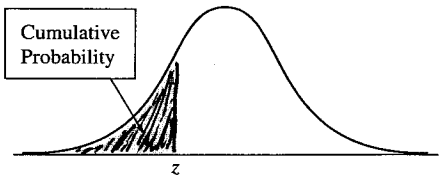
$$Z = \frac{(observed\ value\ - mean)}{SD}$$

Mean value 70 Z score = (70-70)/10=0
80 marks Z score = (80-70)/10=1
60 marks Z score = (60-70)/10=-1
76 marks Z score =(76-70)/10=0.60

In the z table the value of 0.60 is 0.7257.

This is the value of area under curve or the percentile.

APPENDIX A

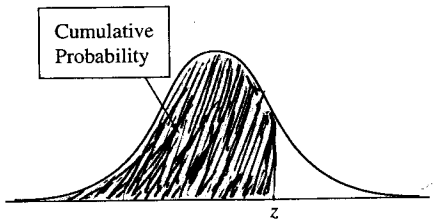


Cumulative probability for z is the area under the standard normal curve to the left of z

z	.00
−5.0	.000000287
−4.5	.00000340
−4.0	.0000317
−3.5	.000233

TABLE A Standard Normal Cumulative Probabilities

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
−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



Cumulative probability for z is the area under the standard normal curve to the left of z

TABLE A Standard Normal Cumulative Probabilities *(continued)*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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

z	.00
3.5	.999767
4.0	.9999683
4.5	.9999966
5.0	.999999713

z-score

A normal curve is determined by mean and SD . If the data follow the normal curve , than knowing mean and SD means knowing the whole histogram.

To compute area under the normal curve , we first standardize the data using Z score formula.

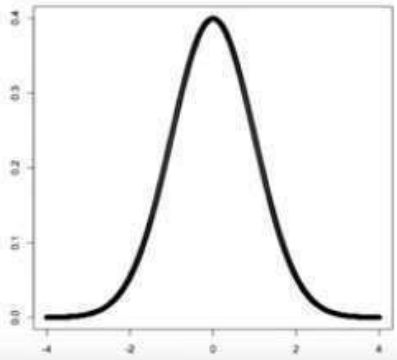
$$Z = \frac{(observed\ value\ - mean)}{SD}$$

Here Z is called the standardized value or Z scaore.

Z has no unit (the observed value, mean,SD have the unit for example meter , inches etc.)

When we standarized the data we have mean 0 and standard deviation equal to 1 →this is the point of stansardization.

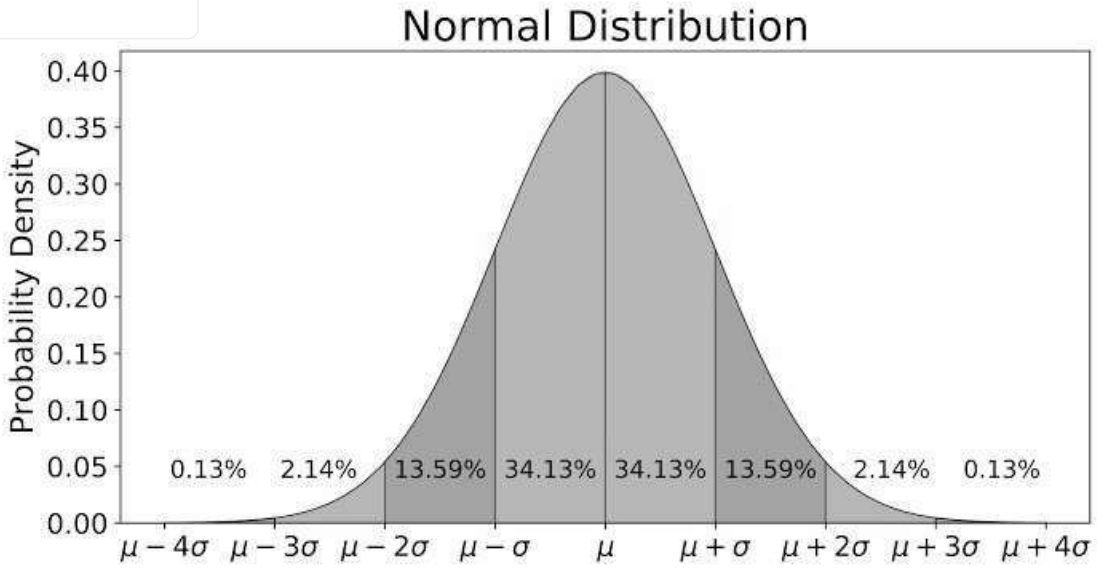
We can convert any normal distribution into the standard normal distribution



This curve is given by the function

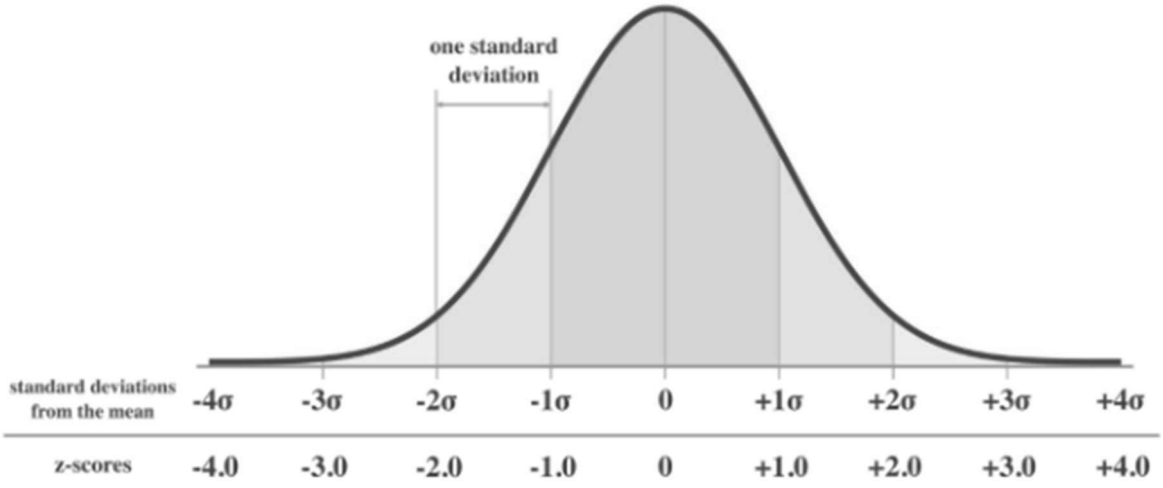
$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

No need to Remember this formula it is just for an understanding.



Standard Normal Distribution

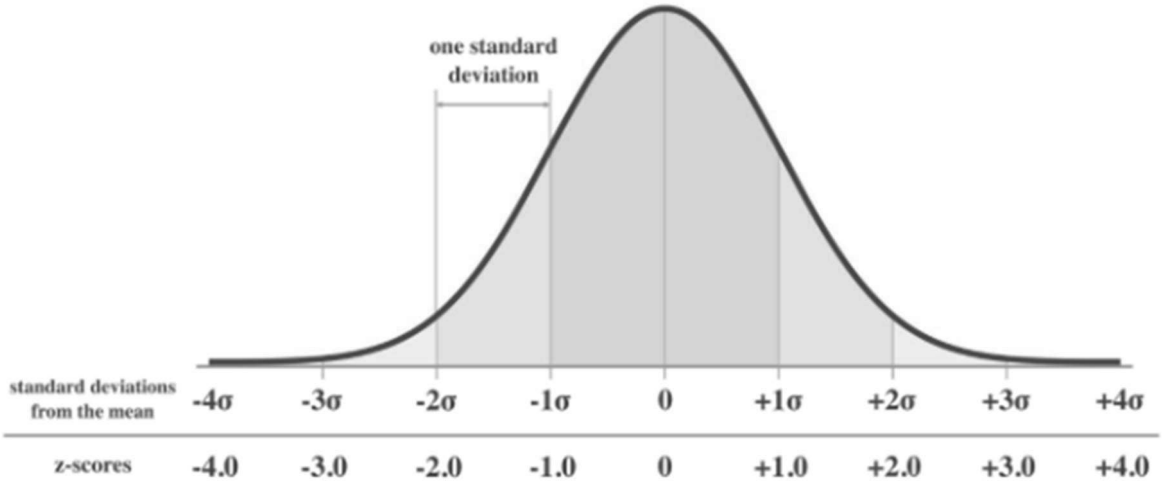
The (z-value/ z-score / z / standard score) represents the number of standard deviations an observation is from the mean for a set of data.



Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the mean μ and also the standard deviation σ.

- **Z-scores can be positive or negative.**

- The sign tells you whether the observation is above or below the mean.
- z-score of +2 indicates that the data point two standard deviations above the mean,
- z-score of -2 signifies it is two standard deviations below the mean.



$$Z = \frac{(observed\ value\ - mean)}{SD}$$

we can use the z-value.

```
import scipy.stats as stats

stats.zscore(data=df, axis=1)

or
df.apply(stats.zscore)
```

Normal approximation:-

Finding areas under the normal curve is called normal approximation.

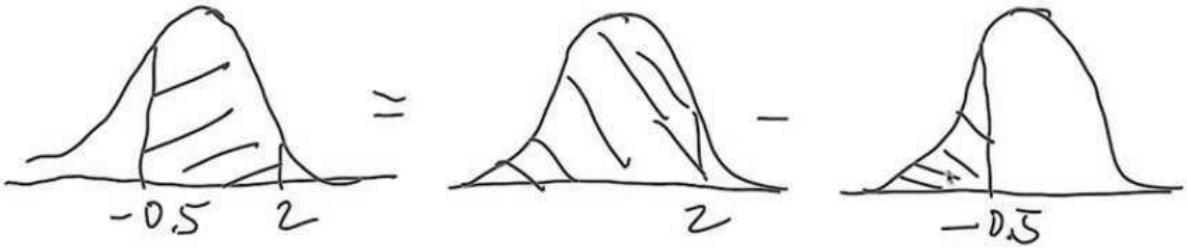
Q. fathers' heights follow the normal curve with a mean of 68.3 inches and a standard deviation of 1.8 inches.
What percentage of fathers have heights between 67.4 inch and 71.9 inch?

1. Standardize:

$$\frac{67.4 \text{ in} - 68.3 \text{ in}}{1.8 \text{ in}} = -0.5 \qquad \frac{71.9 \text{ in} - 68.3 \text{ in}}{1.8 \text{ in}} = 2$$

desired area in a form that can be computed by looked up in a table:

Typically we can look up the area to the left of a given value.



Use a table to find these values: $97.7\% - 30.9\% = 66.8\%$

Q. What is the 30 percentile of the Father's height?

Ans:-

value close to .3000 in z table is .3015 which is -0.52

so z=-0.52
z=(x-x')/std==> -0.52=x-68.3/1.8=> x=67.3

Q. One student score 80 marks in Mathematics and 75 Marks in English.
At this point we can say he has performed excellent in Math as compare to English.
Consider the average class score of Mathematics is 90, SD is 5.
Average class score of English is 60 , SD is 5.
Verify the performance?

Ans:- lets calculate the Z values
Zm=(80-90)/5= -2
Ze=(75-60)/5=3

Z score value is -3 to +3
if it is close to -3 it means lower performance
if z score close to +3 means excellent performance.

By the use of Z-table we can find the Area under the curve / percentile value.

Q3. If you know that -0.1 corresponds to approximately 46 % and 1.8 corresponds to approximately 96.4 % (both percentages are areas under the curve to the left of the value), what percentage of fathers will have heights between 68.1 inc and 71.5 in?

Ans:-

Q5. What proportion of students are between 5.81 feet & 6.3 feet height. Given Mean=5.5, sd=0.5 feet.

Q6. mean height of Gurkhas is 146 cm with Sd of 3 cm . what is the probability of

(a)Height having greater than 152 cm.

(b)Height between 140 and 150 cm.

Q7. Mean demand of an oil is 1000 ltr per month with SD Of 250 ltr.

(a)if 1200 ltrs are stocked , what is the satisfaction level?

(b)For an assurance of 95%. what stock must be kept?

Population & Sample

Researchers are often interested in answering questions about the population like:

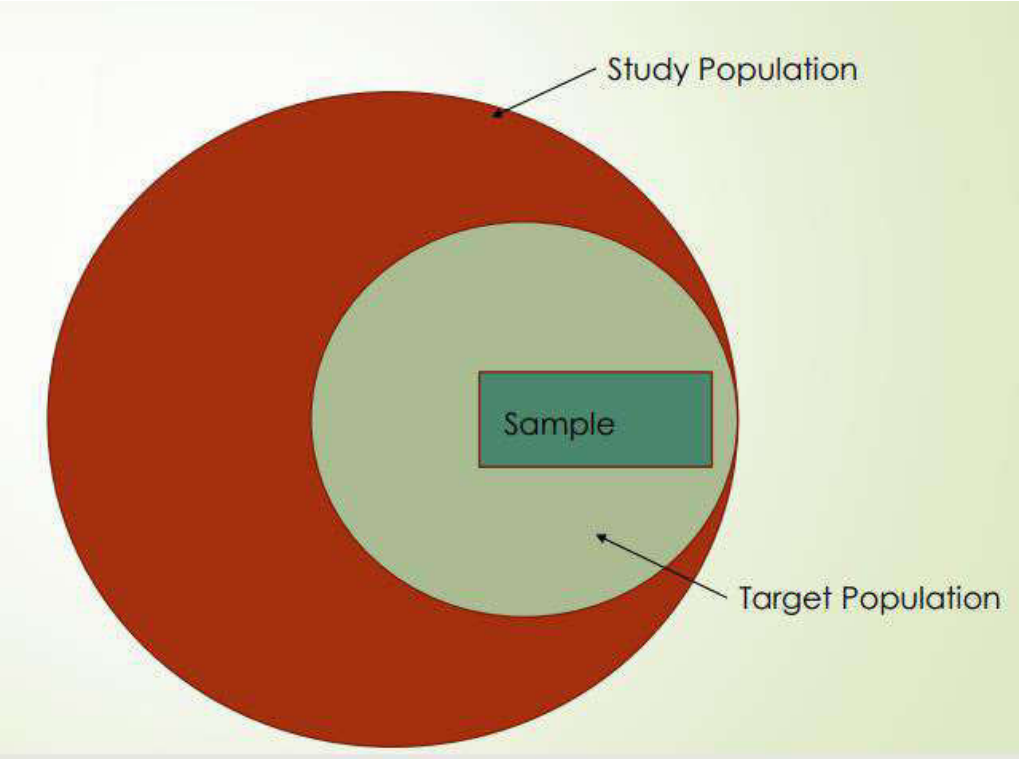
- Which political party will win the election?
- What will be the average cutoff mark in the competitive examination?
- What percentage of citizens in a certain city support a certain law?
- What is the effect of certain medicine on the population of disease patients?

One way to answer these questions is to go around and collect data on every single individual in the population of interest.

However, this is typically too costly and time-consuming which is why researchers instead take a **sample** of the population and use the data from the sample to draw conclusions about the population as a whole.

- ➔ **A population is the entire group that you want to draw conclusions about.**
- ➔ **A sample is the specific group that you will collect data from.**

A well-chosen sample will contain most of the information about a particular population parameter.

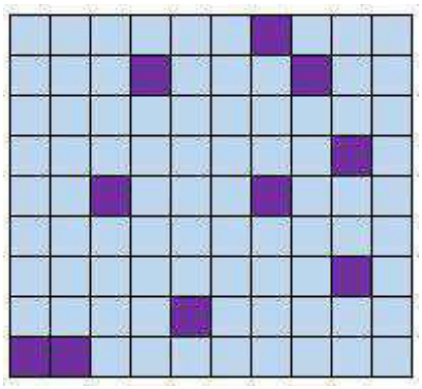


I. Probability Sampling Methods

in prob sampling methods prob of selecting a individual from pop as a sample is same for each datapoint.

Type of Probability Sampling Methods: -

1. Simple random sampling



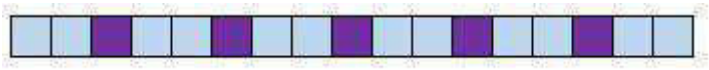
Example: Simple random sampling

You want to select a simple random sample of 10 employees of Company X. You assign a number to every employee in the company database from 1 to 100, and use a random number generator to select 100 numbers.

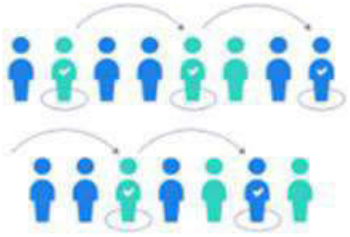
The output may be 2,4,8,12,29,33,45,55,87,99

Definition: Every member of a population has an equal chance of being selected to be in the sample.

2. **Systematic random sample** is similar to simple random sampling, but instead of randomly generating numbers, individuals are chosen at regular intervals.



Definition: Put every member of a population into some order. Choosing a random starting point and select every n^{th} member to be in the sample.



Example: Systematic sampling

One commonly used sampling method is **systematic sampling**, which is implemented with a simple two-step process:

1. Place each member of a population in some order.

2. Choose a random starting point and select every n^{th} member to be in the sample. **If 10th person on the list is selected (6, 16, 26, 36, and so on)**

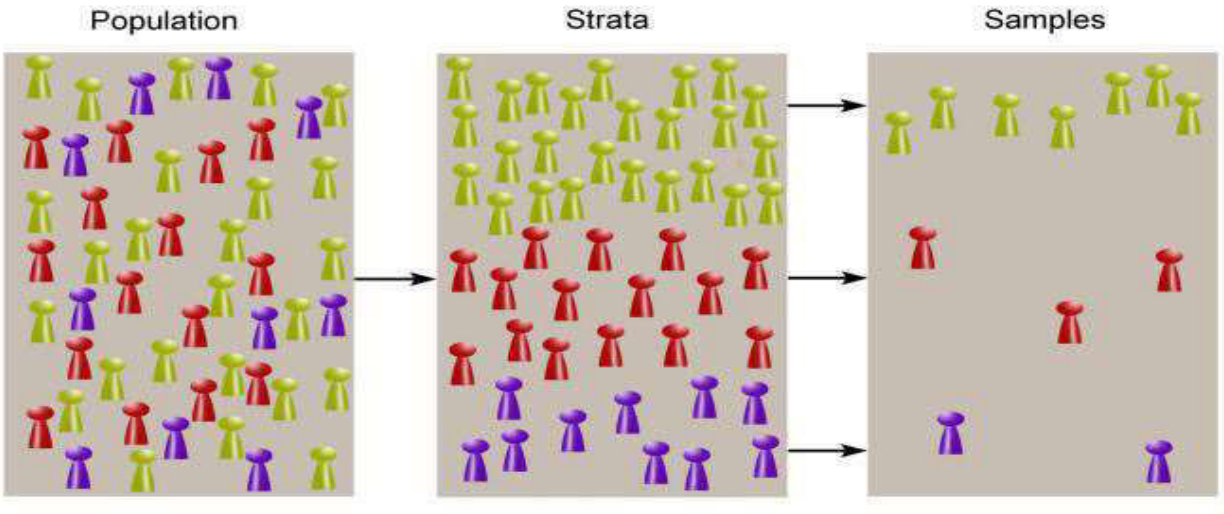
It is important to make sure that there is no hidden pattern in the list.

3. Stratified random sample

for classification problem

if we have imbalance data set we opt statified sampling so we can replicate the pop pattern in sample also. saapose a data set have 8:2 male female imbalance and we need to create a sample with same pattern so we opt stratified.

Split a population into groups. Randomly select some members from each group to be in the sample.



Divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).

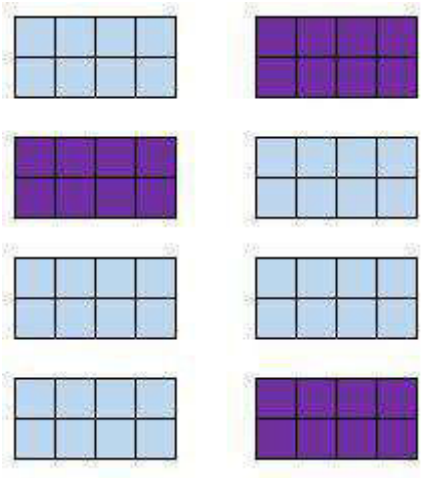
Then can use random or systematic sampling to select a sample from each subgroup.

Example: Stratified sampling

Benefit: Stratified random samples ensure that members from each group in the population are included in the survey.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.



Example: A company that gives **whale watching tours**, wants to survey its customers. Out of ten tours they give one day, they randomly select four tours and ask every customer about their experience.

Benefit: Cluster random samples get every member from some of the groups, which is useful when each group is reflective of the population as a whole.

Difference between stratified sampling and cluster sampling: -

- **In cluster sampling it is not mandatory to select from all the sub groups. We can select from our choice either from 2 group of from more or all groups.**
- **In stratified sampling it is mandatory to select from all the sub-groups.**

II Non-probability Sampling Methods

Not every member in a population has an equal probability of being selected to be in the sample.

This type of sampling method is sometimes used because it's **much cheaper and more convenient** compared to probability sampling methods.

It's often used when researchers simply want to gain an initial understanding of a population.

However, that result from these sampling methods cannot be used to draw inferences about the populations they came from because they typically aren't representative of the entire population.

Type of Non Probability sampling:-

(1)Convenience sample

Example: A researcher stands in front of a Shopping Mall during the day and polls people.

Drawback: Location and time of day will affect the results. More than likely, the sample will suffer from under coverage bias since certain people (e.g. those who work during the day) will not be represented as much in the sample.

(2) Voluntary response sample

Definition: A researcher puts out a request for volunteers to be included in a study and members of a population voluntarily decide to be included in the sample or not.

Example: A radio host asks listeners to go online and take a survey on his website.

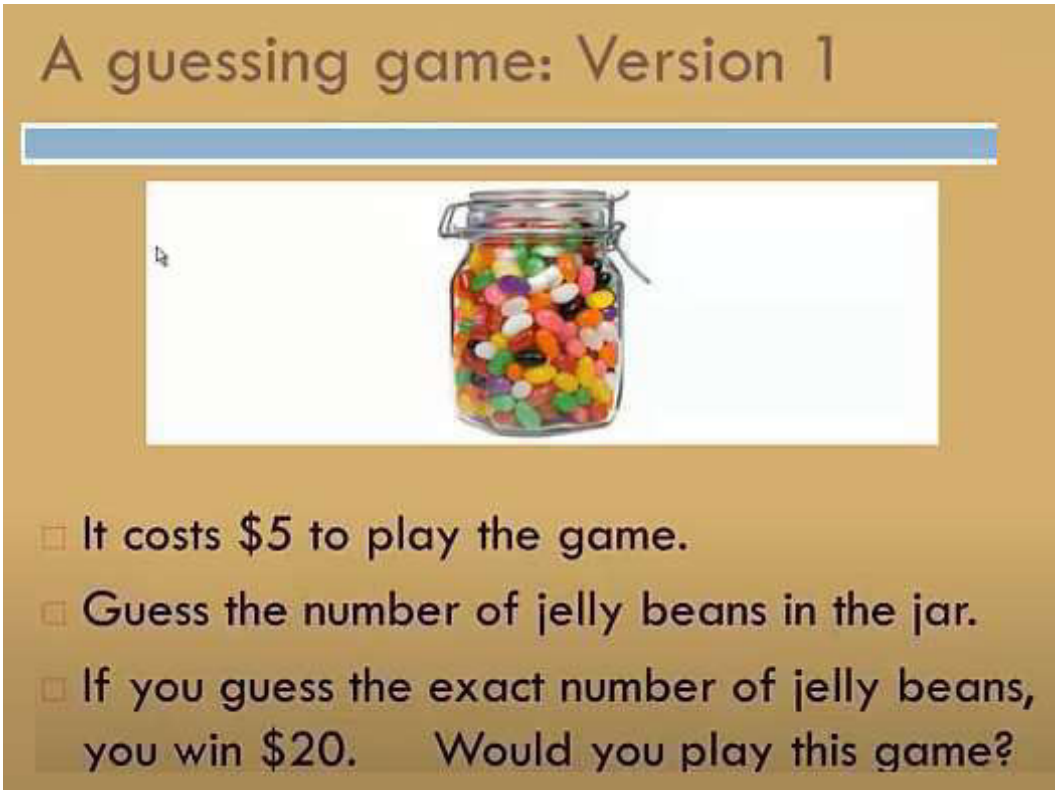
Drawback: People who voluntarily respond will likely have stronger opinions (positive or negative) than the rest of the population, which makes them an unrepresentative sample.

Using this sampling method, the sample is likely to suffer from nonresponsive bias – certain groups of people are simply less likely to provide responses.

Which sampling method is best?

The method of sampling best to use will depend on the nature of the analysis and the data being used. In general, simple random sampling is widely used, but in specific research application stratified sampling can produce a more accurate sample relative to the population under study.

Confidence Level and Margin of Error

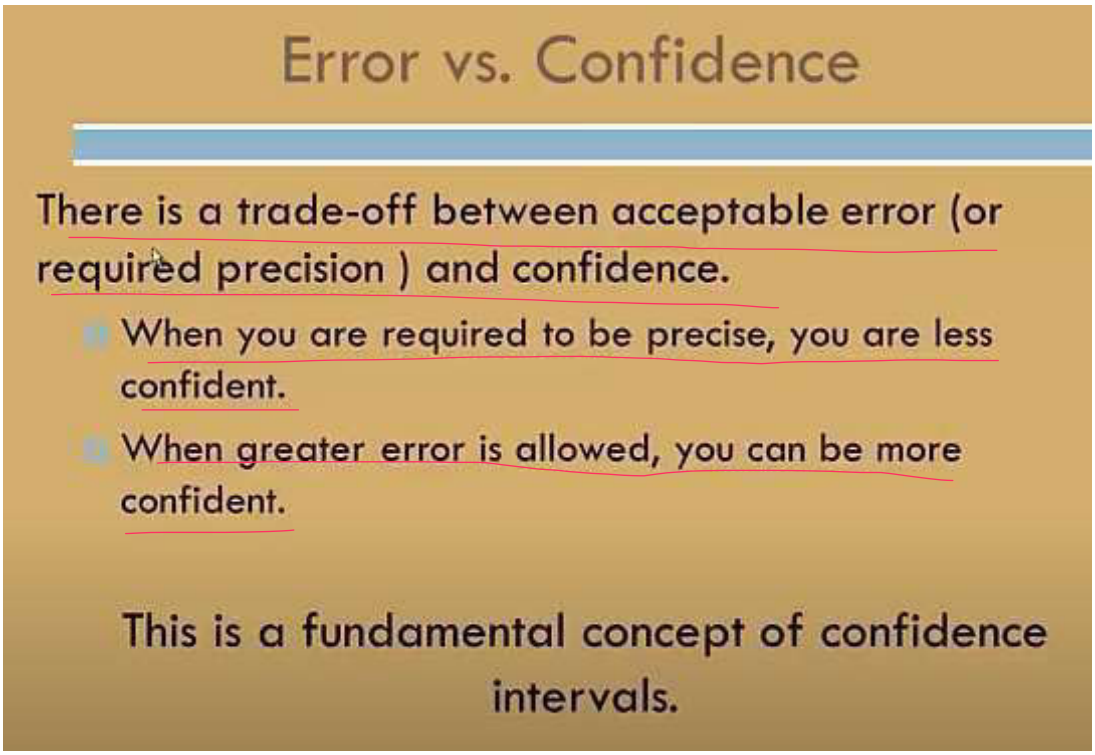


Option 1:- You Guess within 5 no. of actual no. of jelly.

Option 2:- You Guess within 25 no. of actual no. of jelly.

Option 3:- You Guess within 50 no. of actual no. of jelly.

Option 4:- You Guess within 100 no. of actual no. of jelly.



An increase in confidence level results in the increase in the margin of error. We have to choose between the precision and confidence.

we can also say increase in confidence interval will result in confidence level and margin of error. for example if i have the freedom to guess between 245-255 (CI) then assume i have CL of 70% and then game allow me to increase CI to 240 - 260 so my CI increases so my CL increases to 80%, also my margin of error also increases

Estimating the Population Mean

Estimation of sample size

Confidence Level & Confidence Interval: -

Statistical Inference: - The purpose of collecting a sample. Researchers take the sample in order to infer from that data some conclusion about the wider population represented by the sample.

If several random samples were collected, the mean for that variable would be slightly different from one sample to another. Therefore, when researchers estimate population means, instead of providing only one value, they specify a range of values (or an interval) within which this mean is likely to be located.

"A **confidence interval** is the range of values, derived from sample statistics, that is likely to contain the value of an unknown population parameter."

Confidence interval=**Point estimate +/- margin of error.**

Point estimate: - the value of any statistic that estimate the value of a parameter is called Point estimate. preceise value than a range, for example sample mean is point estimate where as population mean is CI

Confidence interval = sample mean ± margin of error

Imagine that a brick maker is concerned whether the mass of bricks he manufactures is in line with specifications. He has measured the average mass of a sample of 100 bricks to be equal to 3 kg.

He has also found the with a **confidence level** of 95% , brick mass confidence interval to be between 2.85 kg and 3.15 kg.

It means that he can be 95% sure that the average mass of all the bricks he manufactures will lie between 2.85 kg and 3.15 kg.

Margin of error: -

The margin of Error used to determine the by which the sample result arrived will differ from the value of the entire population.

A higher margin of error indicates a high chance that the result of the sample may not be the true reflection of the whole population.

Margin of Error = Z * σ / √n

we can see that from the equation, if we increases the sample size margin of error will decreases.

Value from z table

Z =critical factor (for a confidence level)

σ = Population standard deviation

In reality we use sample std in this equation coz pop std is unknow unless its given.

n = No. of observation in the sample.

Here SD of the dataset (Population)

$$\sigma = \sqrt{\sum_{i}^{n} (x_i - \mu)^2 / N}$$

A higher margin indicates that the survey results may stray from the actual views of the total population.

On the other hand, a smaller margin indicates that the results are close to the true reflection of the total population, which builds more confidence in the survey.

Example:- 900 students who were part of a survey, and it was found that the average GPA of the population was 2.7, with a population standard deviation of 0.4. Calculate the margin of error for

90% confidence level
95% confidence level
98% confidence level
99% confidence level

n=900
mue=2.7
pop std=0.4

mor=z*popstd/sqrt(n)
z score of 90% ==>1.645
mor=(1.645*.4)/sqrt(900)=0.0219
z score of 95% ==>1.96
mor=(1.96*.4)/sqrt(900)=0.0261
z score of 98% ==>2.33
mor=(2.33*.4)/sqrt(900)=0.0310
z score of 99% ==>2.58
mor=(2.58*.4)/sqrt(900)=0.0344

Hint: -

For a different confidence level, the critical factor or z-value is :-

Area under curve

Z score

90% CL ---- z value = 1.645
95% CL ---- z value = 1.96
98% CL ---- z value = 2.33
99% CL ---- z value = 2.58

For a 90% Confidence Level

the critical factor or z-value is 1.645 i.e. z = 1.645

Therefore, the margin of error at a 90% confidence level can be made using above the formula as,

ME= 1.645 * 0.4 / √900=0.0219

For a 95% Confidence Level

ME=1.96 * 0.4 / √900=0.0261

For a 98% confidence level

ME= 2.33 * 0.4 / √900 =00.0311

For a 99% confidence level

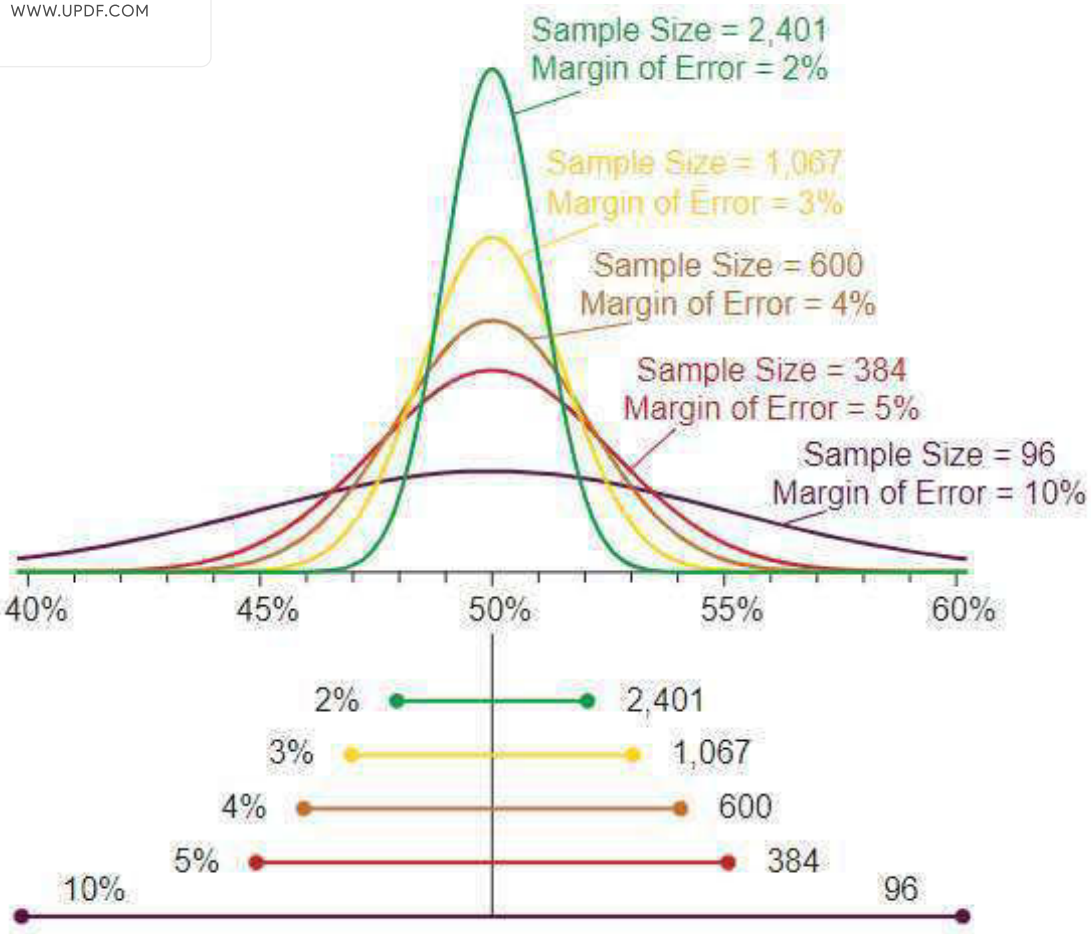
ME=2.58 * 0.4 / √900=0.0344

Consequently, it can be seen that the error of a sample increases with the increase in confidence level.

Calculating Sample Size

- To determine a sample size that will provide the most meaningful results, researchers first determine the preferred margin of error (ME) or the maximum amount they want the results to deviate from the statistical mean.
- It's usually expressed as a percentage, as in plus or minus 5 percent. Researchers also need a confidence level, which they determine before beginning the study.
- This number corresponds to a Z-score, which can be obtained from tables. Common confidence levels are 90 percent, 95 percent and 99 percent, corresponding to Z-scores of 1.645, 1.96 and 2.576 respectively.
- Researchers express the expected standard of deviation (SD) in the results. For a new study, it's common to choose 0.5.

we can see as sample increases moe decreases and CL decreases.



Having determined the margin of error, Z-score and standard deviation, researchers can calculate the ideal sample size by using the following formula:

$$(\text{Z-score})^2 \times \text{SD} \times (1\text{-SD})/\text{ME}^2 = \text{Sample Size}$$

This formula does not depend on the size of the population, only on the size of the sample.

Effects of Small Sample Size

- In the formula, the sample size is directly proportional to Z-score and inversely proportional to the margin of error.
- Consequently, reducing the sample size reduces the confidence level of the study, which is related to the Z-score.
- Decreasing the sample size also increases the margin of error.
- In short, when researchers are constrained to a small sample size for economic or logistical reasons, they may have to settle for less conclusive results.
- Whether or not this is an important issue depends ultimately on the size of the effect they are studying.

Mithilesh Singh

Summary

Confidence interval CI=Point estimate (M) +/- Margin of error.

M=point estimate or samples mean.

confidence interval (CI) roughly speaking, the range of scores (upper and lower value) that is likely to include the true population mean.

Estimating the Population Mean When It Is Unknown: - the best estimate of the population mean is the sample mean.

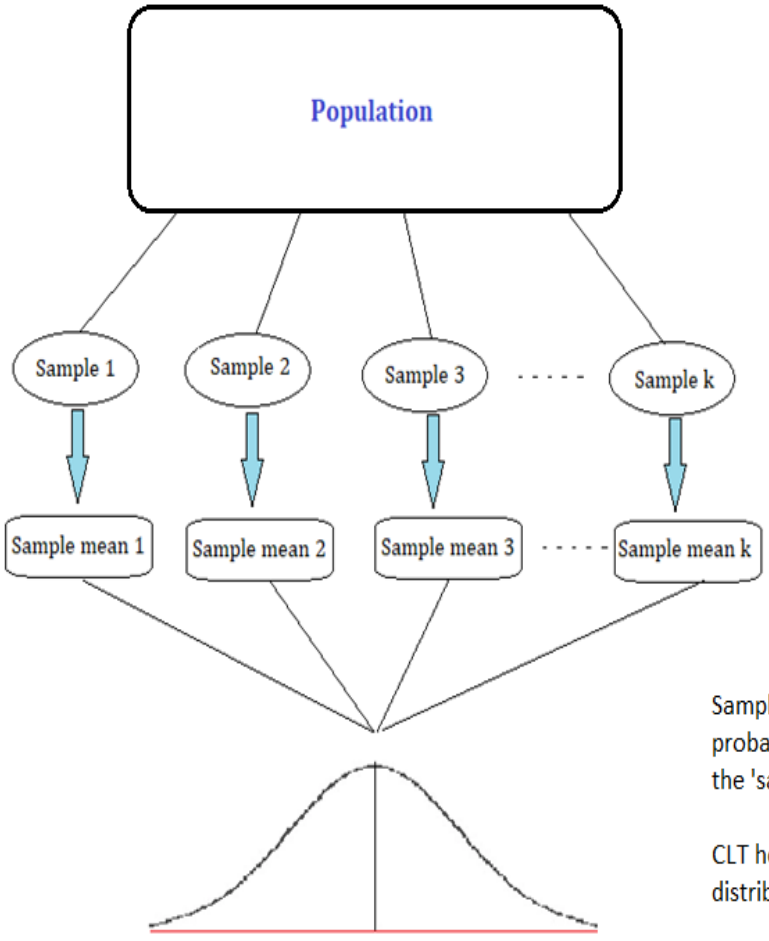
Basically, confidence intervals tell the range of true population mean.

A 95% confidence interval: - We are 95% likely to come from a population with a true mean that happens to be the mean of our sample.

Central Limit Theorem :- The distribution of means (mean of various sample means) always follows normal distribution.

The Central Limit Theorem (CLT) is a statistical theory that indicate -- the mean and standard deviation derived from a sample, will accurately approximate the mean and standard deviation of the population.

law of large number: - The minimum number of members in the sample is 30 or more than 30. **The sample size increases the mean of the sample is close to the mean of the population.**



for each at least sample size should be 30

Sample mean is a random variable itself and thus has a probability distribution like any other random variable called the 'sampling distribution'!

CLT helps us to figure out the parameters of the sampling distribution