2019-08-13

An Introductory Data Analysis: Lecture 2

Oladapo Olaitan
University of Massachusetts Medical School

University of Massachusetts Medical School

Department of Quantitative Health Sciences

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 $\mathbf{x_1, x_2, \ldots, x_n}$ of size **n** drawn from a population of size **N** with average $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$.

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

Note :

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

Standard Error

Given a **sample** $\mathbf{x_1, x_2, \ldots, x_n}$ 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 ≥ 30) :

$$\bar{x} \sim \mathbf{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Note :

If **σ** 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-α)% 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 **X** from a population **(N)** with **μ** and **σ.**

Consider random samples **(n)** drawn with replacement from the population,

thus,

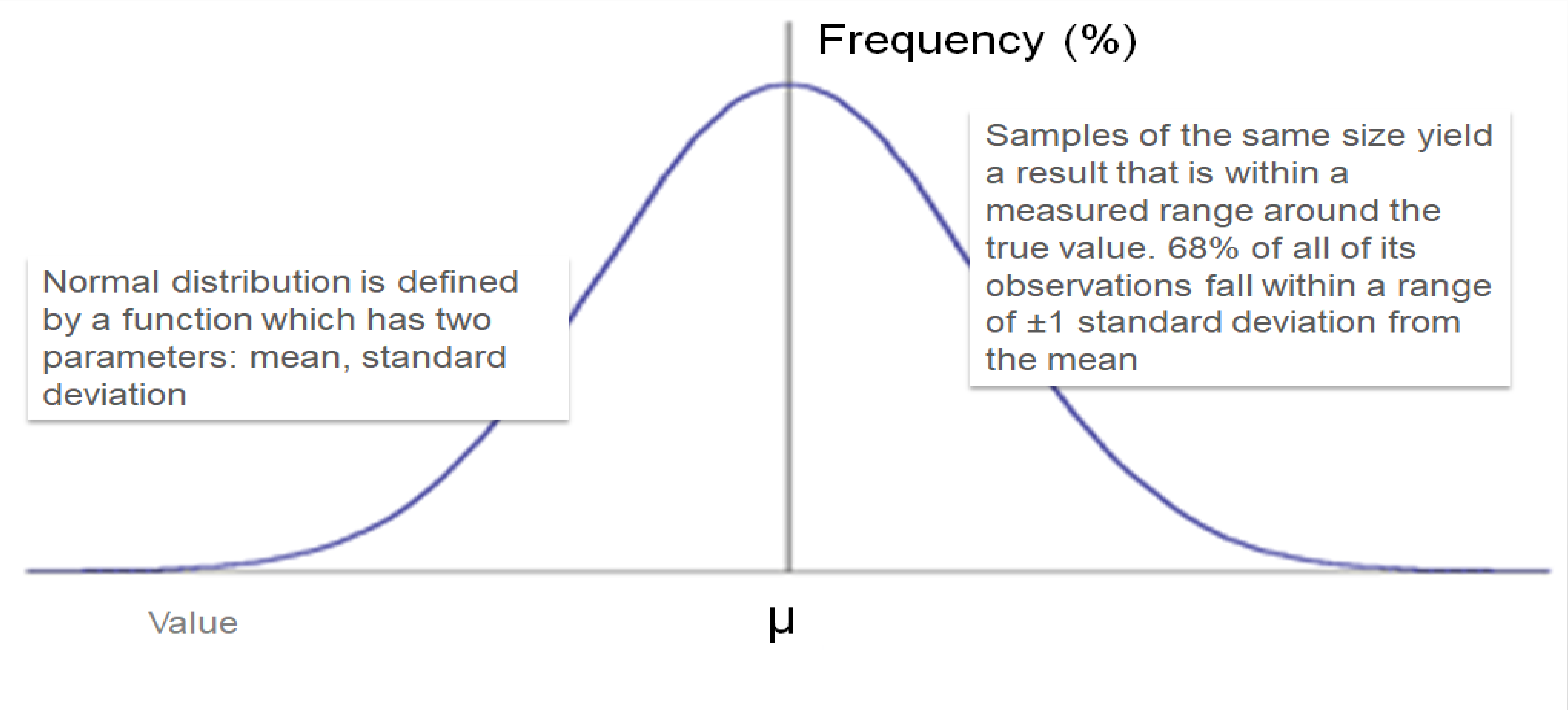
for **large "n",** the distribution of the **sample mean** $\bar{x}$ is approximately normally distributed with

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

Importance :

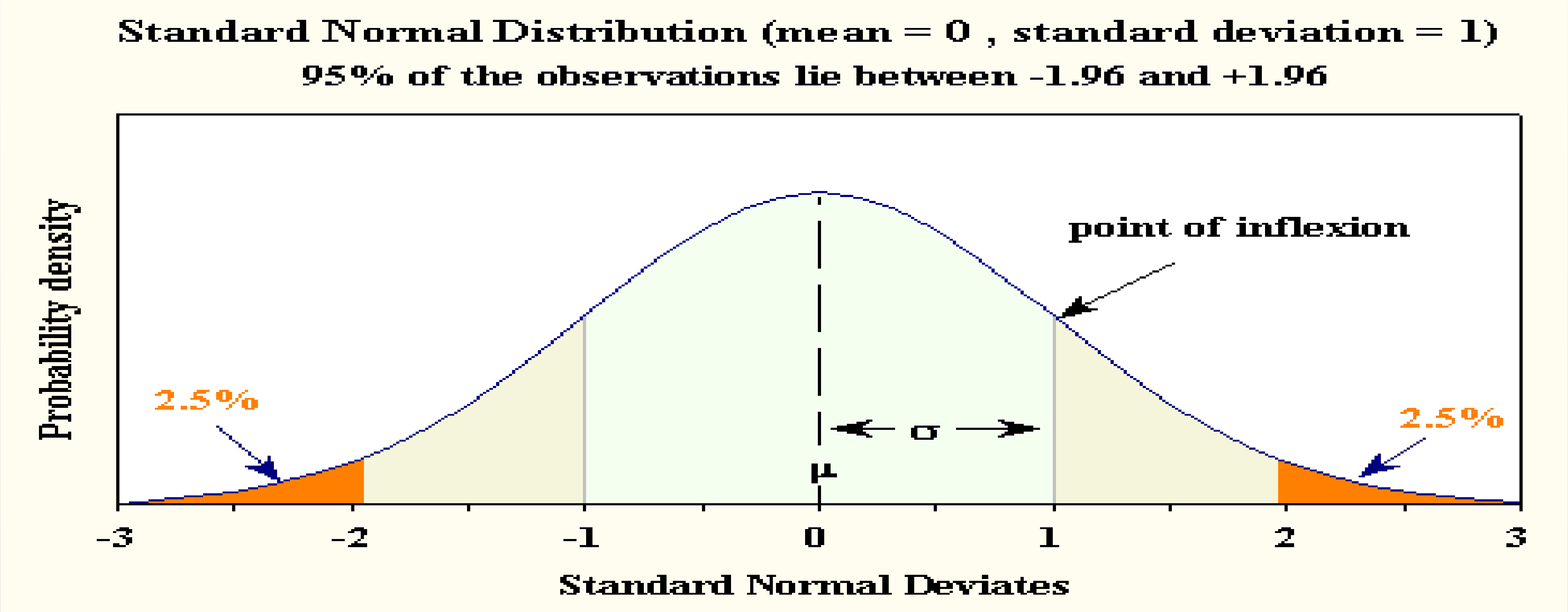
The distribution of the **sample mean** ($\bar{x}$) is approximately normal even if **X** does **not** follow **N(μ, σ).**

Normal Distribution

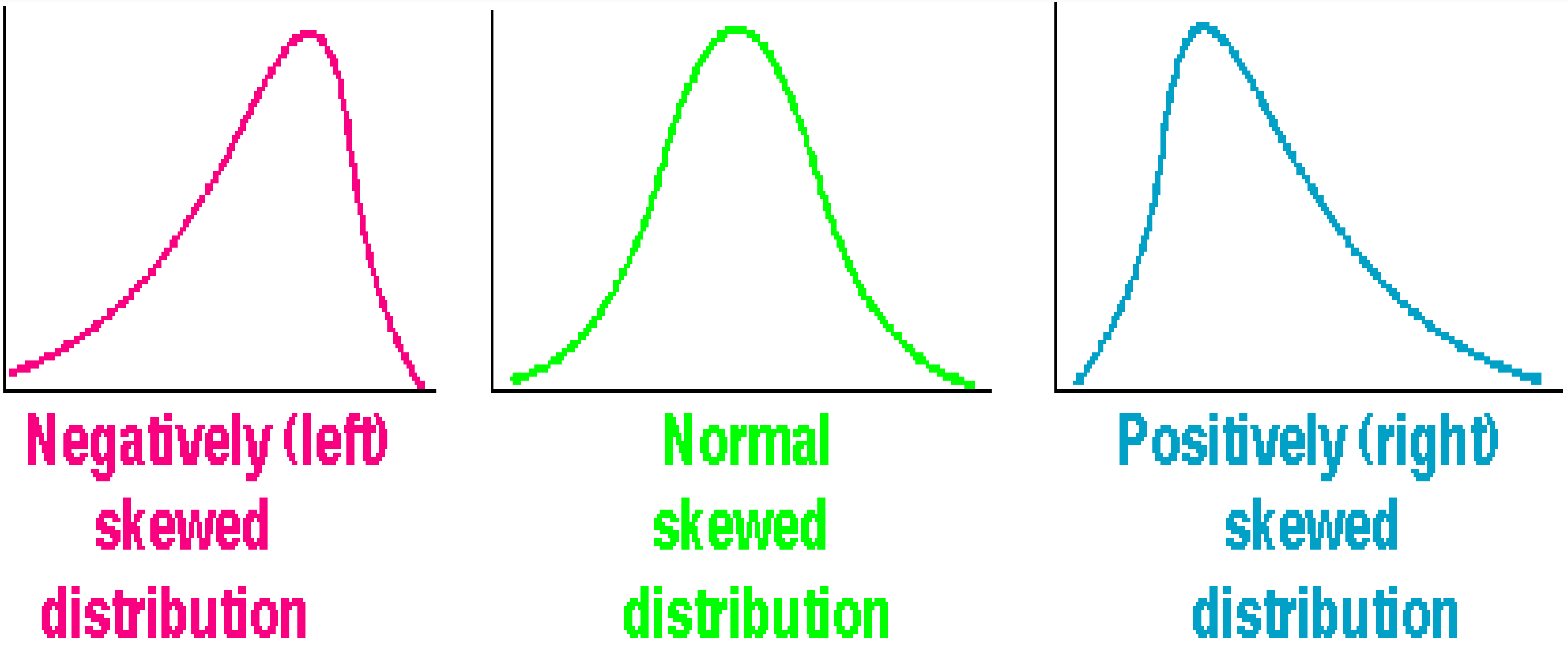


Standard Normal distribution

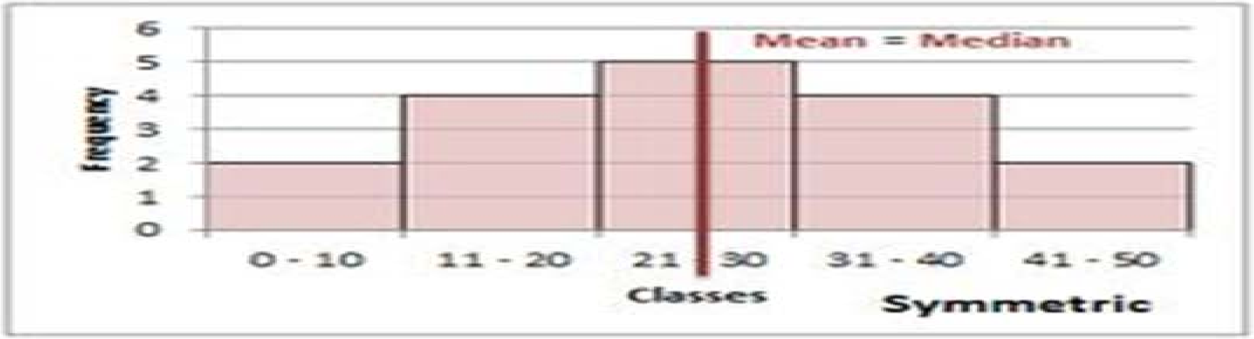
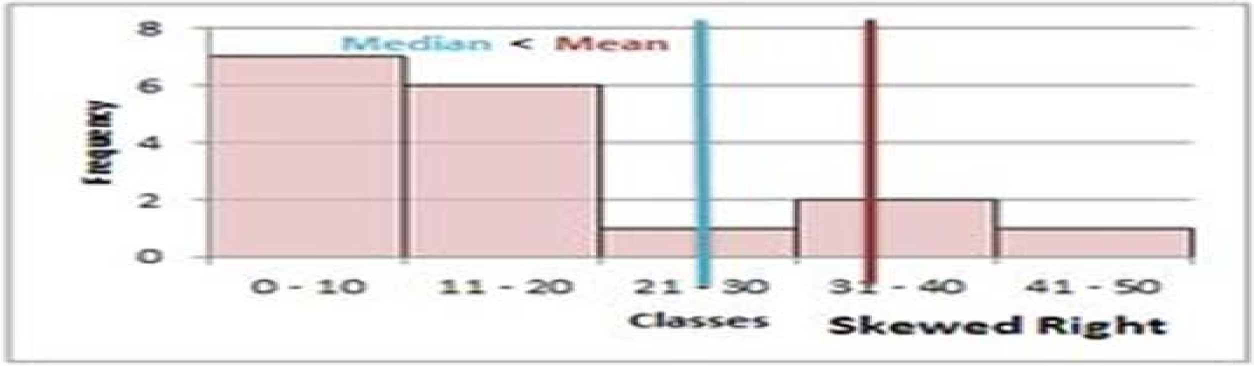
The Standard Normal distribution follows a normal distribution and is perfectly symmetric about zero (**0**).



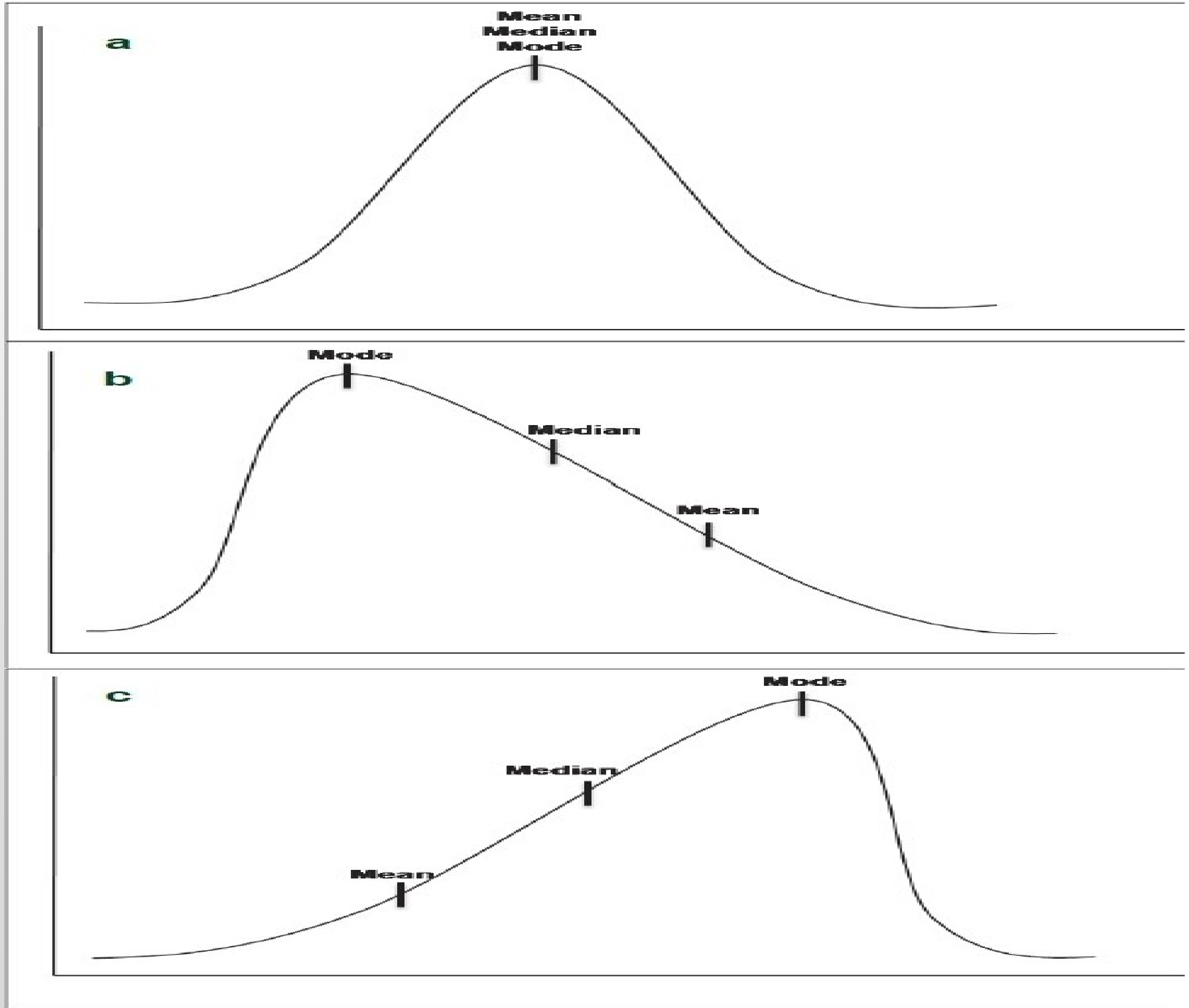
Types Of Distribution

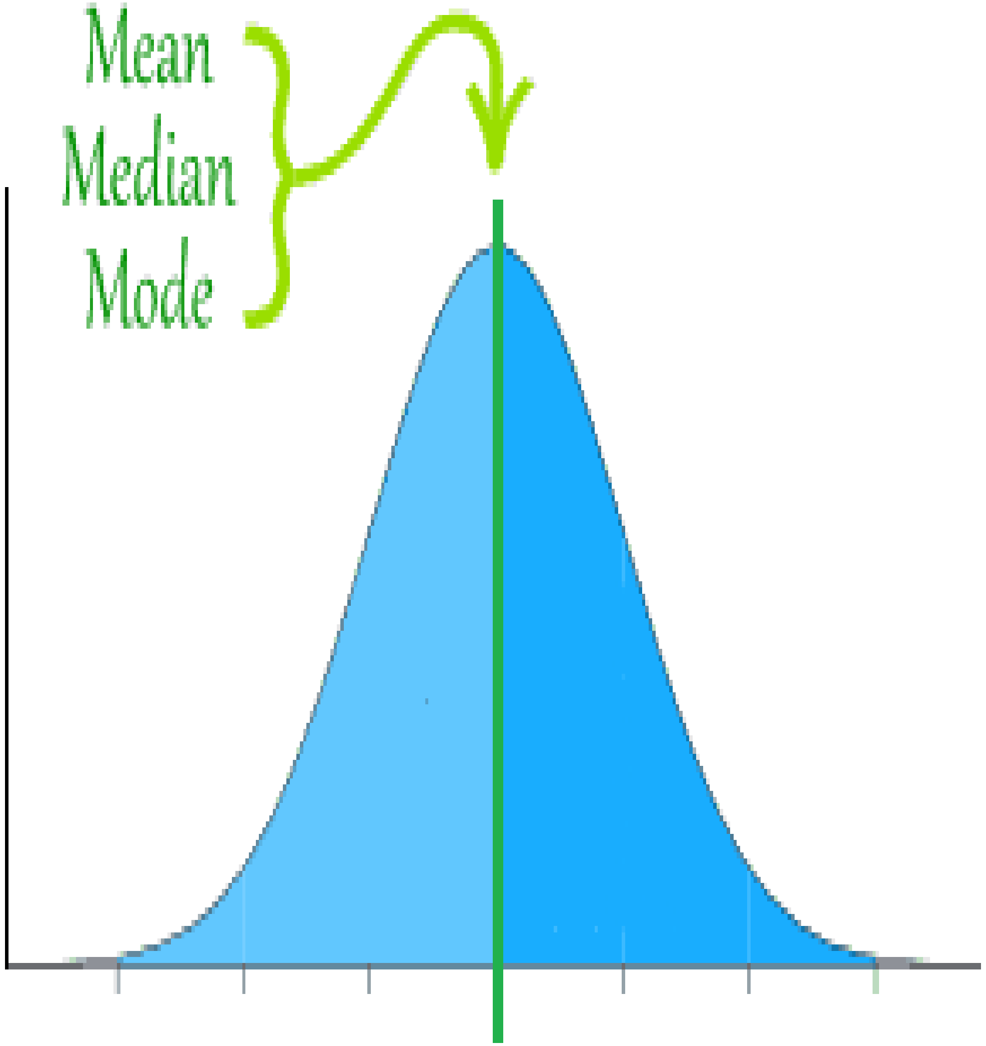


Types of Distribution

	When the data is symmetric, the mean and median are typically close together.
	When the data is skewed left, it is "pulled" to the left, and this drags the mean too low.
	When the data is skewed right, it is "pulled" to the right, and this drags the mean too high.

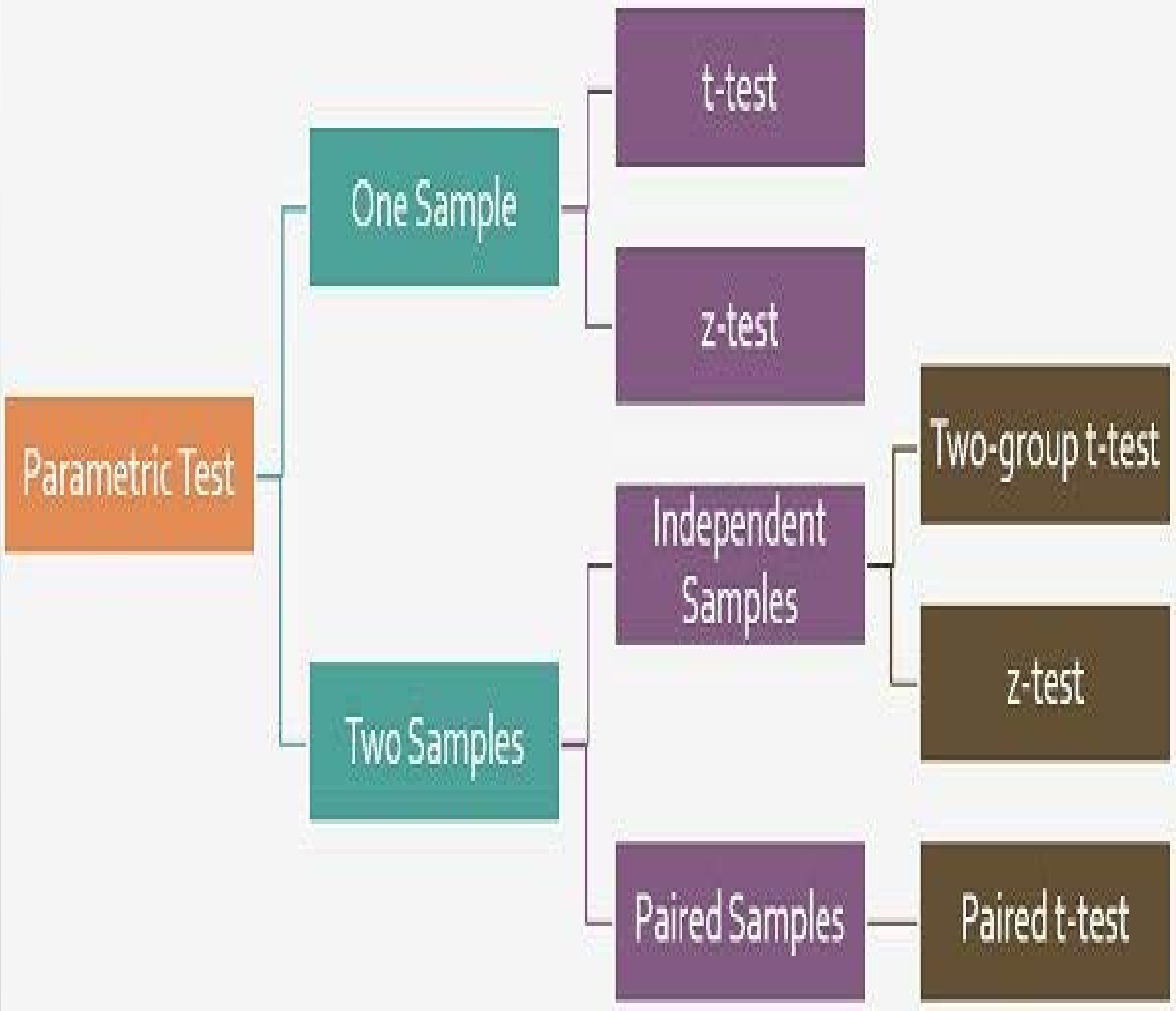
Distribution & Estimates



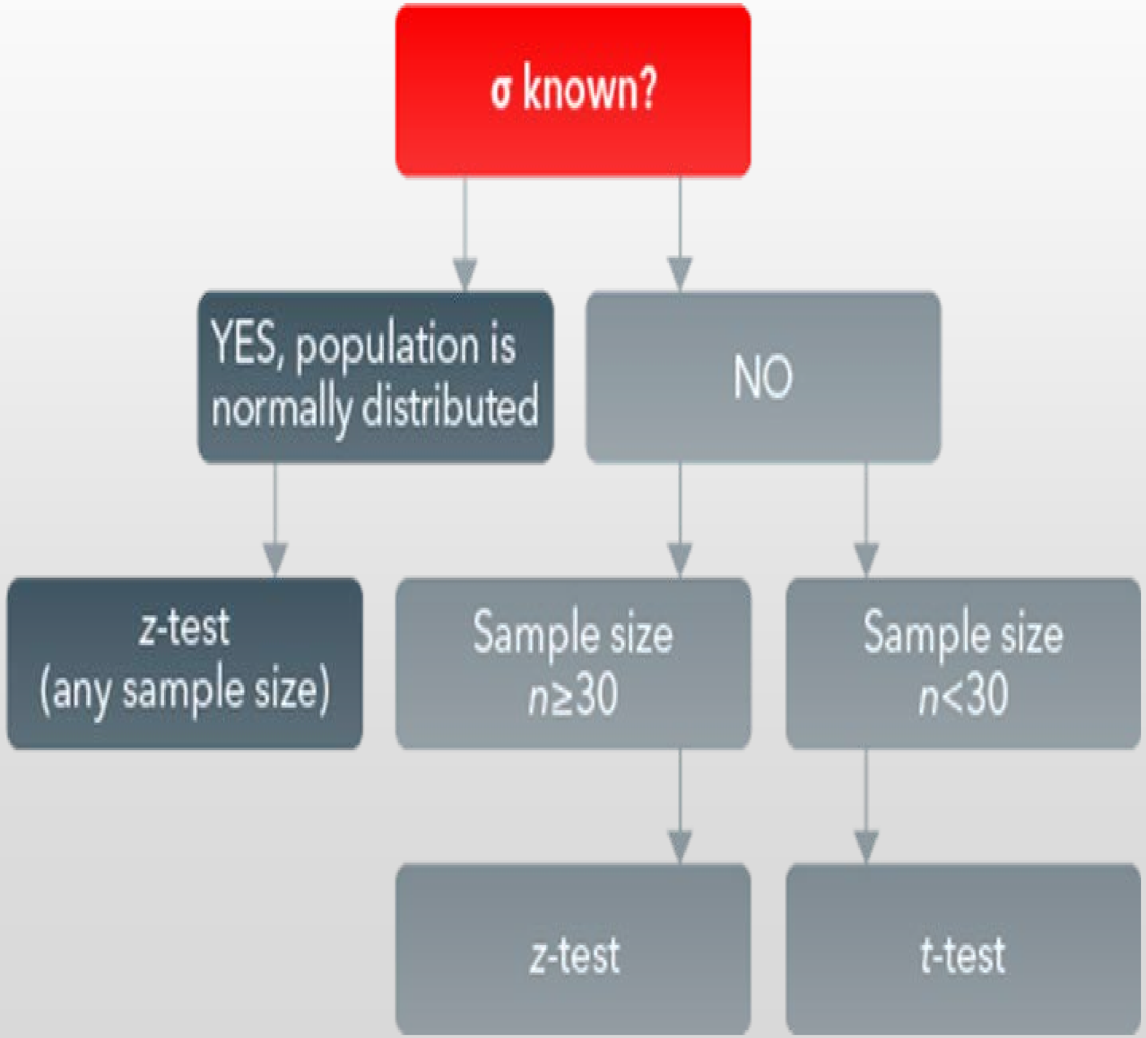


Hypothesis Testing

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

Probability : The probabilities of making errors can be assessed

- P(Type I error) = P(Rejecting $H_0|H_0$ 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, $\mathbf{H_0} : \mu \leq \mu_0$ and $\mathbf{H_a} : \mu > \mu_0$

- **One-tailed** (sided) **lower tail** test, $\mathbf{H_0} : \mu \geq \mu_0$ and $\mathbf{H_a} : \mu < \mu_0$

- **One-tailed** (sided) **upper tail** test, $\mathbf{H_0} : \mu \leq \mu_0$ and $\mathbf{H_a} : \mu > \mu_0$

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 $-z_{\alpha/2} \le z \le z_{\alpha/2}$,

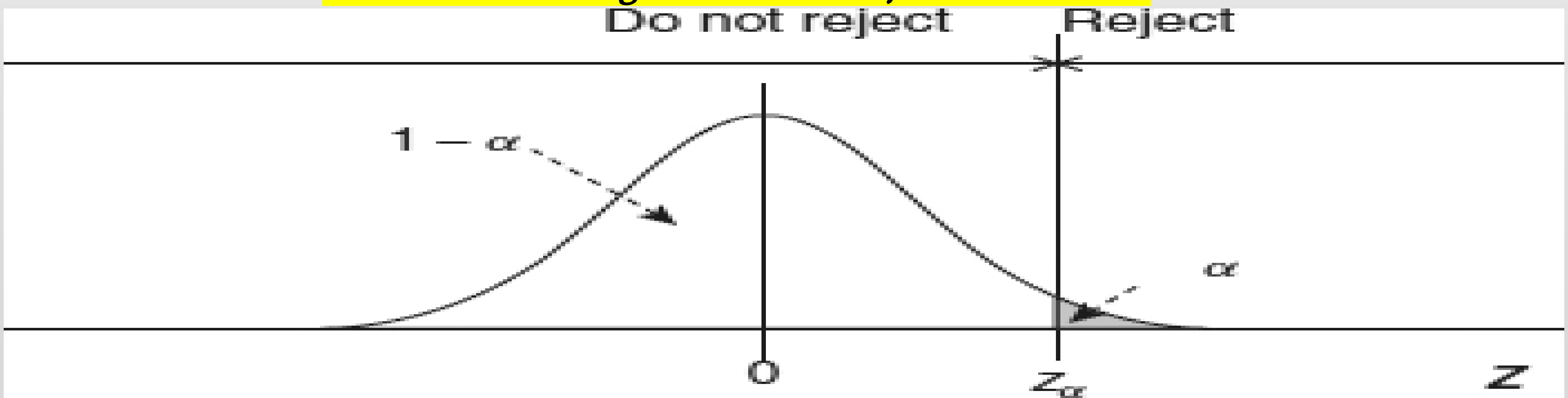
Reject H_0 if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$,

OR

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

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

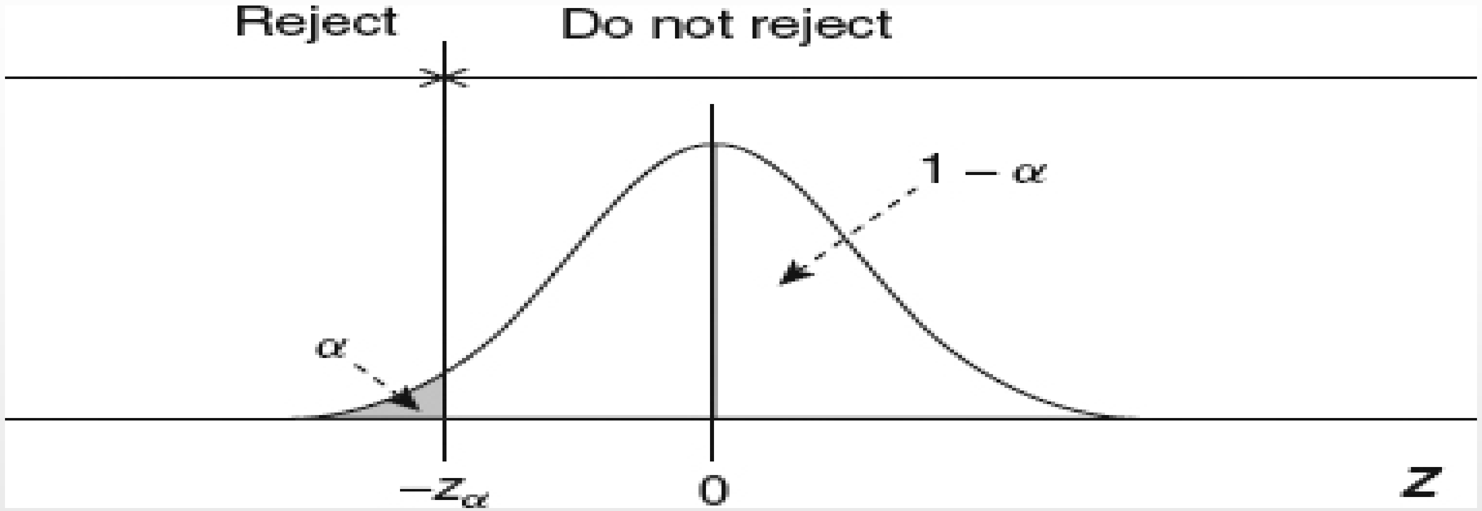
Decision Rule Using Critical Values for a Two-Tailed



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

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

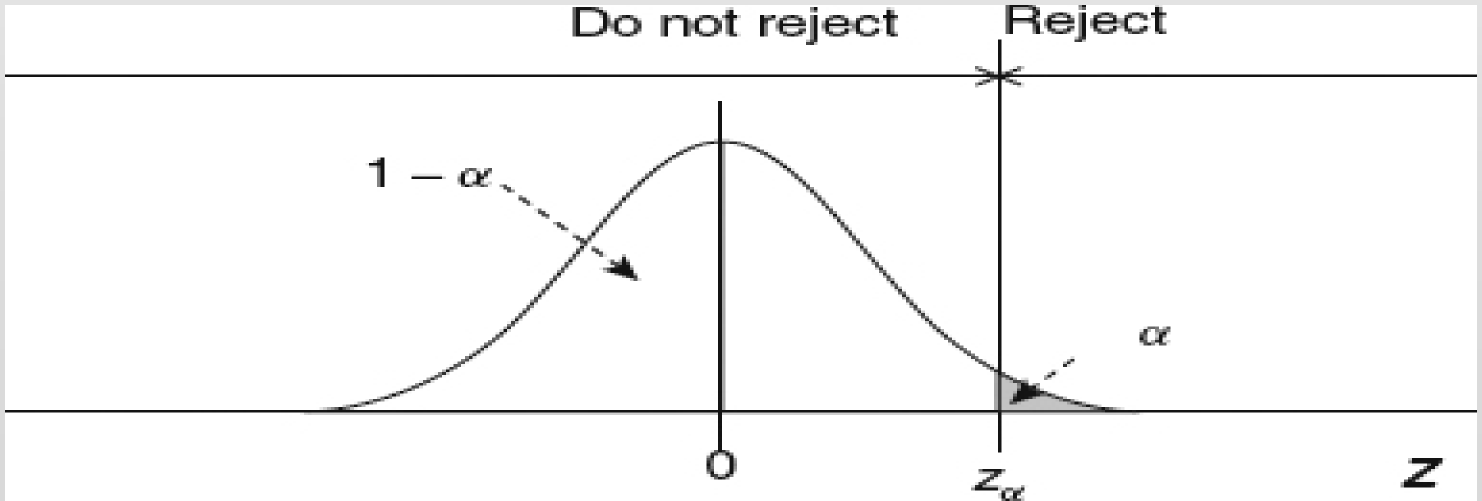
Reject H_o if $z < -z_\alpha$.



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

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

Reject H_o if $z > z_\alpha$.

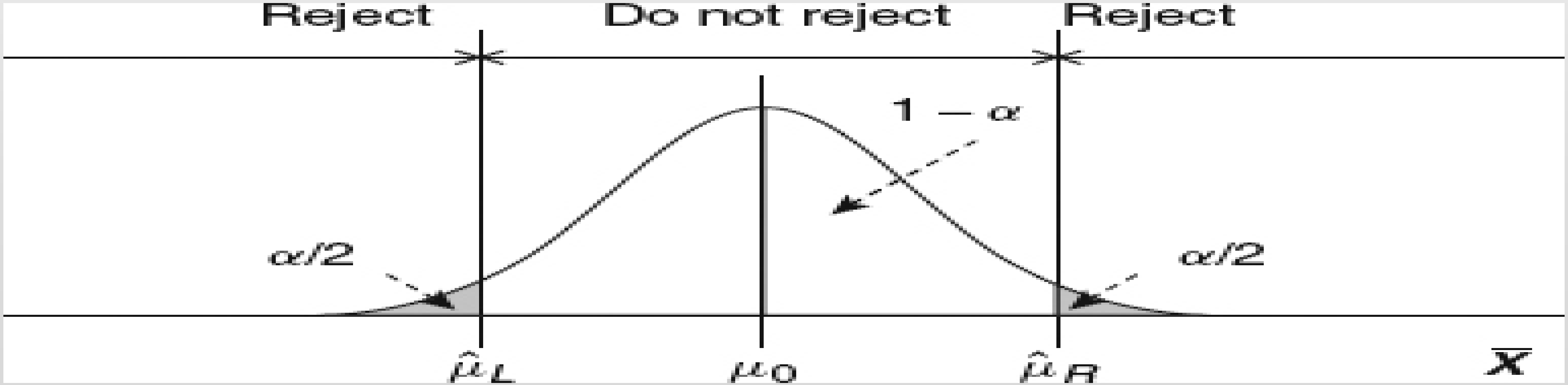


Decision Rules Using Action Limits

A Two-Tailed Test :

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

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

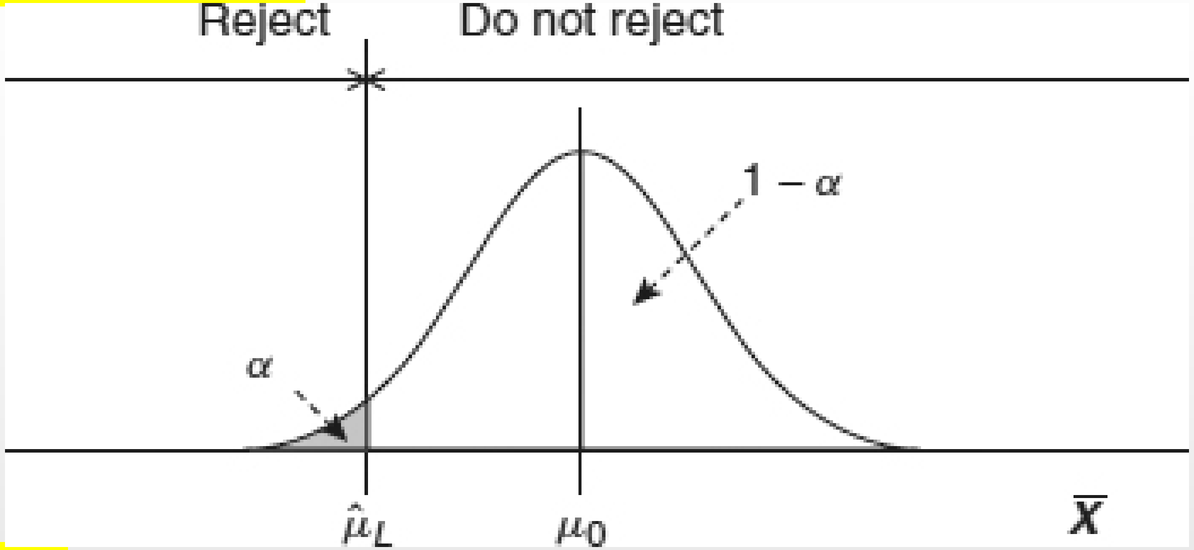


Decision Rule (Action Limits) - A One -Tailed Lower Tail Test

Do not reject H0 if $\overline{x} \geq \hat{\mu}_L$

Reject H0 if $\overline{x} < \hat{\mu}_L$

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

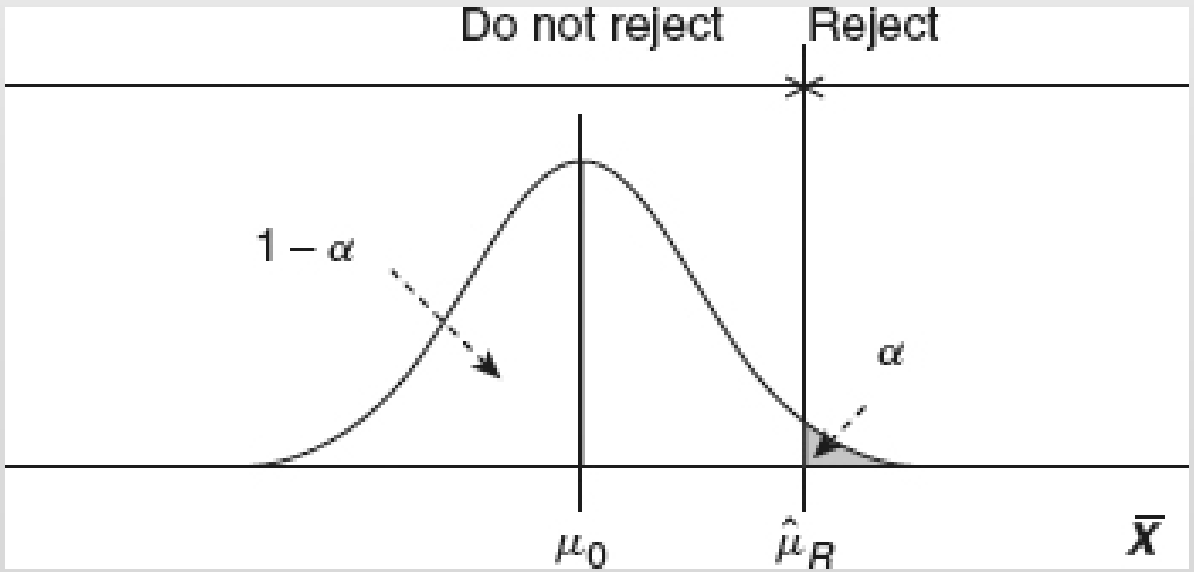


Decision Rule (Action Limits) - A One -Tailed Upper

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

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

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-value $\geq \alpha$

Reject H_0 if p-value $< \alpha$

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

***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

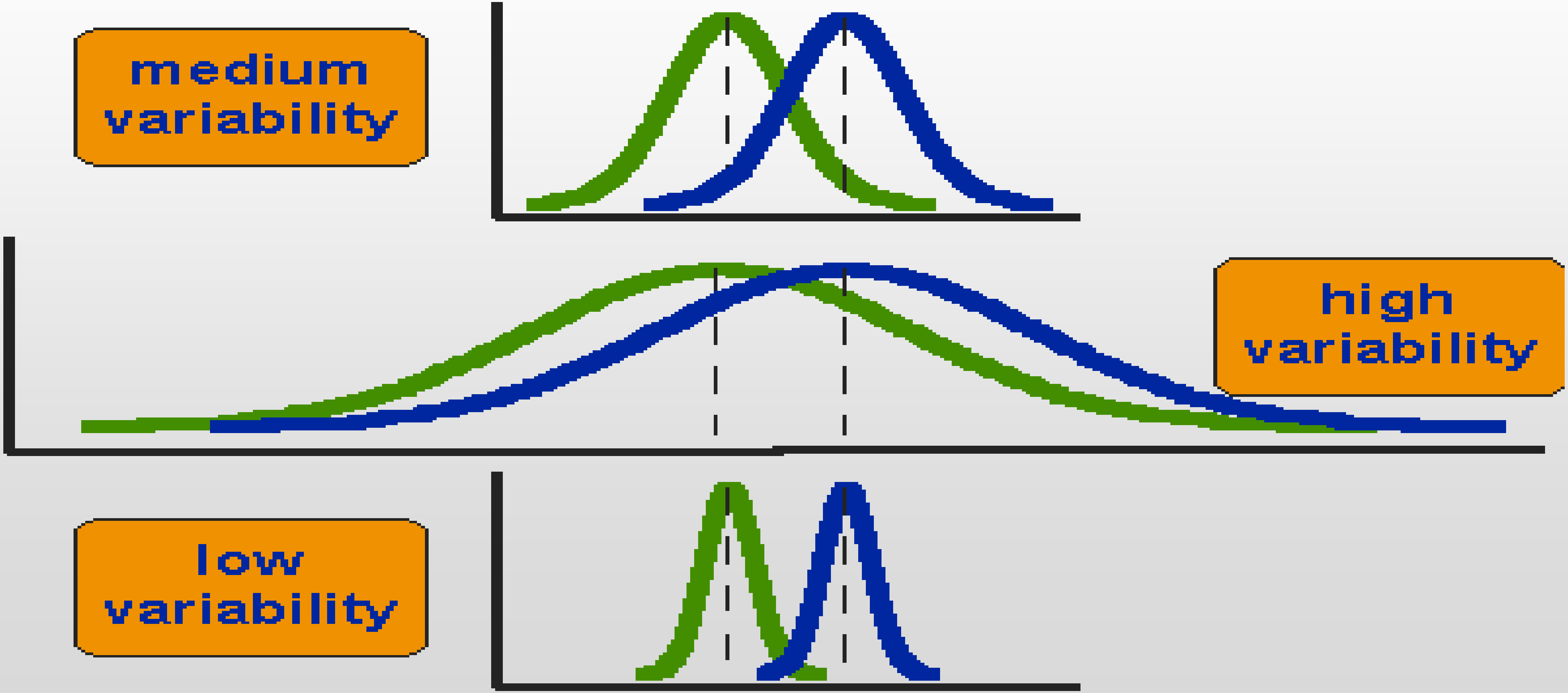
The t-Test is used to test whether the means of two group are *statistically* different from each other.

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 :

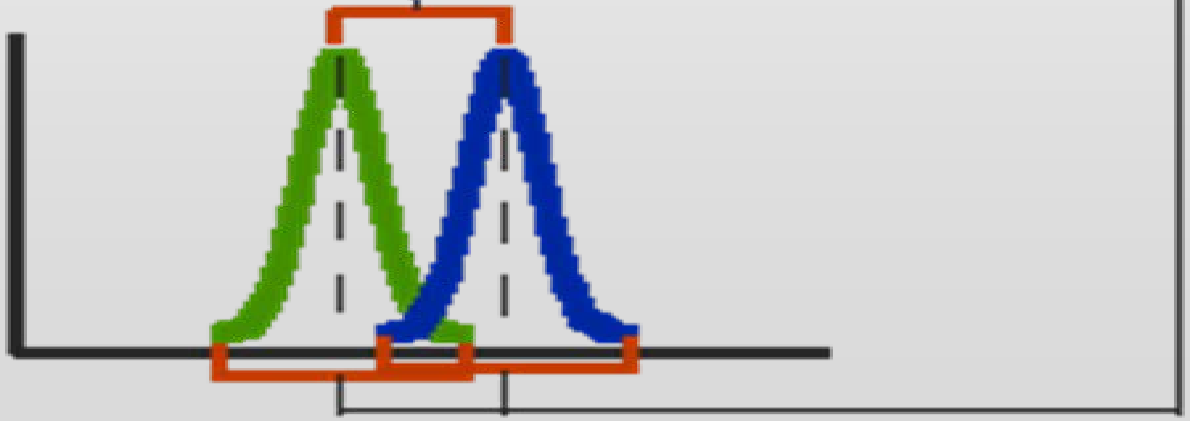


t-Test Estimation :

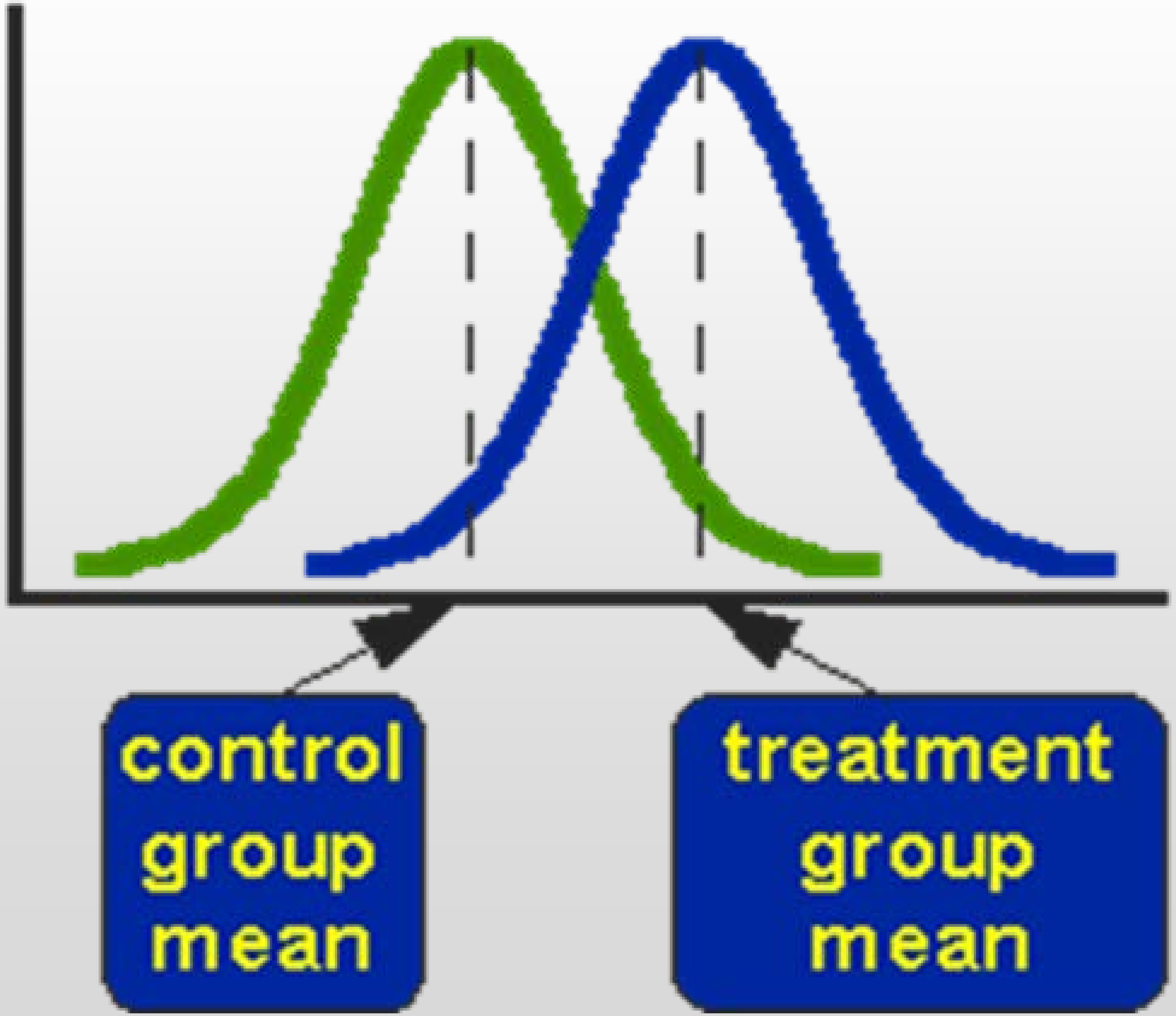
$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

$$= \text{t-value}$$



t – Statistics (Value) :



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

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

One-sample t-Test

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

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

Hypothesis :

Null hypothesis ***Ho*** **:** $\mu \le \mu c$ (The sample data are not significantly different than the hypothesized mean).

Alternate hypothesis ***Ha*** **:** $\mu > \mu c$ (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 ;

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

Null hypothesis ***H0*** **:** $\mu_1 = \mu_2$ (that the means of two groups of observations are identical).

Alternate hypothesis ***Ha*** **:** $\mu_1 \neq \mu_2$ (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.

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

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

Power $= 1 - P(\text{Type II error})$

$= 1 - P(\text{do not reject } H_0 \mid H_1 \text{ is true})$

$= 1 - \beta = P(\text{reject } H_0 \mid H_1 \text{ is true})$

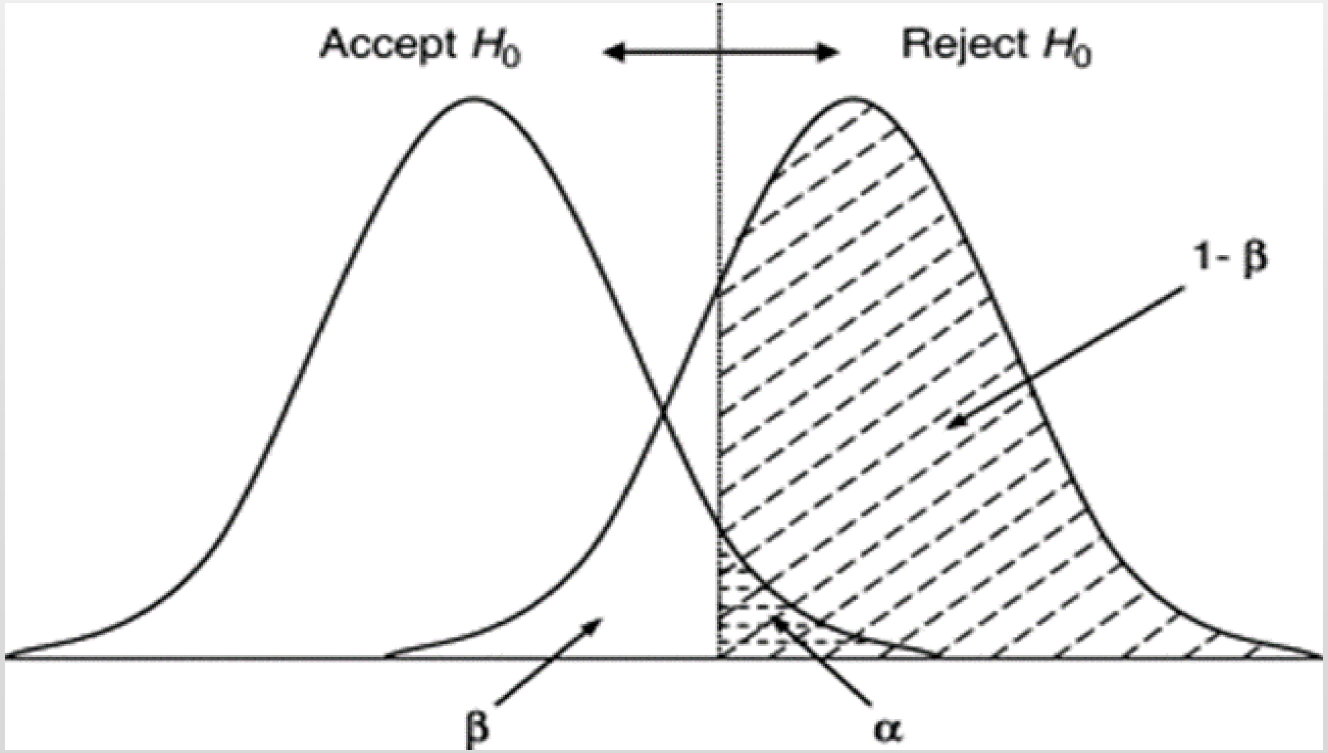
Consider the hypothesis :

$H_0: \mu=\mu_0$ vs $H_1: \mu = \mu_1 > \mu_0$

The power of this test is :

Power $= P(\text{reject } H_0 \mid H_1 \text{ is true})$

$= P(Z_0 > Z_{1-\alpha} \mid \mu = \mu_1 > \mu_0)$



Power and Sample Size Determination

Power is a function of ;

- Standard Deviation (σ),
- Sample Size (n),
- Mean Difference (or effect size),
- Type I error (α).

Power and Sample Size Determination

The power of the test is :

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

depends on :

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

Thank you