

# **An Introductory Data Analysis**

**08/13/2019**

**Dapo Olaitan, Biostatistician**

## Lecture Two

- Inferential Statistics
  - Estimations (Point Estimate, Standard Error and Confidence Intervals
    - Normal Distribution
    - Normal Standard Distribution
    - Hypothesis Testing (z-Test, t-Test, F-test)

# Introductory Inferential Statistics

## Inferential Analysis

Involves two basic areas:

1. Estimation
  - Point Estimate
  - Standard Error
  - Confidence Intervals
2. Hypothesis Testing

# Point Estimate

## Point Estimate :

Given a sample  $x_1, x_2, \dots, x_n$  of size  $n$  drawn from a population of size  $N$  with average  $\mu$  and standard deviation  $\sigma$ .

The point estimate of  $\mu = \bar{x}$ ;

## Note :

- Are derived estimates (sample mean and standard deviation).
- Point estimate may not accurately reflect the actual distribution.

## Standard Error

Given a **sample**  $x_1, x_2, \dots, x_n$  of **size n** with **mean**  $\bar{x}$ , the **standard error** of the mean  $\bar{x}$  is the standard deviation of all possible  $\bar{x}$ 's

From the central limit theorem (CLT), for n large ( $n \geq 30$ ):

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

### Note :

If  $\sigma$  is unknown, it is easily estimated from the sample's standard deviation **S**.

# Confidence Intervals

## Confidence Intervals

- Is the range of values on both sides of an estimate with a  $(1-\alpha)\%$  level of confidence.

## Confidence Intervals (of a Point Estimate)

- Helps to describe the precision of the an estimate.
- Limits the chance of errors in an estimation.
- Determines the degree of uncertainty in a process estimation.

# The Central Limit Theorem (CLT)

Consider a sample  $\mathbf{X}$  from a population ( $\mathbf{N}$ ) with  $\mu$  and  $\sigma$ .

Consider random samples ( $\mathbf{n}$ ) drawn with replacement from the population,

thus,

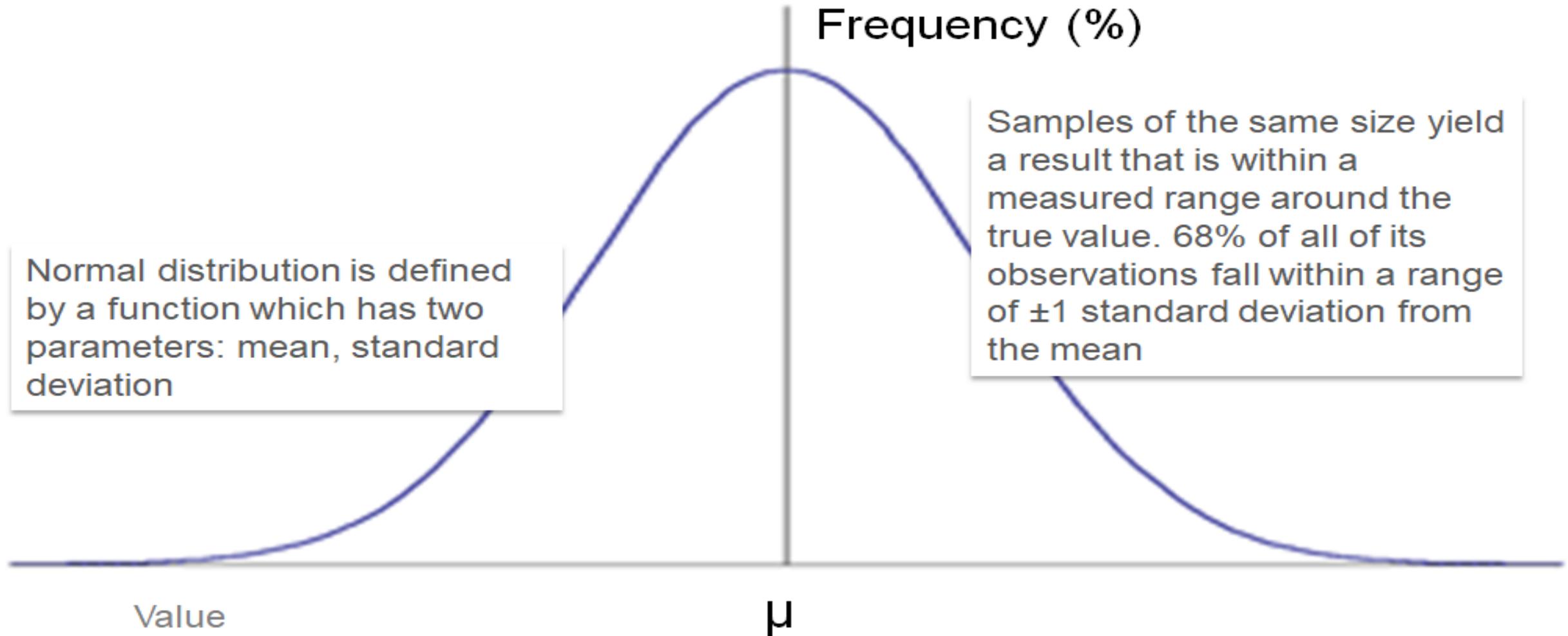
for **large “n”**, the distribution of the **sample mean  $\bar{x}$**  is approximately normally distributed with

$$\mu_{\bar{x}} = \mu \quad , \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad , \quad \text{and} \quad \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

## Importance :

The distribution of the **sample mean ( $\bar{x}$ )** is approximately normal even if  $\mathbf{X}$  does **not** follow  $N(\mu, \sigma)$ .

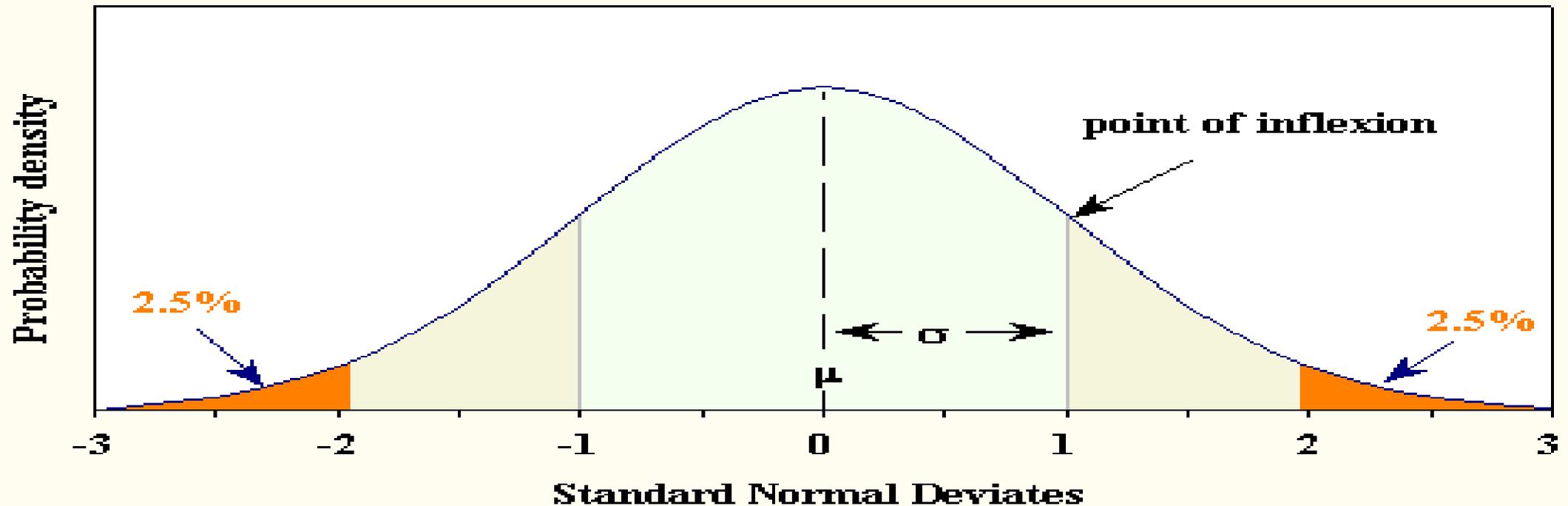
# Normal Distribution



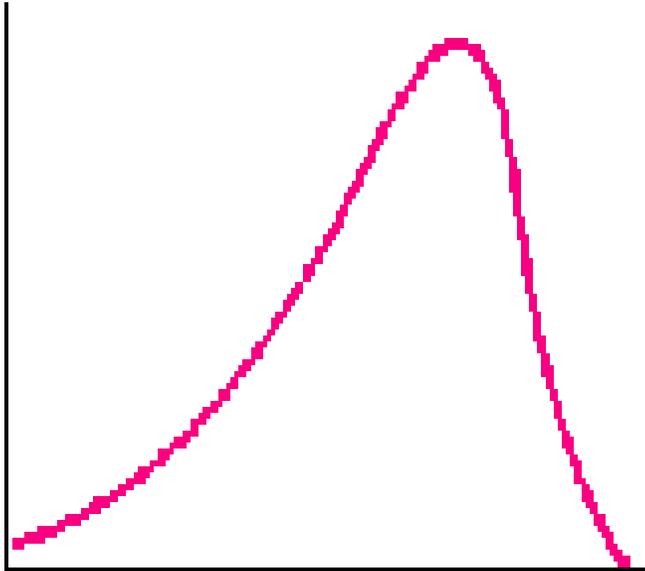
# Standard Normal distribution

The Standard Normal distribution follows a normal distribution and is perfectly symmetric about zero (0).

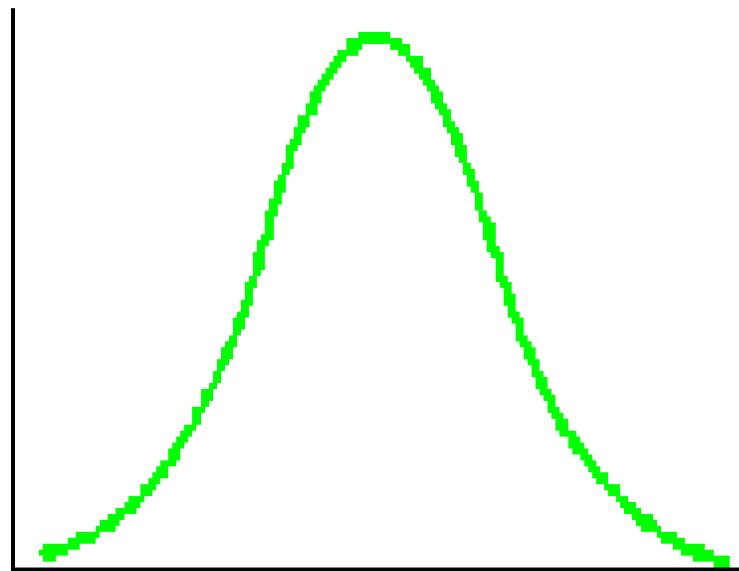
**Standard Normal Distribution (mean = 0 , standard deviation = 1)  
95% of the observations lie between -1.96 and +1.96**



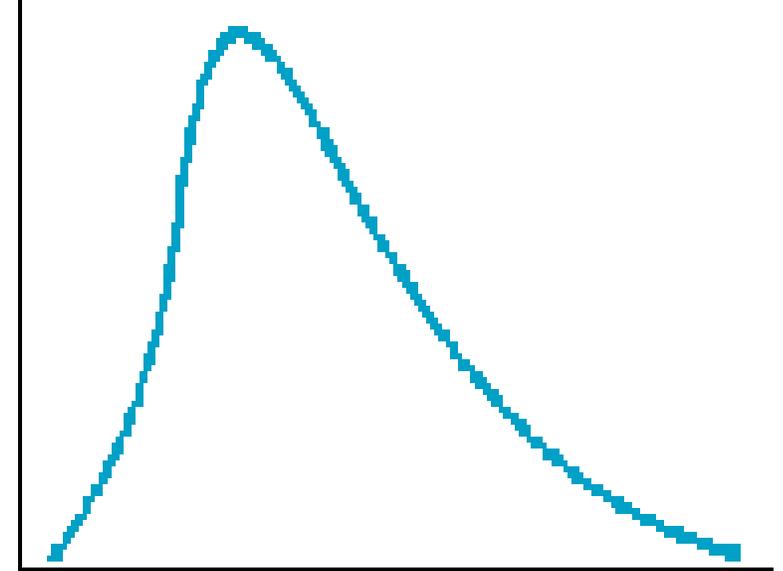
# Types Of Distribution



**Negatively (left)  
skewed  
distribution**

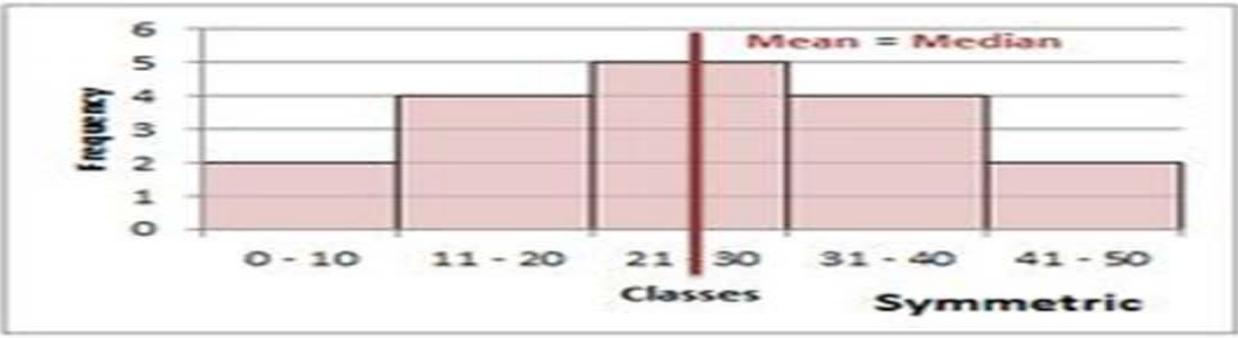
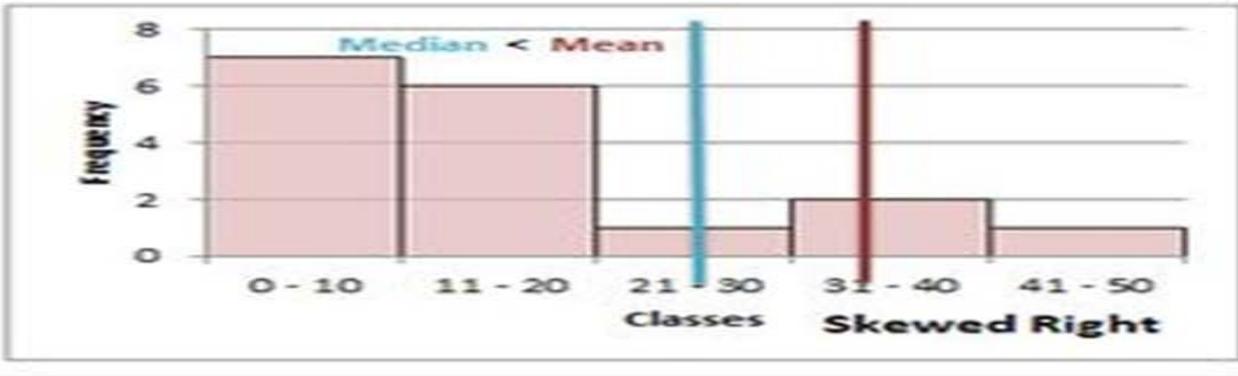


**Normal  
skewed  
distribution**

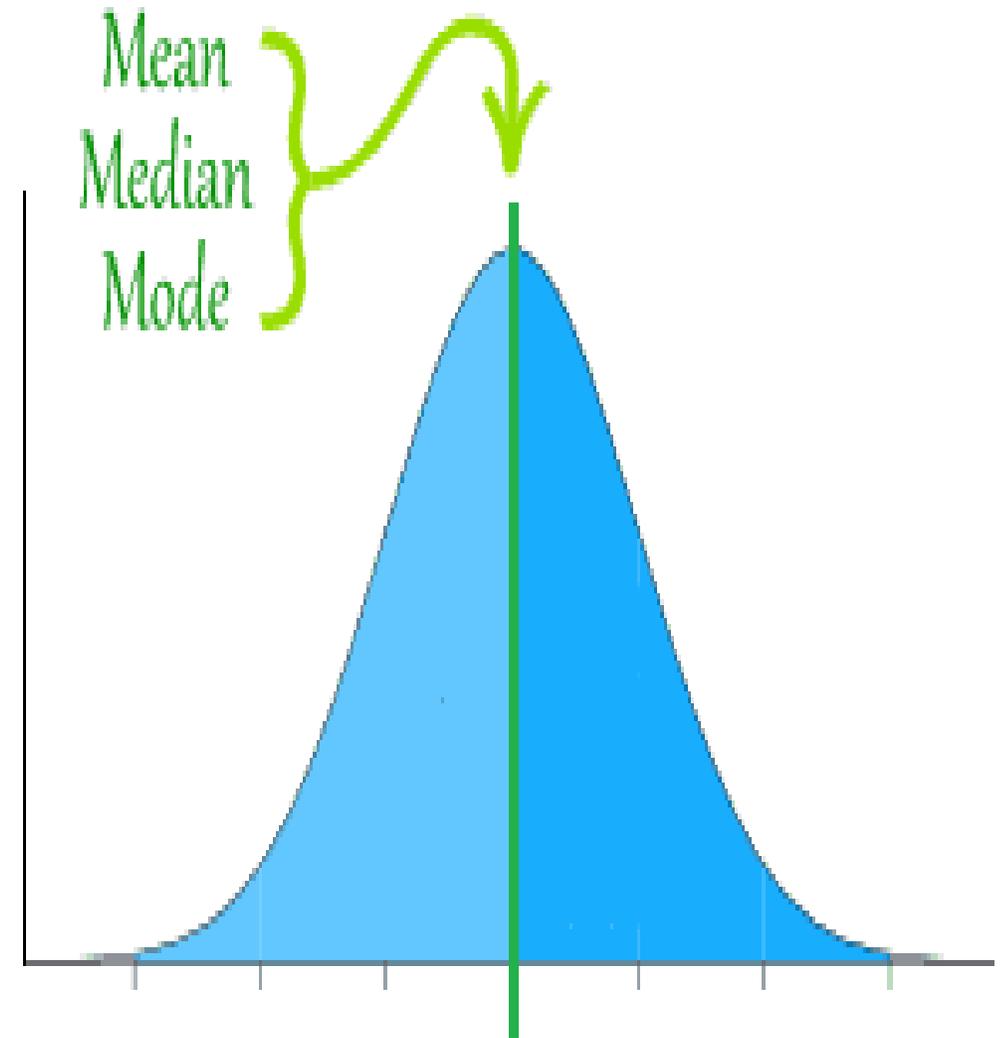
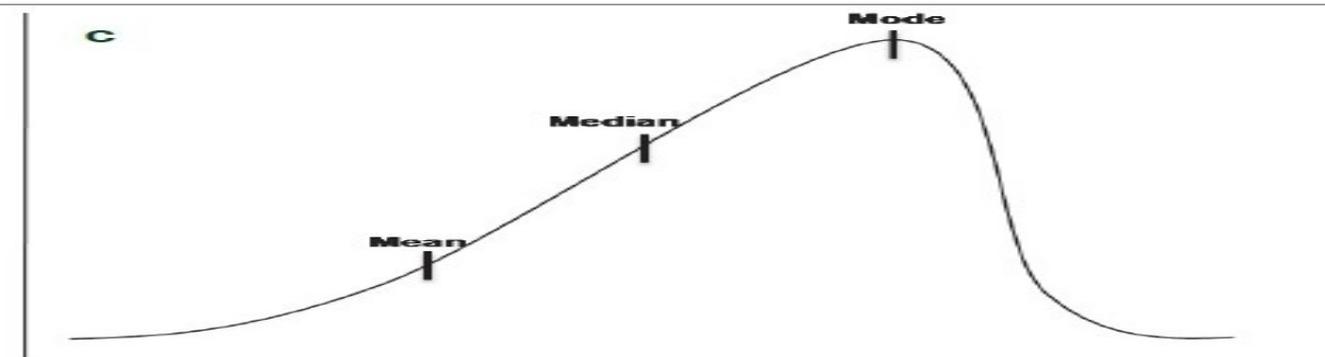
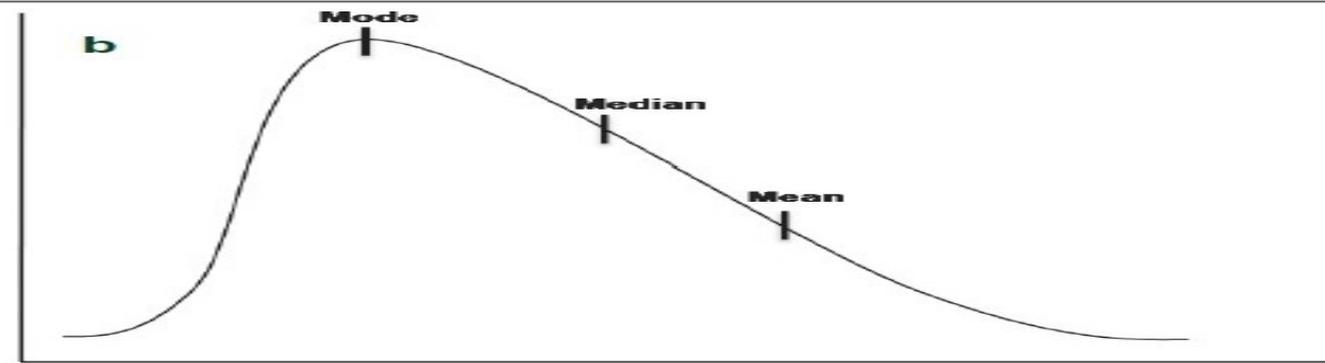
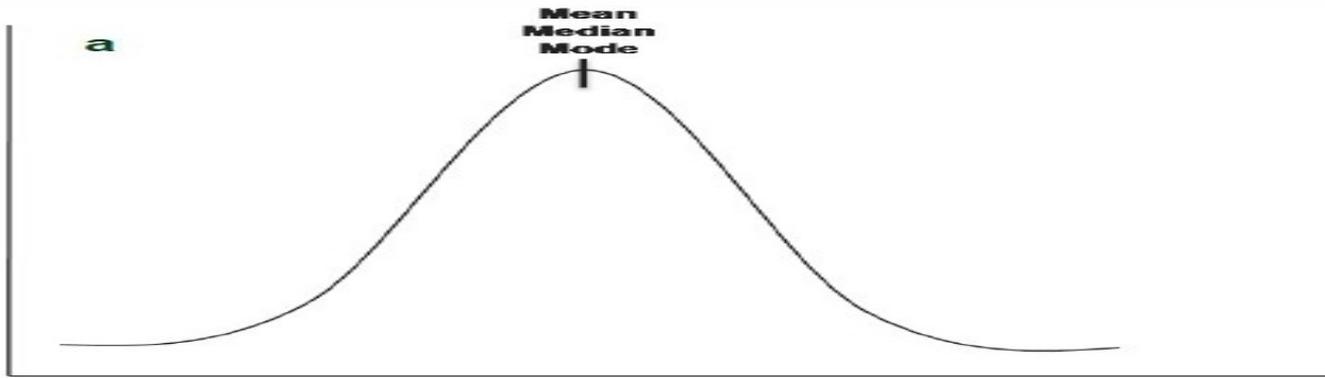


**Positively (right)  
skewed  
distribution**

# Types of Distribution

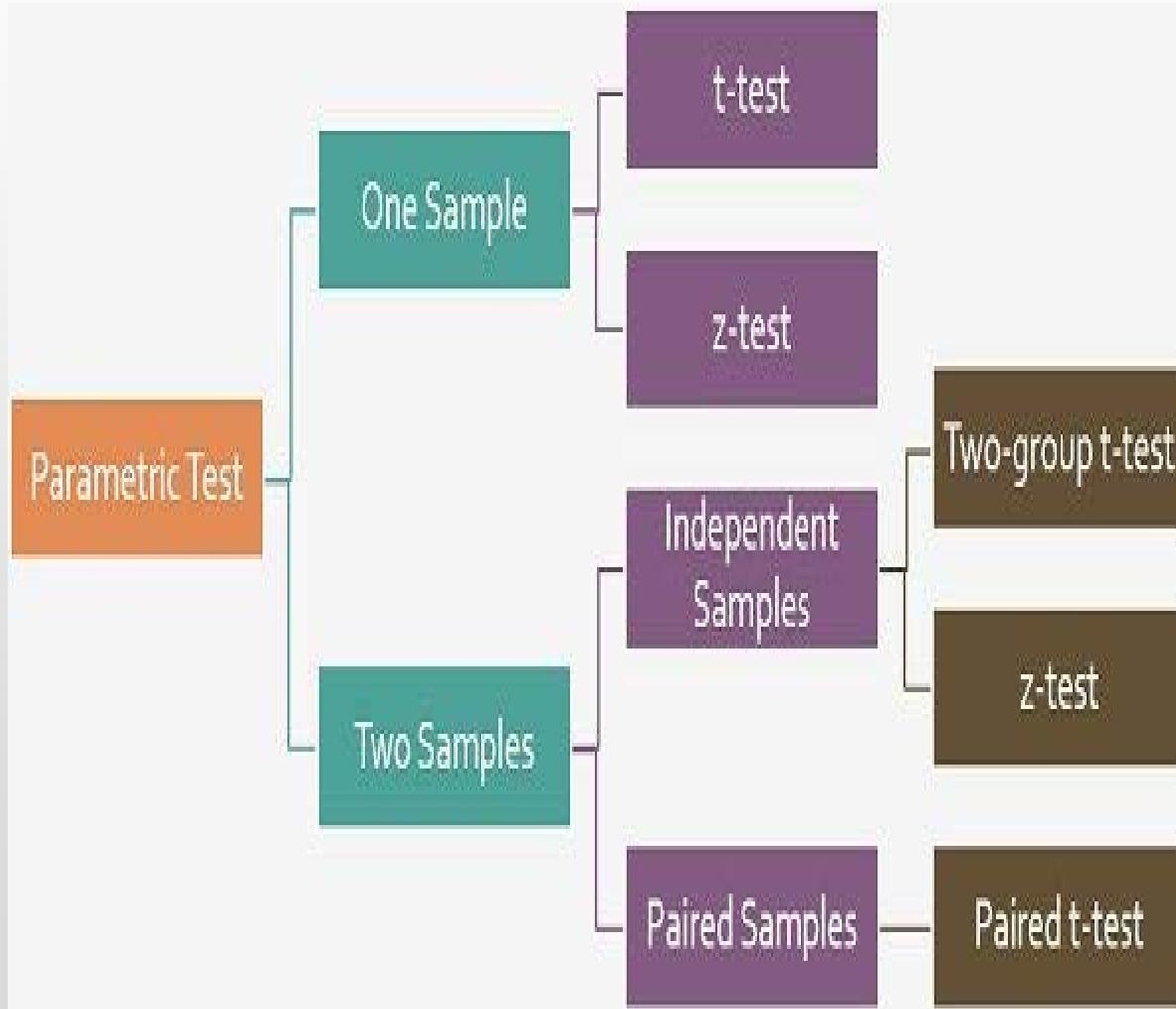
 <p>Frequency</p> <p>Classes</p> <p>Symmetric</p> <p>Mean = Median</p>	<p>When the data is symmetric, the mean and median are typically close together.</p>
 <p>Frequency</p> <p>Classes</p> <p>Skewed Left</p> <p>Mean &lt; Median</p>	<p>When the data is skewed left, it is "pulled" to the left, and this drags the mean too low.</p>
 <p>Frequency</p> <p>Classes</p> <p>Skewed Right</p> <p>Median &lt; Mean</p>	<p>When the data is skewed right, it is "pulled" to the right, and this drags the mean too high.</p>

# Distribution & Estimates

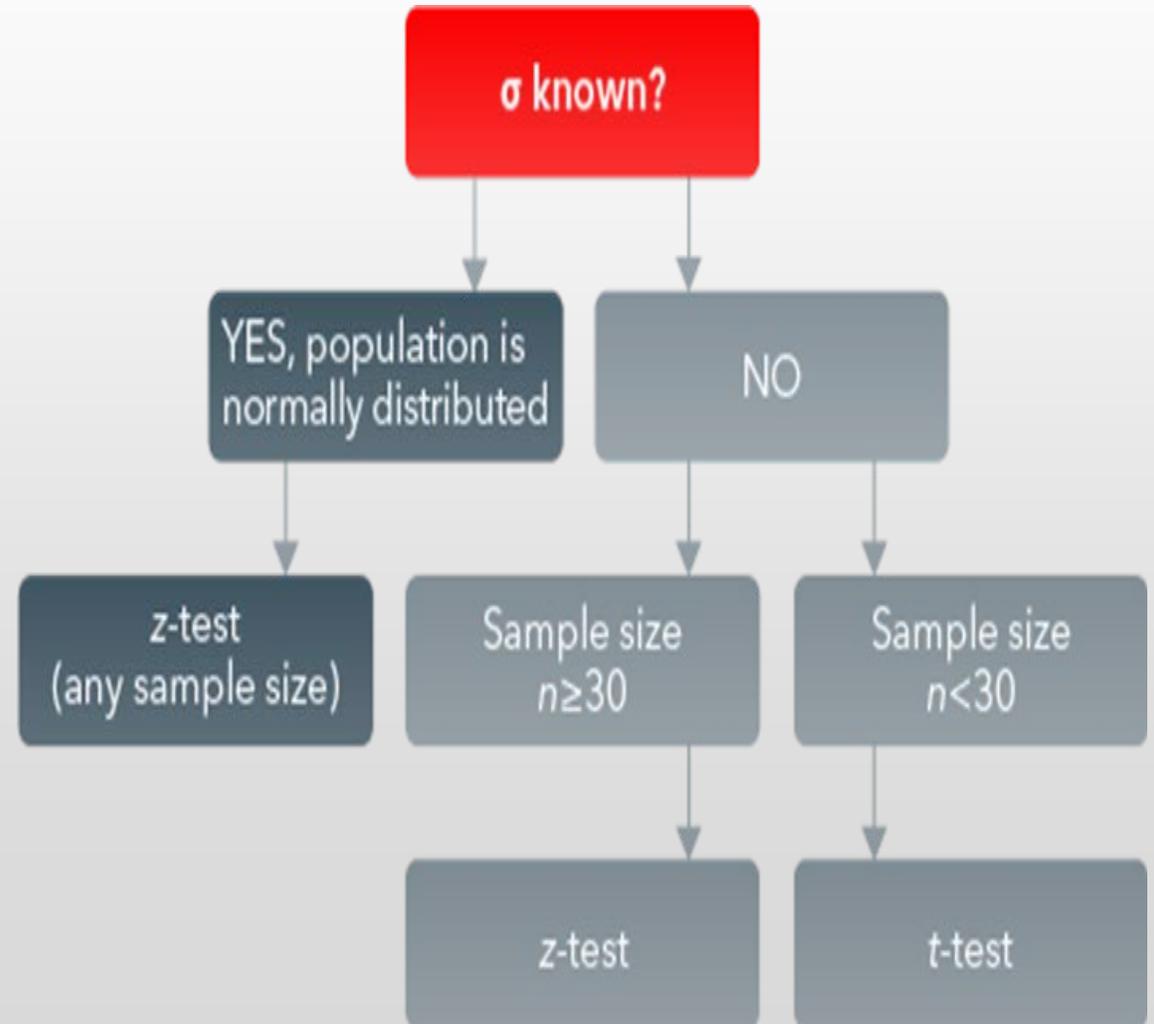


# Hypothesis Testing

## Normal Distribution



## Sample size and Parameter



# Hypothesis Testing

- The statistical inference about a population is based on a sample drawn from the population.
- The confidence interval estimates a population parameter with respect to the point estimate.
- Statistical inference about the parameter estimate is known as the ‘Hypothesis testing’.

# Hypothesis Testing

Comparison (Sample Mean / Normal Mean Value) :

- Large difference between the means,
  - the difference has not occurred by chance .
- Small variability about the mean,
  - the observed sample mean likely represents the true mean of the population.
- Large sample size,
  - the more accurate the sample mean will represent the true population mean.

# Hypotheses

There are **two hypotheses** involved in a hypothesis test;

- The ***null hypothesis,  $H_0$***  (cannot be viewed as false unless sufficient evidence to the contrary is obtained).
- The ***alternative hypothesis,  $H_a$***  (hypothesis against which the null hypothesis is tested and is viewed as true when the null hypothesis is declared as false).

## Decision :

A correct decision is made if a true hypothesis is accepted or a false hypothesis is rejected.

## Types of Error :

- *Type I error* (rejecting a null hypothesis,  $H_0$  when it is actually true)
- *Type II error* (rejecting a null hypothesis,  $H_a$  when it is actually false)

**Probability :** The probabilities of making errors can be assessed

- $P(\text{Type I error}) = P(\text{Rejecting } H_0 | H_0 \text{ is true})$ , called the *level of significance*.

## z Test

The **z test** uses samples with normal distribution to test hypotheses about ;

- the mean of a population based on a single sample
- the proportion of successes in a population based on a single sample
- the difference between the means (or proportions of successes ) of two populations based on samples from each population.

## z test (Test of a Population Mean)

### Hypotheses :

- **Two-tailed (sided) test**,  $H_0 : \mu \leq \mu_0$  and  $H_a : \mu > \mu_0$
- **One-tailed (sided) lower tail test**,  $H_0 : \mu \geq \mu_0$  and  $H_a : \mu < \mu_0$
- **One-tailed (sided) upper tail test**,  $H_0 : \mu \leq \mu_0$  and  $H_a : \mu > \mu_0$

## z test – Decision Rules

### Decision Rules :

- Using a **Critical Values**
- Using the **Action Limits**
- Using a ***p* Value**

Using a Critical Values by computing Z value :

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Do not reject  $H_0$  if  $-z_{\alpha/2} \leq z \leq z_{\alpha/2}$ ,

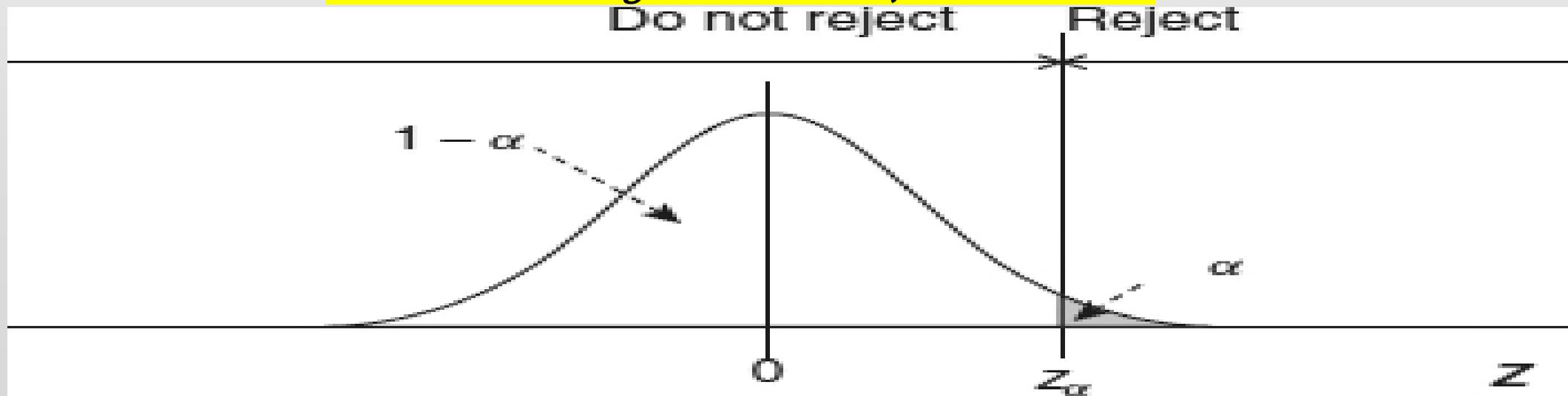
Reject  $H_0$  if  $z < -z_{\alpha/2}$  or  $z > z_{\alpha/2}$

OR

Do not reject  $H_0$  if  $|z| \leq z_{\alpha/2}$ ,

Reject  $H_0$  if  $|z| > z_{\alpha/2}$ .

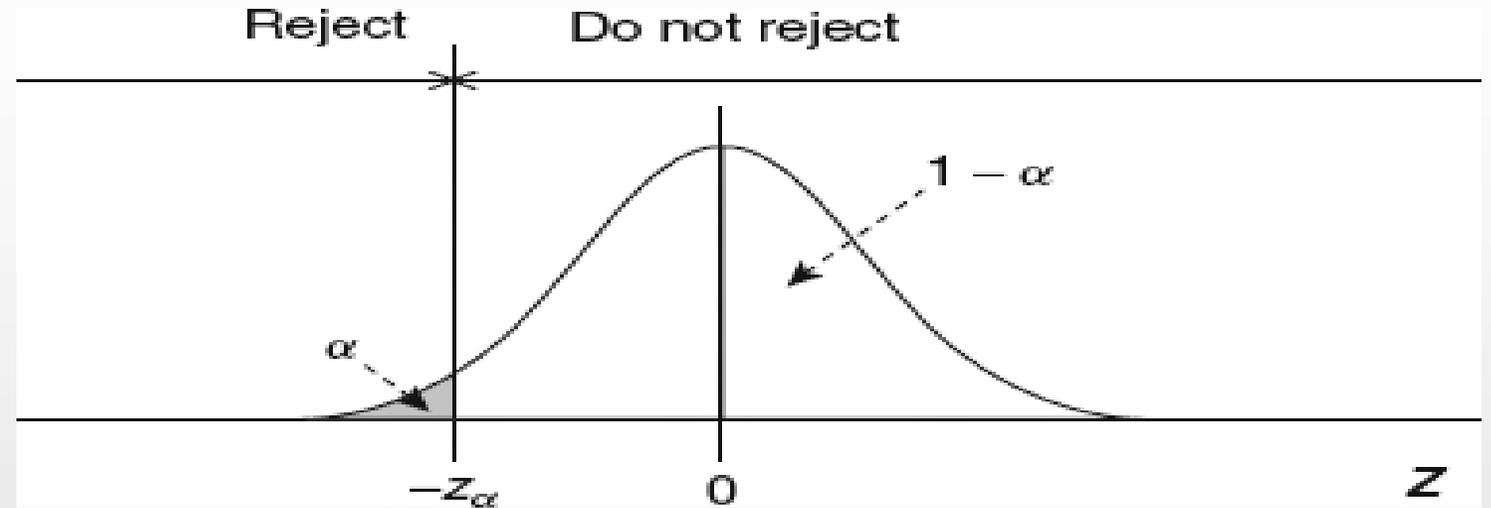
Decision Rule Using Critical Values for a Two-Tailed



### Decision Rule (Critical Value) - A One-Tailed Lower Tail Test

Do not reject  $H_0$  if  $z \geq -z_\alpha$ ,

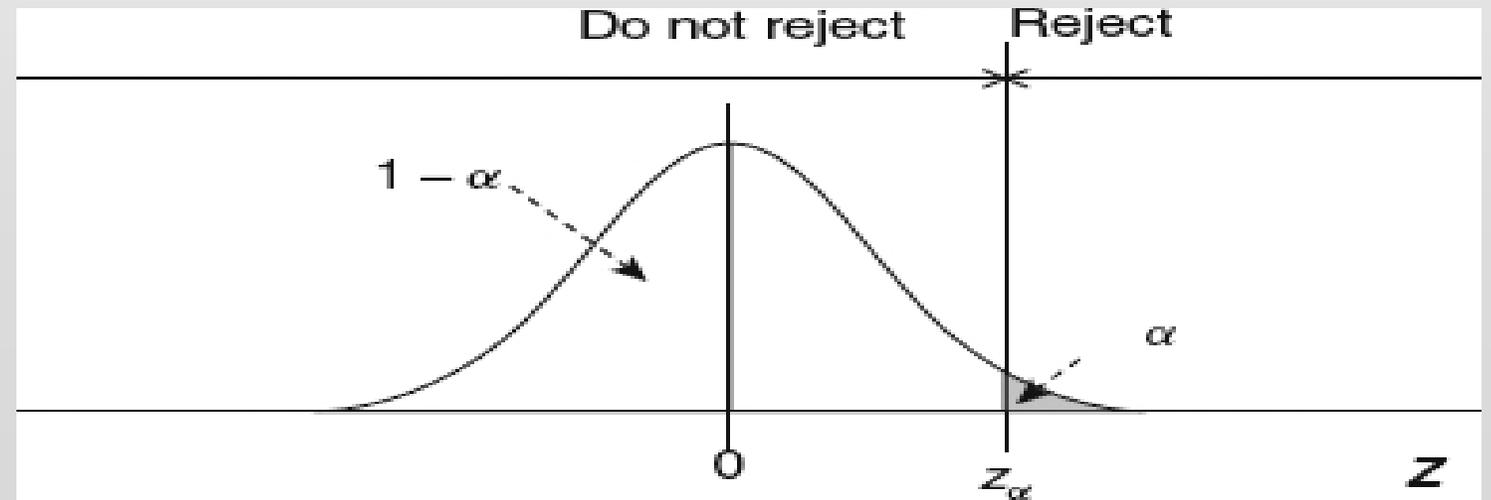
Reject  $H_0$  if  $z < -z_\alpha$ .



### Decision Rule (Critical Value) - A One-Tailed Upper Tail Test

Do not reject  $H_0$  if  $z \leq z_\alpha$ ,

Reject  $H_0$  if  $z > z_\alpha$ .

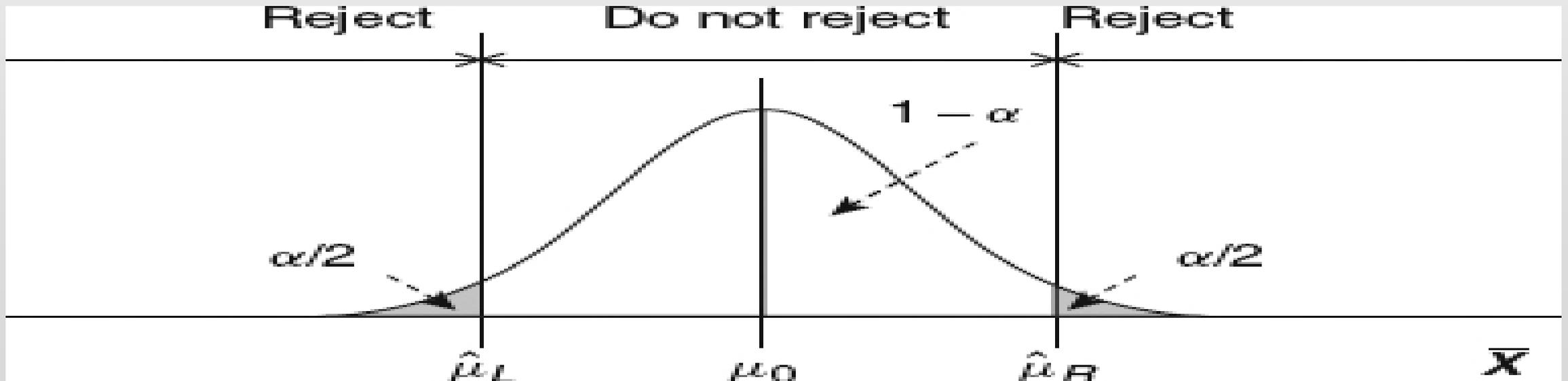


## Decision Rules Using Action Limits

A Two-Tailed Test :

Do not reject  $H_0$  if  $\hat{\mu}_L \leq \bar{x} \leq \hat{\mu}_R$  , Reject  $H_0$  if  $\bar{x} < \hat{\mu}_L$  or  $\bar{x} > \hat{\mu}_R$

where  $\hat{\mu}_L = \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  and  $\hat{\mu}_R = \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  (see Figure 4).



**Decision Rule (Action Limits) - A One-Tailed Lower Tail Test**

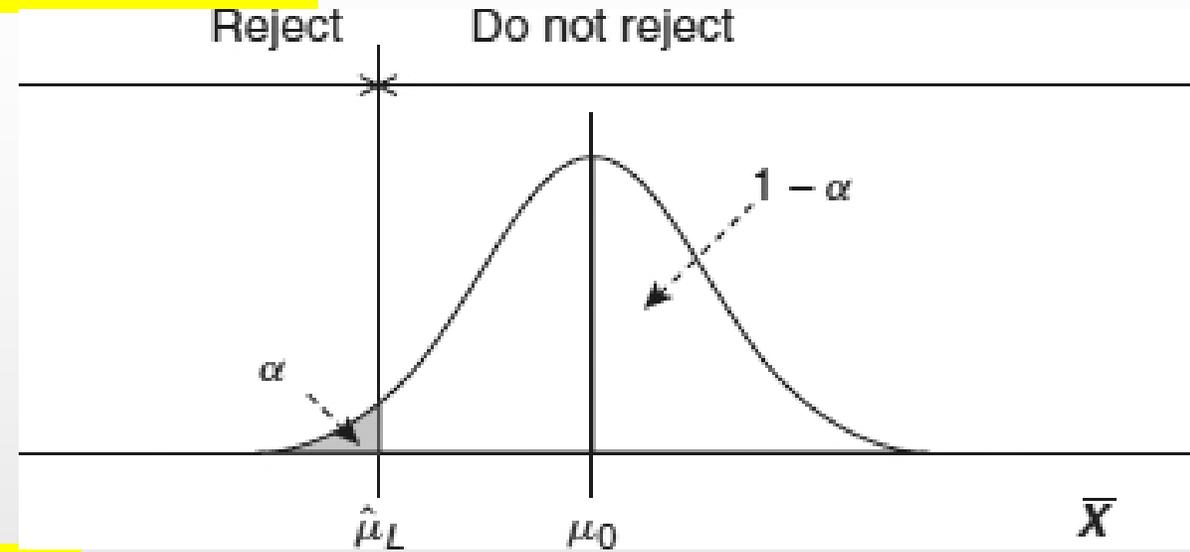
Do not reject  $H_0$  if

$$\bar{x} \geq \hat{\mu}_L$$

Reject  $H_0$  if

$$\bar{x} < \hat{\mu}_L$$

Where  $\hat{\mu}_L = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$

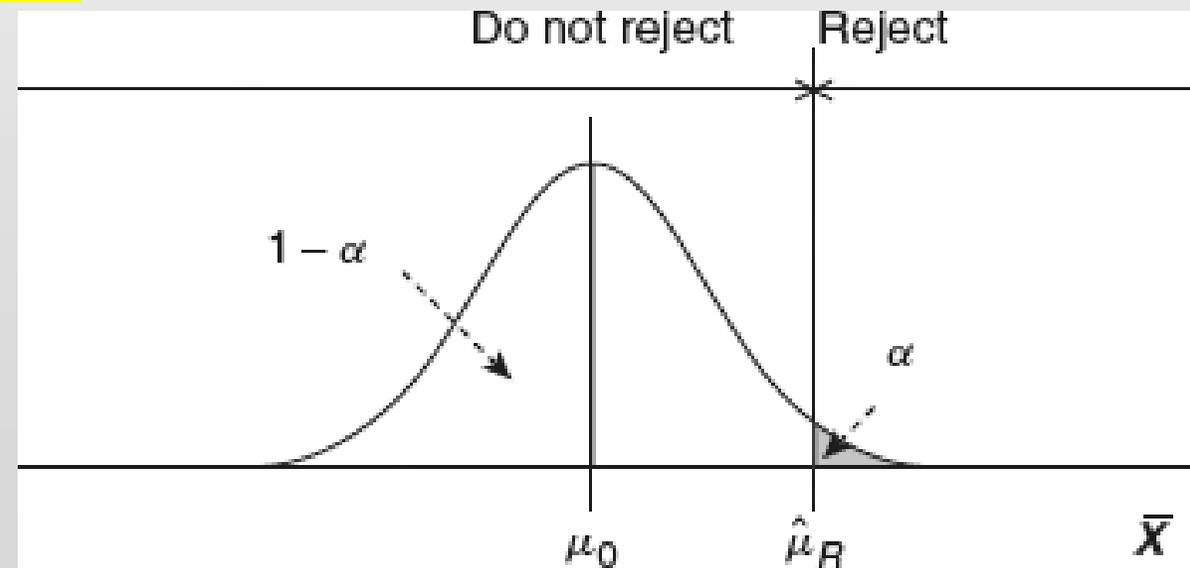


**Decision Rule (Action Limits) - A One-Tailed Upper**

Do not reject  $H_0$  if  $\bar{x} \leq \hat{\mu}_R$

Reject  $H_0$  if  $\bar{x} > \hat{\mu}_R$

Where  $\hat{\mu}_R = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$



## Decision Rules Using $p$ -Value

### **$p$ -value (of a test) :**

Is the probability of obtaining a value of the test statistic, a value that might have been more extreme in the appropriate direction of the rejection region than was actually observed.

### **A Two-tailed Test :**

Do not reject  $H_0$  if  $p\text{-value} \geq \alpha$

Reject  $H_0$  if  $p\text{-value} < \alpha$

$$p \text{ value} = \begin{cases} 2p(\bar{X} > \bar{x}), & \text{if } \bar{x} > \mu_0 \\ 2p(\bar{X} < \bar{x}), & \text{if } \bar{x} < \mu_0. \end{cases}$$

**p-value :**

**One-tailed Lower Tail Test -**

$$p \text{ value} = p(\bar{X} < \bar{x})$$

**p-value :**

**One-tailed Upper Tail Test -**

$$p \text{ value} = p(\bar{X} > \bar{x})$$

# t-Test

The t-Test is used to test whether the means of two group are *statistically* different from each other.

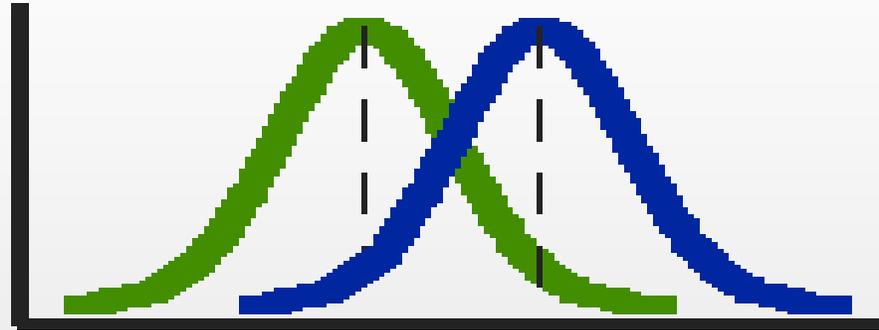
## Note :

### t-Test :

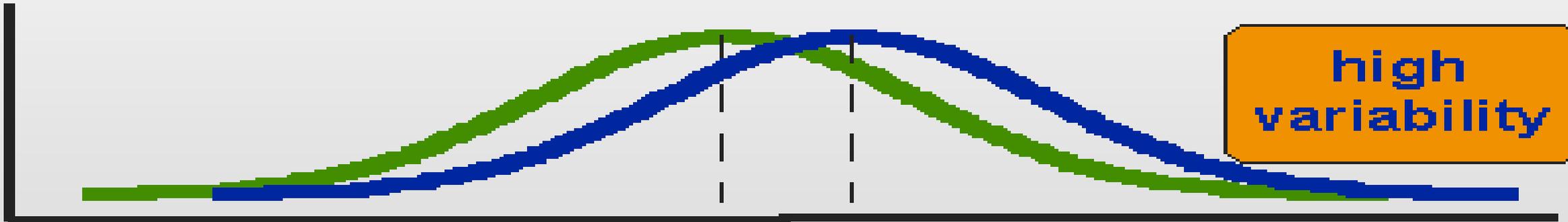
- Is based on statistical theory.
- Uses an approximation method to the sampling distribution based on the Central Limit Theorem.
- The larger the sample size, the closer to normality is the sample distribution.

## t-Test Assumptions :

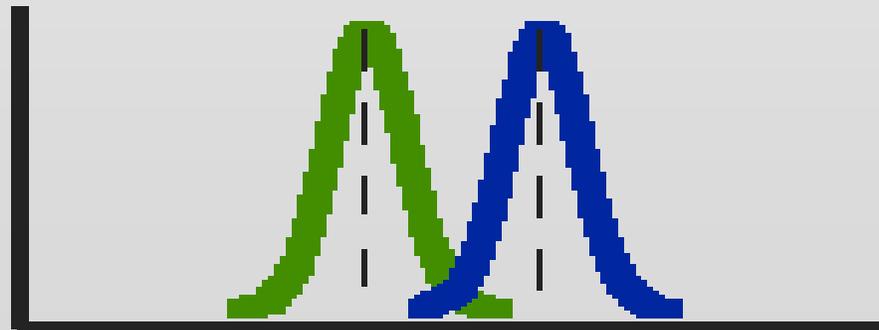
medium  
variability



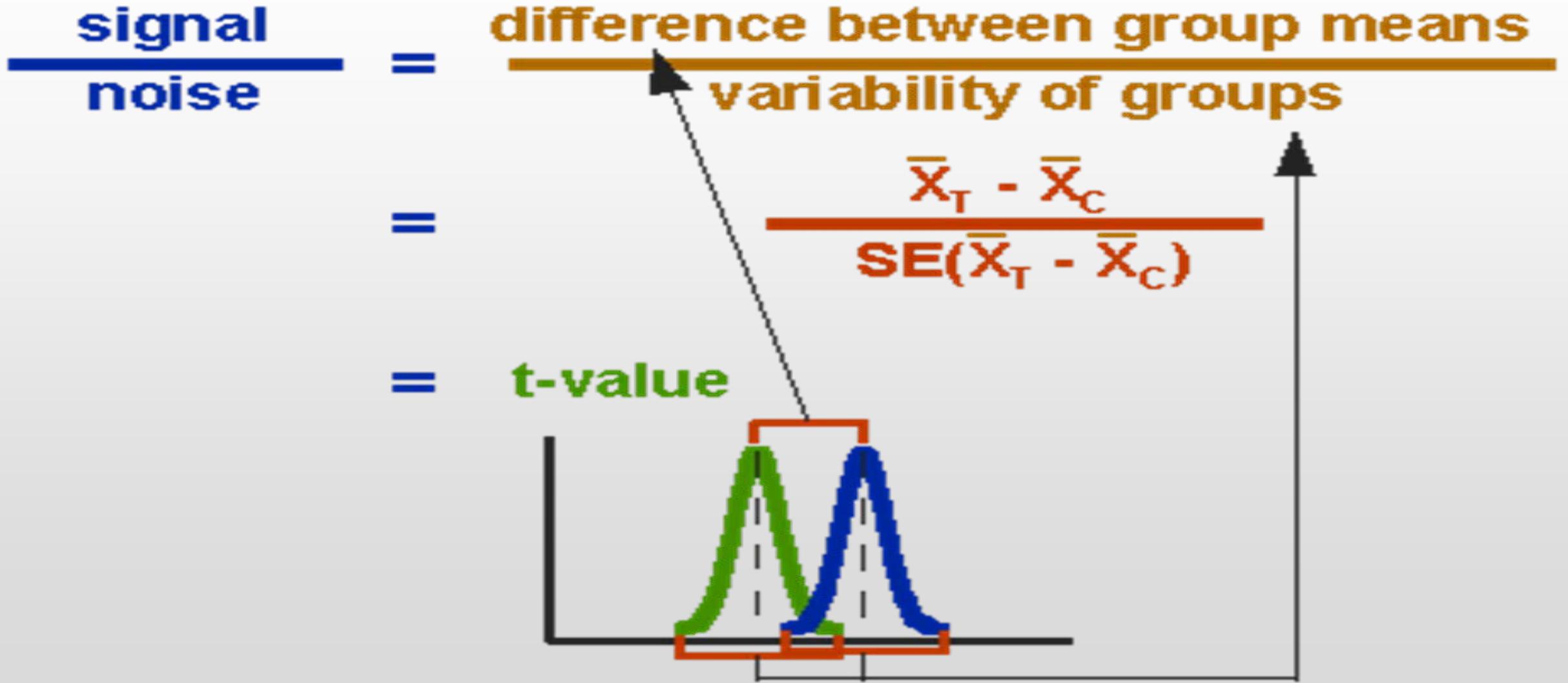
high  
variability



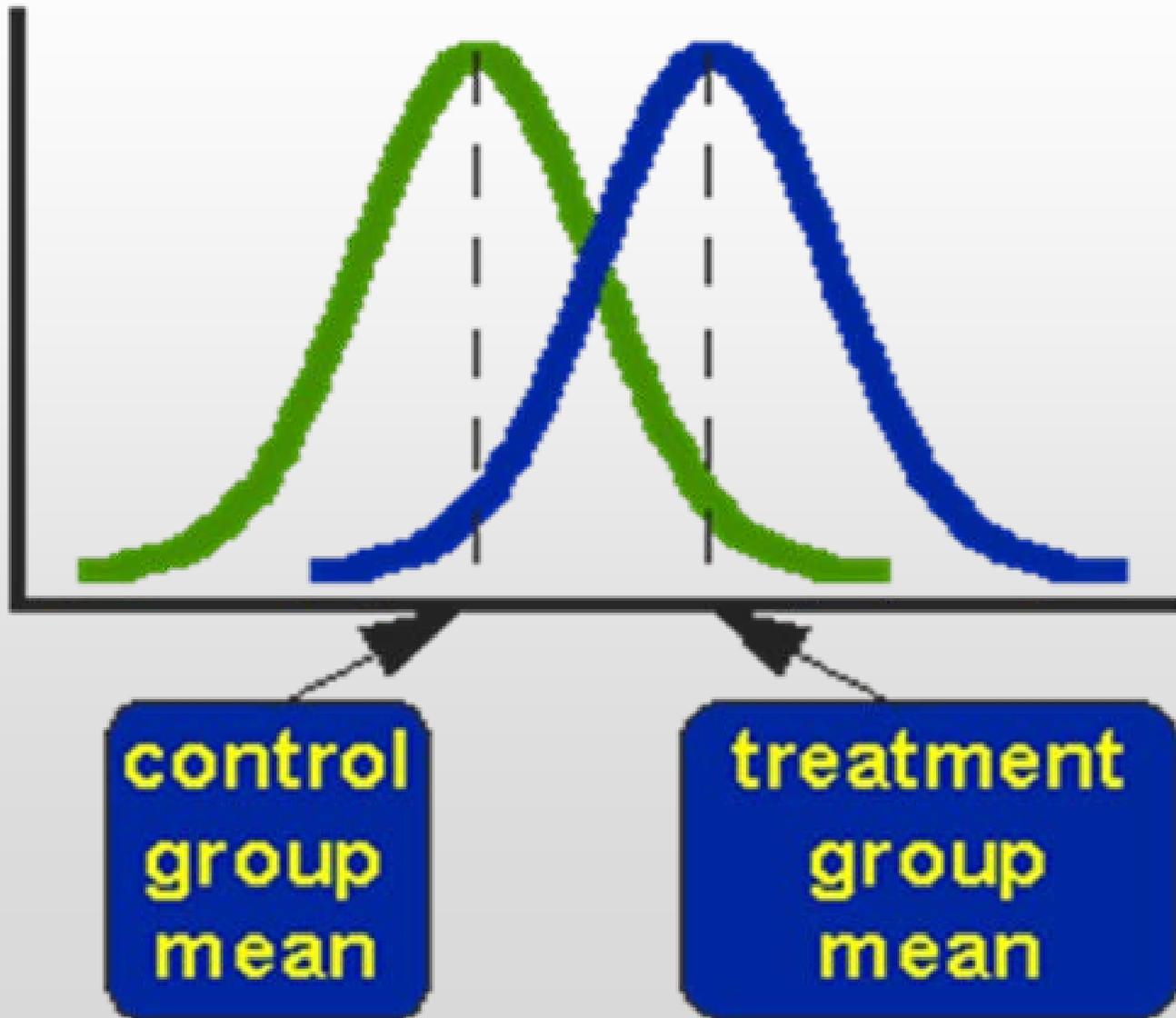
low  
variability



## t-Test Estimation :



## t – Statistics (Value) :



$$SE(\bar{X}_T - \bar{X}_C) = \sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}$$

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}}$$

# One-sample t-Test

- A hypothesis test to test whether the mean of a population is different from some known value,

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

## Hypothesis :

Null hypothesis  $H_0 : \mu \leq \mu_C$  (The sample data are not significantly different than the hypothesized mean).

Alternate hypothesis  $H_a : \mu > \mu_C$  (The sample data are significantly different than the hypothesized mean).

## Two-sample t-Test

The independent samples **t-test** is a hypothesis test for determining whether the population means of two independent groups are the same, given by the **t value**

formula ;

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Null hypothesis **H<sub>0</sub>** :  $\mu_1 = \mu_2$  (that the means of two groups of observations are identical).

Alternate hypothesis **H<sub>a</sub>** :  $\mu_1 \neq \mu_2$  (that the means of two groups of observations are not identical).

## Paired t-Test

A paired samples  $t$ -test is a hypothesis test for determining whether the population means of two dependent groups are the same.

$$t = \frac{\bar{x}_d}{s_d / \sqrt{n_d}}$$

where  $d$  is the sample mean difference score,

$sd$  is the standard deviation of the sample difference scores,

$nd$  is the number of paired observations in the sample.

## Analysis of Variance (ANOVA)

- Compares the means of three or more groups of an experiment.
- Determines whether any of those means are statistically and significantly different from each other.
- Uses F-test (F-ratio, an overall test) rather than an individual t-Test.
- Estimates the variance (the variability that may exist) in a data.
- Reflects the different experimental designs and situations for which they have been developed.

## ANOVA (F - Test)

- A significant F - test means that there are some differences among the components of the data.
- A non-significant F - test means that there no differences.
- The larger the sample size, the smaller the F-ratio value.
- The smaller the sample size, the larger F-ratio.

# Power and Sample Size Determination

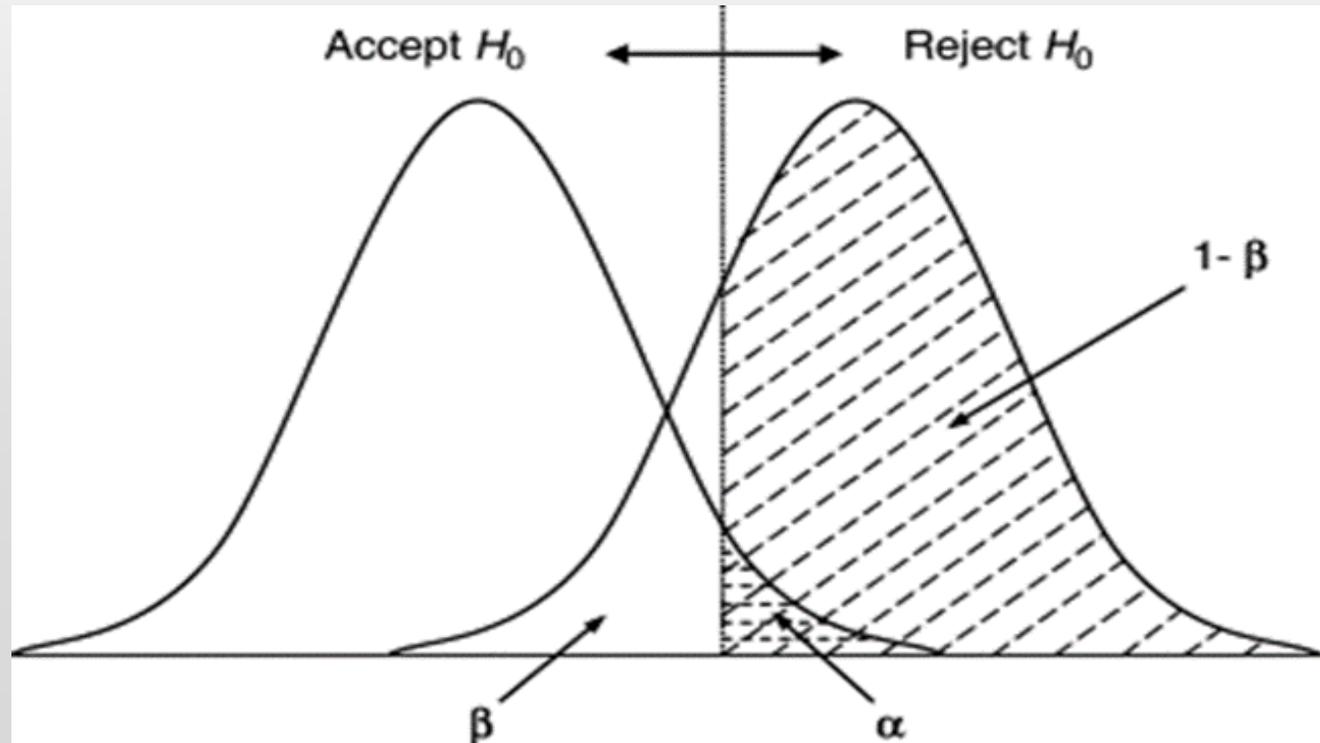
$$\begin{aligned}\text{Power} &= 1 - P(\text{Type II error}) \\ &= 1 - P(\text{do not reject } H_0 \mid H_1 \text{ is true}) \\ &= 1 - \beta = P(\text{reject } H_0 \mid H_1 \text{ is true})\end{aligned}$$

Consider the hypothesis :

$$H_0: \mu = \mu_0 \text{ vs } H_1: \mu = \mu_1 > \mu_0$$

The power of this test is :

$$\begin{aligned}\text{Power} &= P(\text{reject } H_0 \mid H_1 \text{ is true}) \\ &= P(Z_0 > Z_{1-\alpha} \mid \mu = \mu_1 > \mu_0)\end{aligned}$$



# Power and Sample Size Determination

Power is a function of ;

- Standard Deviation ( $\sigma$ ),
- Sample Size ( $n$ ),
- Mean Difference (or effect size),
- Type I error ( $\alpha$ ).

# Power and Sample Size Determination

The power of the test is :

$$\text{Power} = P\left(Z_1 > Z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) = P(\text{reject } H_0 \mid H_1 \text{ is true}), \text{ and it}$$

depends on :

1.  $\sigma$  (standard deviation),  $\sigma \uparrow \Rightarrow \text{Power} \downarrow$
2.  $n$  (sample size),  $n \uparrow \Rightarrow \text{Power} \uparrow$
3.  $\alpha$  (significance level),  $\alpha \downarrow \Rightarrow \text{Power} \downarrow$
4.  $\mu_1 - \mu_0$  (Effect Size),  $ES \uparrow \Rightarrow \text{Power} \uparrow$

**Thank you**