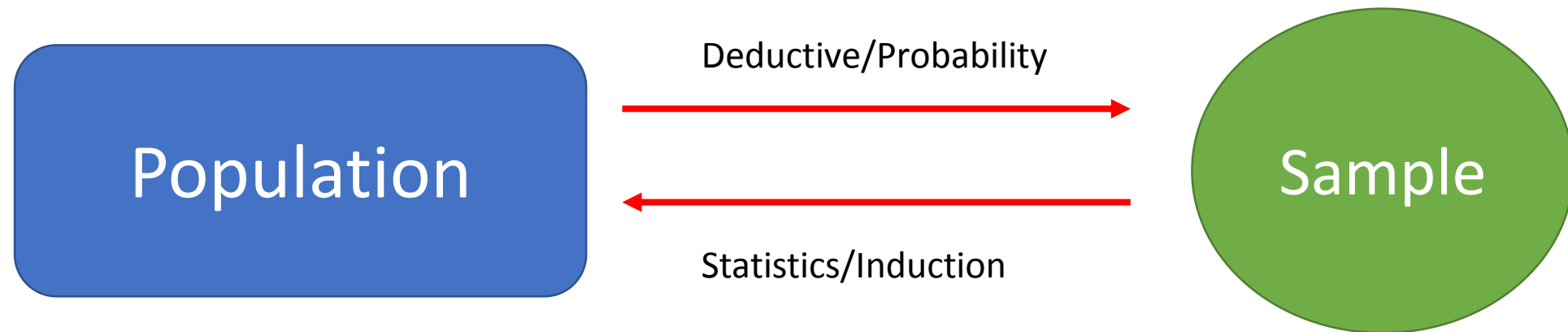


# Statistical Estimation

# Introduction

- Estimation is a technique for calculating information about a bigger group from a smaller sample, and statistics are crucial to analyzing data.
  - *educated guesses*
  - *Characteristics – mean, variance, proportions*
  
- *Deductive vs Inductive Reasoning*

# Inductive Vs Deductive Reasoning



# Parameter

- Population – described by probability distribution &/or parameters
  - Quantitative population – Mean and standard deviation
  - Binomial distribution – parameter ‘p’

----- What if the value for the parameter is unknown – ‘inferential statistics’

# Purpose of Estimation

- Statistical estimation is essential for making **inferences** about populations using sample data, helping to determine parameters like mean and variance without individual measurements.
- This evaluation is vital for **decision-making** in business and healthcare, informing strategies and treatment options.
- It is closely linked to **hypothesis testing**, contributing to scientific development, political decisions, public health, and economic choices.
- **Risk assessment** benefits from evaluation in managing probabilities and risk in finance and insurance.
- **Quality control** also relies on evaluation to ensure products and services meet standards by identifying and correcting deviations.

# Inferential statistics

The part of statistics that allows researchers to generalize their findings to a larger population beyond data from the sample collected.

It mainly consists of two parts:

- Estimation – involves the use of data in the sample to calculate the corresponding parameter in the population from which the sample was drawn
- Testing of Hypothesis

# Types of Estimation

## Point Estimation

- Identifying a single number to represent a large group is like a point estimate.
- Population mean is estimated using the sample mean
- Other attributes like percentages of specific characteristics in a population.
- not always precise - offer good understanding of the group's traits

## Interval Estimation

- give a range likely to contain the true parameter. This method recognizes data variability and estimation uncertainty.
- allows for uncertainty in the estimate and acknowledges the margin of error (CI)

**Confidence intervals (CI) give us a sense of freedom in our estimations, while point estimates only provide a single number without considering this uncertainty.**

# Good Estimator

*Estimator* – is a rule (usually a formula) that tells you how to calculate the estimate based on the sample

Properties of good estimator:

1. Unbiasedness
2. Consistency
3. Sufficiency
4. Efficiency



# Unbiasedness

Any sample statistic is said to be an unbiased estimator for the population parameter if on an average the value sample statistic is equal to the parameter value.

e.g.  $E(\bar{x}) = \mu$  i.e. sample mean is an unbiased estimator of population mean

# Consistency

An estimator is said to be a consistent estimator for the parameter if the value of statistics gets closer to the value of the parameter and the respective variance of statistics get closer to zero as sample size increases.

e.g.  $E(\bar{x}) \rightarrow \mu$  and  $V(\bar{x}) = \frac{\sigma^2}{n} \rightarrow 0$  as sample size  $n$  increases

# Sufficiency

If a statistic contain almost all information regarding the population parameter that is contained in the population then the statistic is called sufficient estimator for the parameter.

# Efficiency

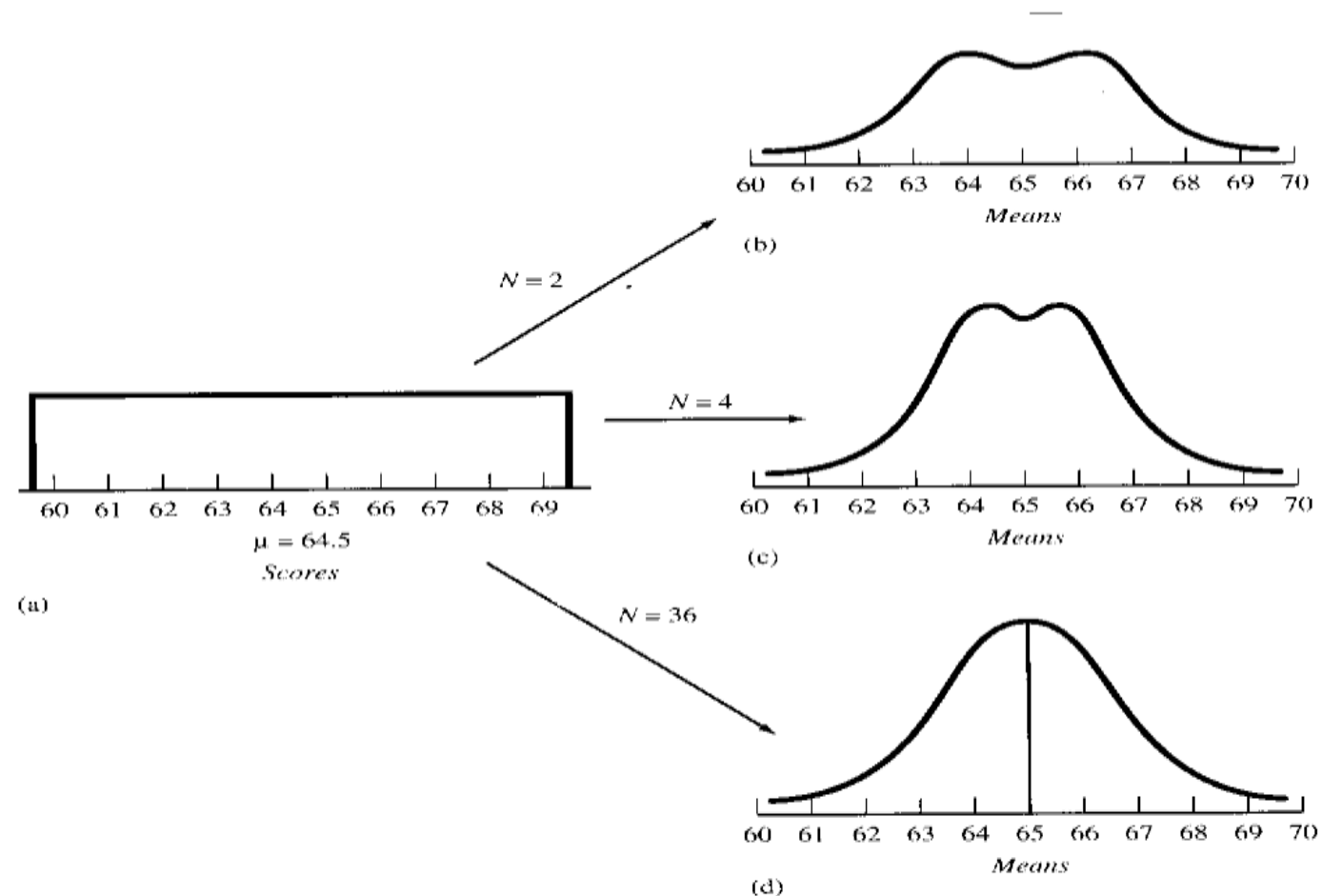
An estimator is said to be an efficient estimator if it contains smaller variance among all variances of all other estimators.

# Sampling distribution

From a population of size  $N$ , number of samples of size  $n$  can be selected and these samples give different values of a statistics. These different values of statistic can be arranged in form of a frequency distribution which is known as **sampling distribution** of that statistics.

We can have sampling distribution of sample mean, sampling distribution of sample proportion etc.

# Sampling distribution as N- increases Mean



**FIGURE 10-5**

Relationship between sample size and the shape of the sampling distribution of means (after Clarke et al., 1965): (a) frequency distribution of scores in the population; (b) sampling distribution from samples of size 2; (c) sampling distribution from samples of size 4; (d) sampling distribution from samples of size 36

# Methods Used to Calculate Point Estimators

- Point estimators can be calculated using various methods, depending on the nature of the parameter being estimated and the characteristics of the sample data.
- Common methods include the method of moments, maximum likelihood estimation, and Bayesian estimation.
- In the method of moments, the estimator is chosen to match the sample moments (e.g., mean, variance) with the corresponding population moments.

# Methods Used to Calculate Point Estimators

- Maximum likelihood estimation involves finding the parameter value that maximizes the likelihood function, which measures the probability of observing the sample data given different values of the parameter.
- Bayesian estimation incorporates prior beliefs about the parameter into the estimation process, updating these beliefs based on the observed data to obtain a posterior distribution for the parameter.

# How to calculate (formula)

- Formulae used to measure point estimators depend on the specific estimator and parameter being estimated.
- However, in general, a point estimator can be represented as a function of the sample data, denoted by a symbol such as  $\hat{\theta}$ .

For example, the sample mean ( $\bar{x}$ ) is a point estimator for the population mean ( $\mu$ ), and its formula is:

$$\bar{x} = (\sum x_i) / n$$

$x_i$  - each Individual Observation in Sample

$n$  - Sample Size

Similarly, the sample variance ( $s^2$ ) is a point estimator for the population variance ( $\sigma^2$ ), and its formula is:

$$s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$$



# What are the Values Needed to Calculate Point Estimators?

- Sample data from the population of interest.
- The specific values needed depend on the estimator being used.
- To calculate the sample mean, you need the individual observations from the sample.
- To calculate the sample variance, you need both the individual observations and the sample mean.
- Similarly, other estimators require different types of sample data, such as counts for proportions or ordered data for medians.
- Sample size is often a crucial factor in calculating point estimators, as it determines the precision and reliability of the estimates.

# Some Common Point Estimators Used in Statistics

Common point estimators include:

- Sample Mean ( $\bar{x}$ ) for estimating Population Mean ( $\mu$ )
- Sample Variance ( $s^2$ ) for estimating Population Variance ( $\sigma^2$ )
- Sample Proportion ( $\hat{p}$ ) for estimating Population Proportions

# Estimation Methods

Several techniques can be used to generate estimates:

- Method of Moments
- Maximum Likelihood Estimation (MLE)

# Method of Moments

- This method compares the moments (central tendency and spread) that are computed from the sample data to the corresponding moments in the population.
- The population parameters can be estimated by working out the resulting equations.

# Maximum Likelihood Estimation (MLE)

- Maximum likelihood estimation (MLE) aims to find parameter values that give the highest chance of observing the data in a statistical model.
- It involves identifying values that maximize the likelihood of the observed data.

# Maximum Likelihood Estimation (MLE)

- MLE is a method used to find the most probable values of variables based on given data.
- It involves starting with an initial estimate for a parameter and iteratively adjusting it to maximize the likelihood of observing the data.
- By comparing different estimates to the dataset, the MLE process helps identify the parameter values that best fit the data.
- This statistical method is valuable in accurately estimating unknown variables by increasing the probability of occurrence in the dataset through adjusting parameter values in a model.

# Estimators as Random Variables

An estimator in statistics is considered a random variable as it's computed from random data samples, leading to varying values.

# Variability

- **Sample Variability:** When sampling from a population, we randomly select a subset of people or observations. Estimators such as sample mean vary between samples.
- **Sampling Distribution:** The sampling distribution of an estimator shows the potential values it can take when calculated from various samples of a certain size from a population, providing insights into its characteristics and variability.



# Variability

- Bias and Variance: Estimators can have bias, consistently overestimating or underestimating the true parameter. Variance measures the spread of estimator values around its predicted value. Both variance and bias impact the accuracy of estimators.
- Mean and Variance of Estimators: Estimators have the same mean and variance as random variables. The mean of an estimator should be equal to the parameter it is estimating. The variance of an estimator indicates its precision  $\sigma^2$

# Variability

- Efficiency and Consistency: Efficiency measures an estimator's accuracy in estimating a parameter with sample data. A smaller variance indicates better efficiency. Consistency means the estimator approaches the correct parameter value as sample size increases.
- Central Limit Theorem: Central Limit Theorem states that regardless of population distribution, the sampling distribution of many estimators becomes normal as sample size increases. Understanding this theorem is essential to grasp the behavior of estimators.

# Interval Estimate

**Confidence interval (interval estimate)** – A range of values defined by the confidence level within which the population parameter is estimated to fall.

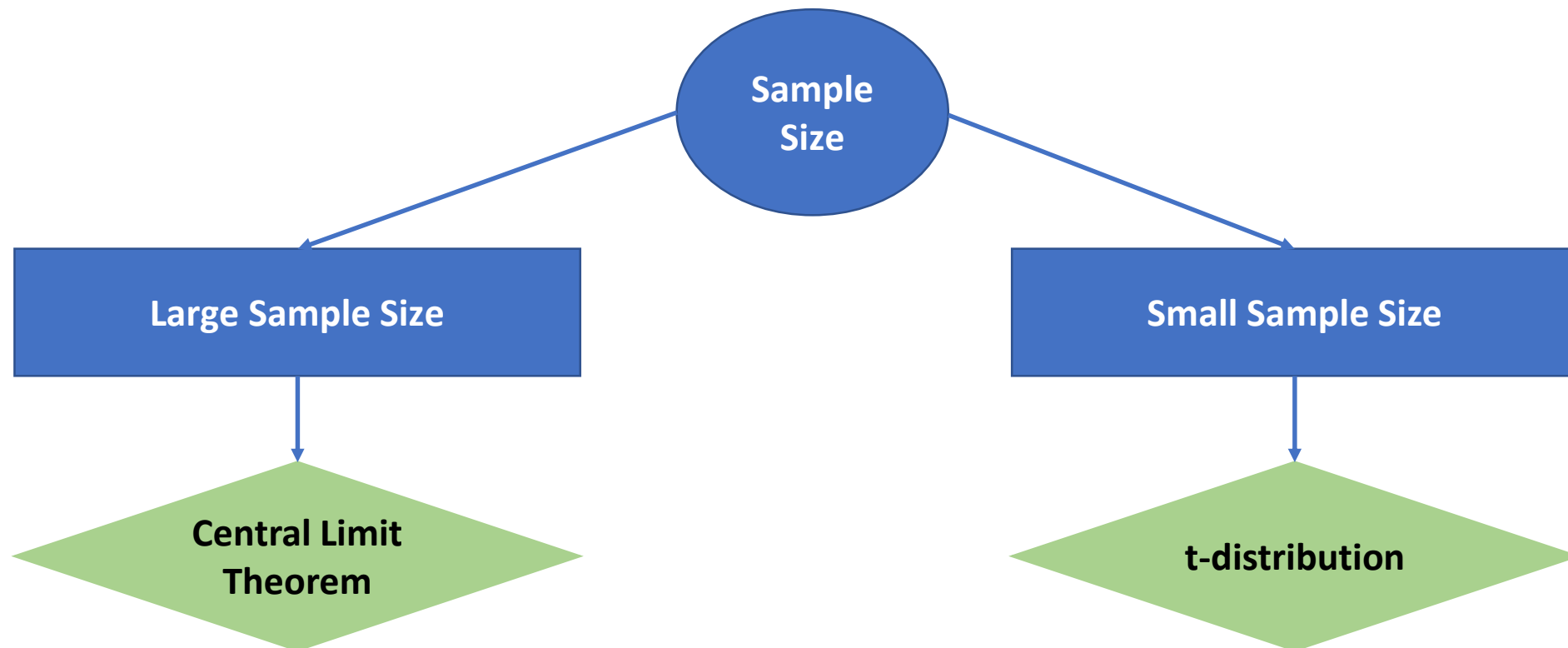
- The interval estimate is less precise, but gives more confidence.
- Two numbers are calculated to create an interval within the parameter is expected to lie. It is constructed so that, with a chosen degree of confidence, the true unknown parameter will be captured inside the interval

# Confidence Interval

- The point estimate is going to be different from the population parameter because due to sampling error, and there is no way to know how close it is to the actual parameter.
- For this reason, in stats, interval estimate is chosen to estimate the actual parameter in form of range of values.
- Has a specific level of confidence (90%, 95%, 99%)

# Confidence Interval

- CI depends upon the sampling distribution
- The shape of the sampling distribution

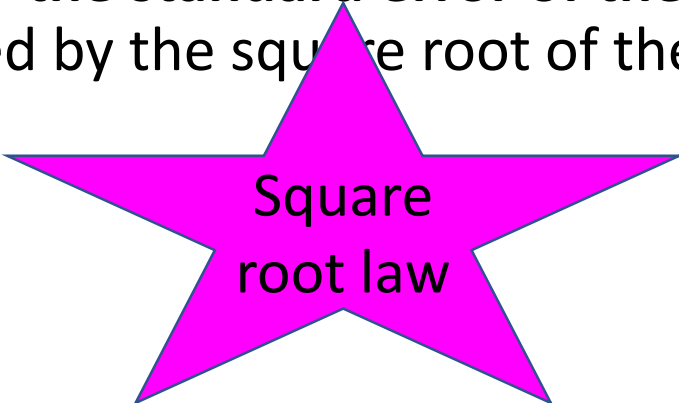


# Central Limit Theorem

States that the sampling distribution of means, for samples of 30 or more:

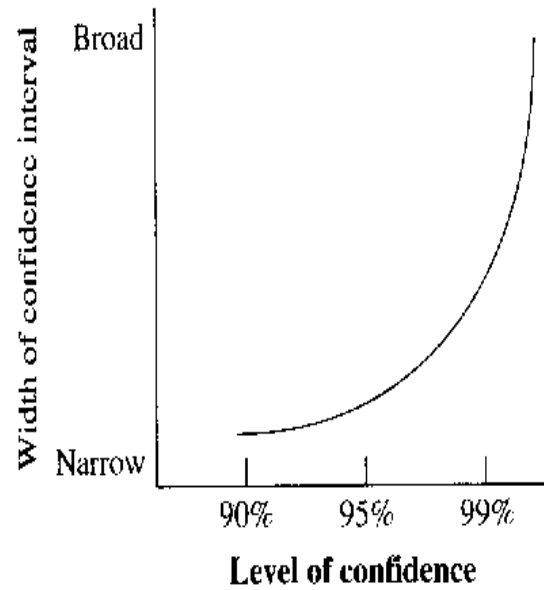
- Is normally distributed (regardless of the shape of the population from which the samples were drawn)
- Has a mean equal to the population mean, “mu” regardless of the shape population or of the size of the sample
- Has a standard deviation--**the standard error of the mean**--equal to the population standard deviation divided by the square root of the sample size

$$SEM = \sigma / \sqrt{n}$$

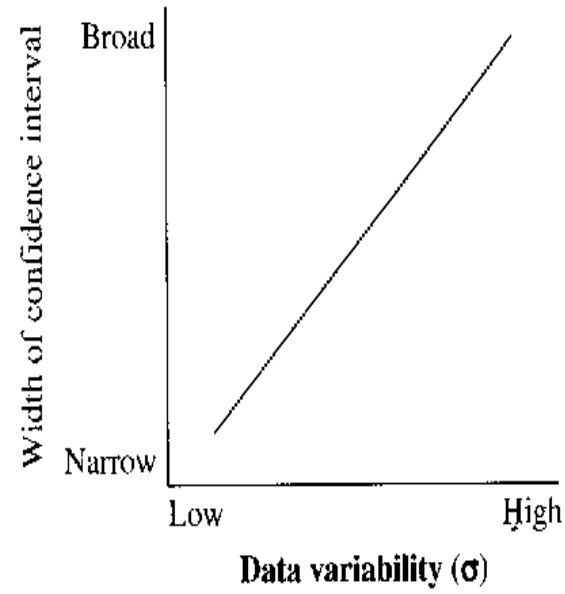


Square  
root law

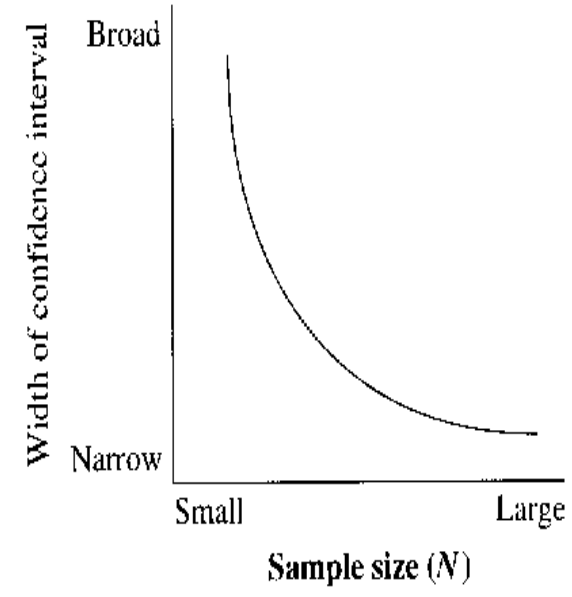
# Distribution of Means and Standard Error of the Means



(a)

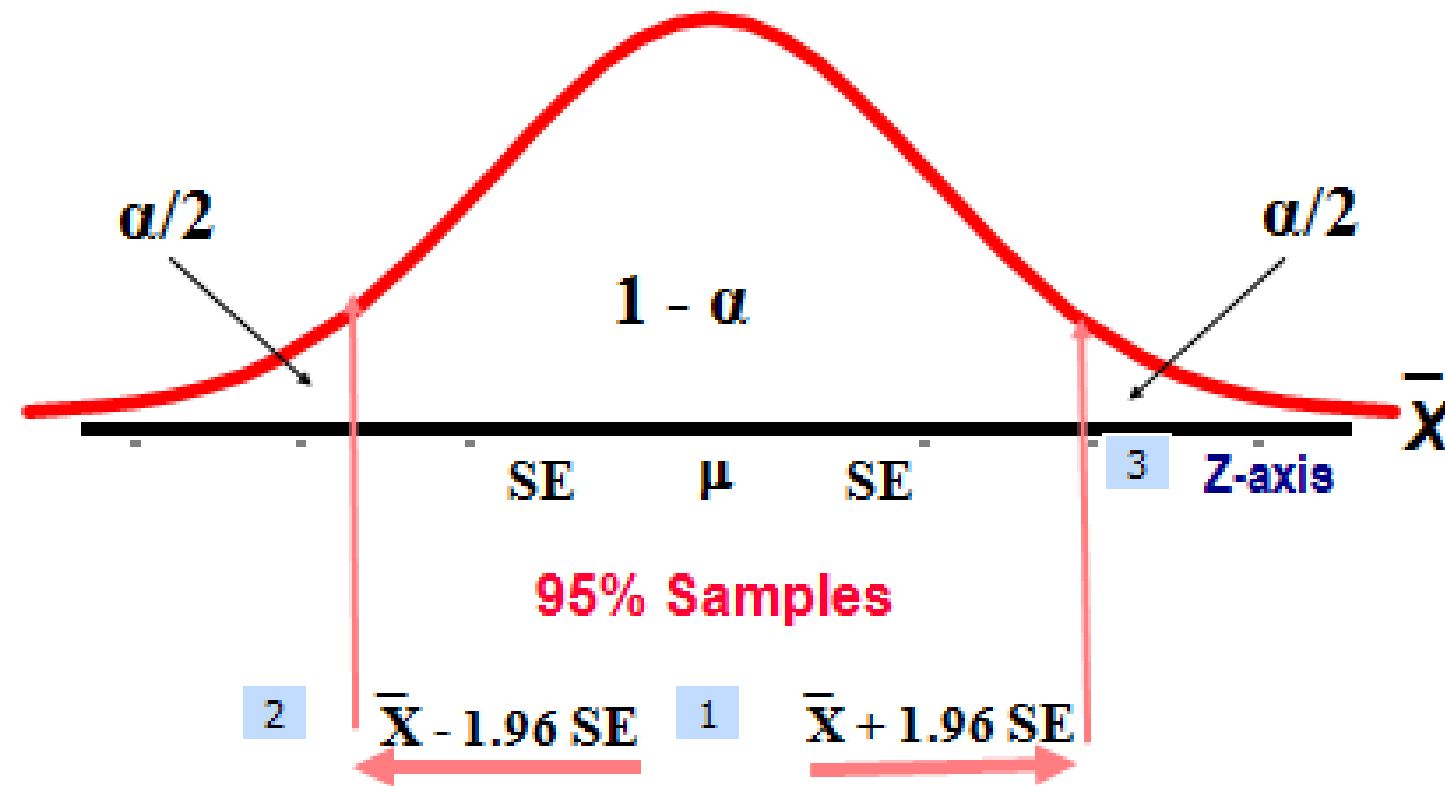


(b)



(c)

# Confidence interval





# Confidence limits

- The  $\alpha$  (“alpha”) level represents the “lack of confidence”
- $(1-\alpha)100\%$  represent the **confidence level of a confidence interval**
- **Confidence interval** =  $\bar{x} \pm (z_{1-\alpha/2})(SEM)$
- *$Z_{1-\alpha/2}$  instead of  $Z_{1-\alpha}$  in this formula is because the random error (imprecision) is split between right and left tail*

# Z values for different confidence level

$(1-\alpha)100\%$	$\alpha$	$Z_{1-\alpha/2}$
90%	.10	$Z_{1-.10/2} = Z_{.95} = 1.64$
95%	.05	$Z_{1-.05/2} = Z_{.975} = 1.96$
99%	.01	$Z_{1-.01/2} = Z_{.995} = 2.58$



Area under the curve

# Z table 2 tailed

Second decimal places

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962			
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

Area under the curve

1.96=1.9+0.06

Z.975 = 1.96

# Process for Constructing Confidence Intervals



- Compute the sample statistic (e.g. a mean)
- Compute the standard error of the mean
- Make a decision about level of confidence that is desired (usually 95% or 99%)
- Find tabled value for 95% or 99% confidence interval
- Multiply standard error of the mean by the tabled value
- Form interval by adding and subtracting calculated value to and from the mean

# Example - Estimating Mean Radiation Dose

Q. Assume we want to estimate the mean dose of radiation given to patients with a confidence level of 95%.

- Sample Mean: 65 Gy
- Population Std. Dev.: 10 Gy
- Sample Size (n): 30.
  
- Standard Error =  $10 / \sqrt{30} \approx 1.83$  Gy.
- Confidence Interval:  $65 \pm 1.96 * 1.83 \rightarrow (61.41, 68.59$  Gy).

# Interpreting results...

- The 95% CI (61.41 Gy, 68.59 Gy) means we are 95% confident that the true mean dose lies within this range.
- Helps quantify uncertainty in clinical research and treatment planning.
- Wider intervals imply more uncertainty; narrower intervals imply more precision.

# Factors Affecting Confidence Intervals

- Sample size: Larger samples result in narrower (more precise) intervals.
- Confidence level: Higher confidence levels (e.g., 99%) result in wider intervals.
- Population variability: Greater variability leads to wider intervals.