College of Science
Department of Statistics & OR

STAT 109
Biostatistics

Chapter 6:

Using Sample Data to Make Estimations About Population Parameters

August 2025

NOTE: This presentation is based on the presentation prepared thankfully by Professor Abdullah al-Shiha.

6.1 Introduction

6.2 Confidence Interval for a Population Mean (μ)

6.3 The t Distribution: (Confidence Interval Using t)

6.4 Confidence Interval for the Difference between Two Population Means ($\mu_1 - \mu_2$)

6.5 Confidence Interval for a Population Proportion (p)

6.6 Confidence Interval for the Difference Between Two Population Proportions (p_1-p_2)

6.1 Introduction

Statistical Inferences: (Estimation and Hypotheses Testing)

It is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

There are two main purposes of statistics:

- Descriptive Statistics: (Chapter 1 & 2): Organization & summarization of the data
- Statistical Inference: (Chapter 6 and 7): Answering research questions about some unknown population parameters.

1) Estimation: (chapter 6)

Approximating (or estimating) the actual values of the unknown parameters:

- **Point Estimate:** A point estimate is single value used to estimate the corresponding population parameter.

- **Interval Estimate (or Confidence Interval):** An interval estimate consists of two numerical values defining a range of values that most likely includes the parameter being estimated with a specified degree of confidence.

2) Hypothesis Testing: (chapter 7)

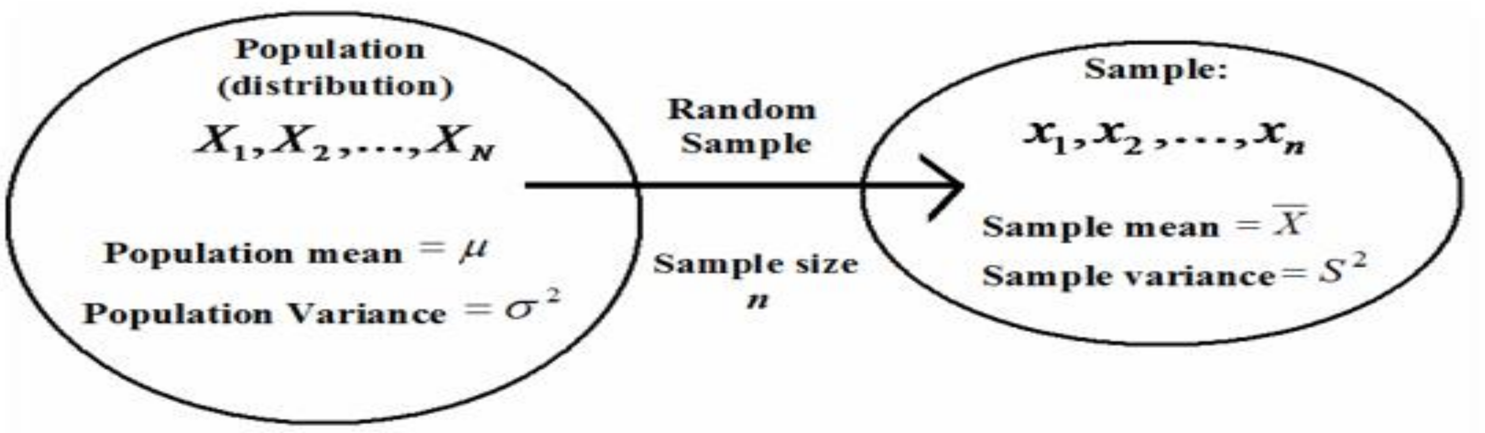
Answering research questions about the unknown parameters of the population (confirming or denying some conjectures or statements about the unknown parameters).

The Point Estimates of the Population Parameters :

	Population Parameters	**Point Estimator**
Mean	μ	$\bar{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S
Proportion	P	$\hat{p}$
The Difference between Two Means	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$
The Difference between Two Proportion	$P_1 - P_2$	$\hat{p}_1 - \hat{p}_2$

6.2 Confidence Interval for a Population Mean (μ) :

In this section we are interested in estimating the mean of a certain population (μ).



Population:

Population Size = N

Population Values: $X_1, X_2, \ldots, X_N$

Population Mean: $\mu = \dfrac{\sum_{i=1}^{N} X_i}{N}$

Population Variance: $\sigma^2 = \dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$

Sample:

Sample Size = n

Sample values: $x_1, x_2, \ldots, x_n$

Sample Mean: $\overline{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Sample Variance: $S^2 = \dfrac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1}$

(i) Point Estimation of μ:

A point estimate of the mean is a single number used to estimate (or approximate) the true value of μ .

- Draw a random sample of size n from the population:

$$x_1, x_2, \ldots, x_n$$

- Compute the sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Result:

The sample mean $\mathbf{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i}$ is a "good" point estimator of the population mean (μ).

(ii) Confidence Interval (Interval Estimate) of μ:

An interval estimate of μ is an interval (L,U) containing the true value of μ "with a probability of $1-\alpha$ ".

* $1-\alpha$ = is called the confidence coefficient (level) (confidence level), degree of confidence.

* L = lower limit of the confidence interval

* U = upper limit of the confidence interval

How to calculate the value of α

$(1-\alpha)$% Confident level

Example (1) :

If we are 95% Confident level , find the value of α ?

$$\alpha = \frac{5}{100} = 0.05$$

Example (2) :

If we are 99% Confident level , find the value of α ?

$$\alpha = \frac{1}{100} = 0.01$$

Example (3) :

If we are 80% Confident level , find the value of α ?

$$\alpha = \frac{20}{100} = 0.20$$

Example (4) :

If we are 92% Confident level , find the value of α ?

$$\alpha = \frac{8}{100} = 0.08$$

Result: (For the case when σ is known)

(a) If $x_1, x_2, \ldots, x_n$ is a random sample of size n from a normal distribution with mean μ and known variance σ^2 , then:

A (1 − α) 100% confidence interval for μ is:

$$\overline{\mathrm{X}} \pm \mathrm{Z}_{1-\frac{\alpha}{2}}\ \sigma_{\overline{\mathrm{X}}}$$

$$\overline{\mathrm{X}} \pm \mathrm{Z}_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

$$\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

(b) If $x_1, x_2, \ldots, x_n$ is a random sample of size n from a non-normal distribution with mean μ and known variance σ^2 , and if the sample size n is large $(n \geq 30)$, then:

An approximate (1 − α) 100% confidence interval for μ is:

$$\overline{\mathrm{X}} \pm \mathrm{Z}_{1-\frac{\alpha}{2}} \ \hat{\sigma}_{\overline{\mathrm{X}}}$$

$$\overline{\mathrm{X}} \pm \mathrm{Z}_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

$$\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Note that :

1. We are (1 − α) 100% confident that the true value of μ belongs to the interval $\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$

2. Upper limit of the confidence interval =

$$\bar{\mathbf{X}} + \mathbf{Z}_{1-\frac{\alpha}{2}}\frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$$

3. Lower limit of the confidence interval

$$\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

4. $Z_{1-\frac{\alpha}{2}}$ = Reliability Coefficient

5. $Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ = margin of error = precision of the estimate.

6. In general, the interval estimate (confidence interval) may be expressed as follows:

estimator ± (reliability coefficient) × (standard Error)

estimator ± margin of error

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

6.3 The t Distribution: (Confidence Interval Using t)

We have already introduced and discussed the t distribution.

Result:

(For the case when σ is unknown + normal population + n small (n<30))

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and unknown variance σ^2, then:

A (1−α) 100% confidence interval for μ is:

$$\overline{\mathrm{X}} \pm \mathrm{t}_{1-\frac{\alpha}{2}}\ \hat{\sigma}_{\overline{\mathrm{X}}}$$

$$\overline{\mathrm{X}} \pm \mathrm{t}_{1-\frac{\alpha}{2}}\ \frac{s}{\sqrt{n}}$$

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\ ,\overline{X} + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$$

where the degrees of freedom is: df = ν = n-1

.

Note that:

1. We are (1−α) 100% confidence that the true value of μ belongs to the interval

$$\left(\underbrace{\bar{X}-t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}}_{\text{Lower limit}}\ ,\qquad \underbrace{\bar{X}+t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}}_{\text{Upper limit}}\right)$$

2. $\hat{\sigma}_{\bar{X}}=\frac{S}{\sqrt{n}}$ (estimate of the standard error of $\bar{X}$)

3. $t_{1-\frac{\alpha}{2}}$ = Reliability Coefficient.

4. In this case, we replace σ by S and Z by t.

5. In general, the interval estimate (confidence interval) may be expressed as follows:

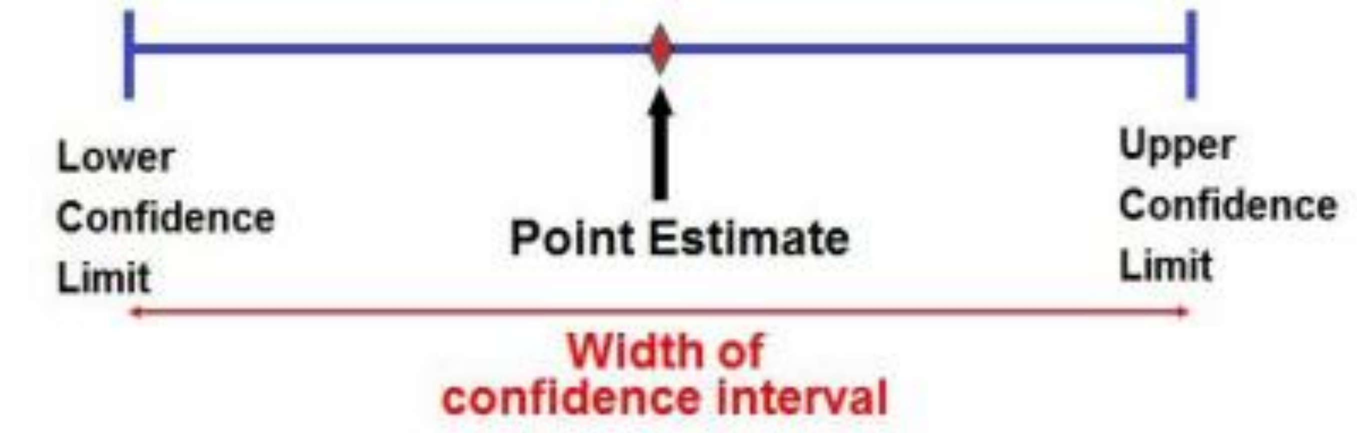
Estimator ± (Reliability Coefficient) × (Estimate of the Standard Error)

estimator ± margin of error

$$\bar{X} \pm t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

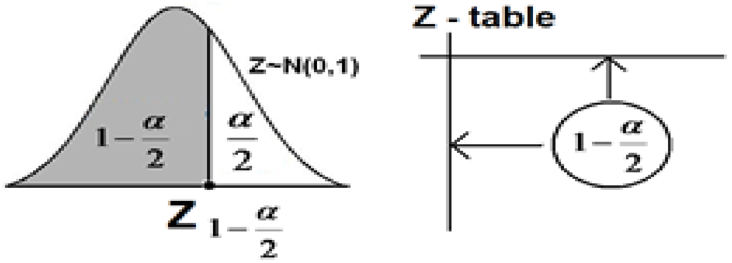
The length of the $(1-\alpha)\%$ confidence interval

Length = Upper limit - Lower limit

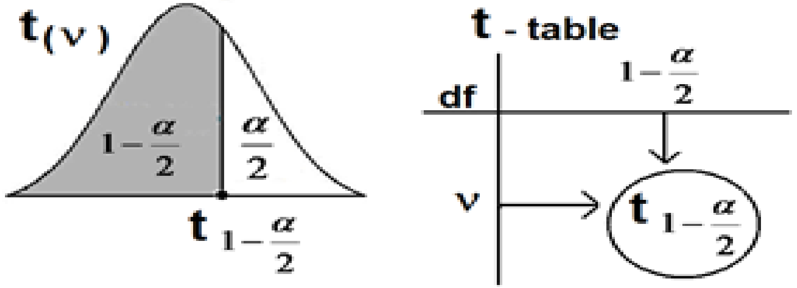


Notes: (Finding Reliability Coefficient)

(1) We find the reliability coefficient $Z_{1-\frac{\alpha}{2}}$ from the Z-table as follows:



(2) We find the reliability coefficient $t_{1-\frac{\alpha}{2}}$ from the t-table as follows: $(df = \nu = n-1)$



Example:

Suppose that $Z \sim N(0,1)$. Find $Z_{1-\frac{\alpha}{2}}$ for the following cases:

(1) $\alpha = 0.1$ (2) $\alpha = 0.05$ (3) $\alpha = 0.01$

Solution:

(1) For $\alpha = 0.1$:

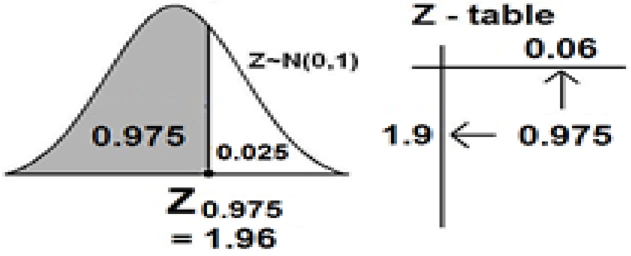
$$1-\frac{\alpha}{2} = 1-\frac{0.1}{2} = 0.95 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$$

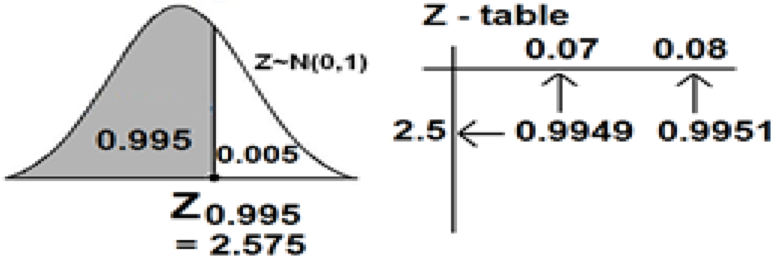
(2) For $\alpha = 0.05$:

$$1-\frac{\alpha}{2} = 1-\frac{0.05}{2} = 0.975 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.975} = 1.96.$$

(3) For $\alpha = 0.01$:

$$1-\frac{\alpha}{2} = 1-\frac{0.01}{2} = 0.995 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.995} = 2.575.$$





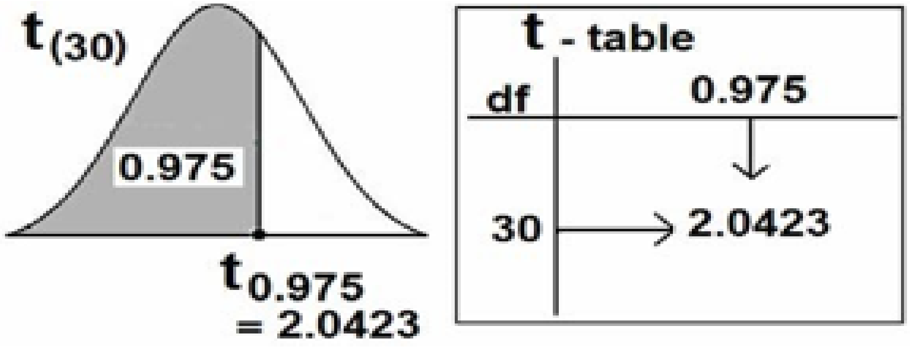
Example:

Suppose that t ~ t(30). Find $t_{1-\frac{\alpha}{2}}$ for α =0.05.

Solution:

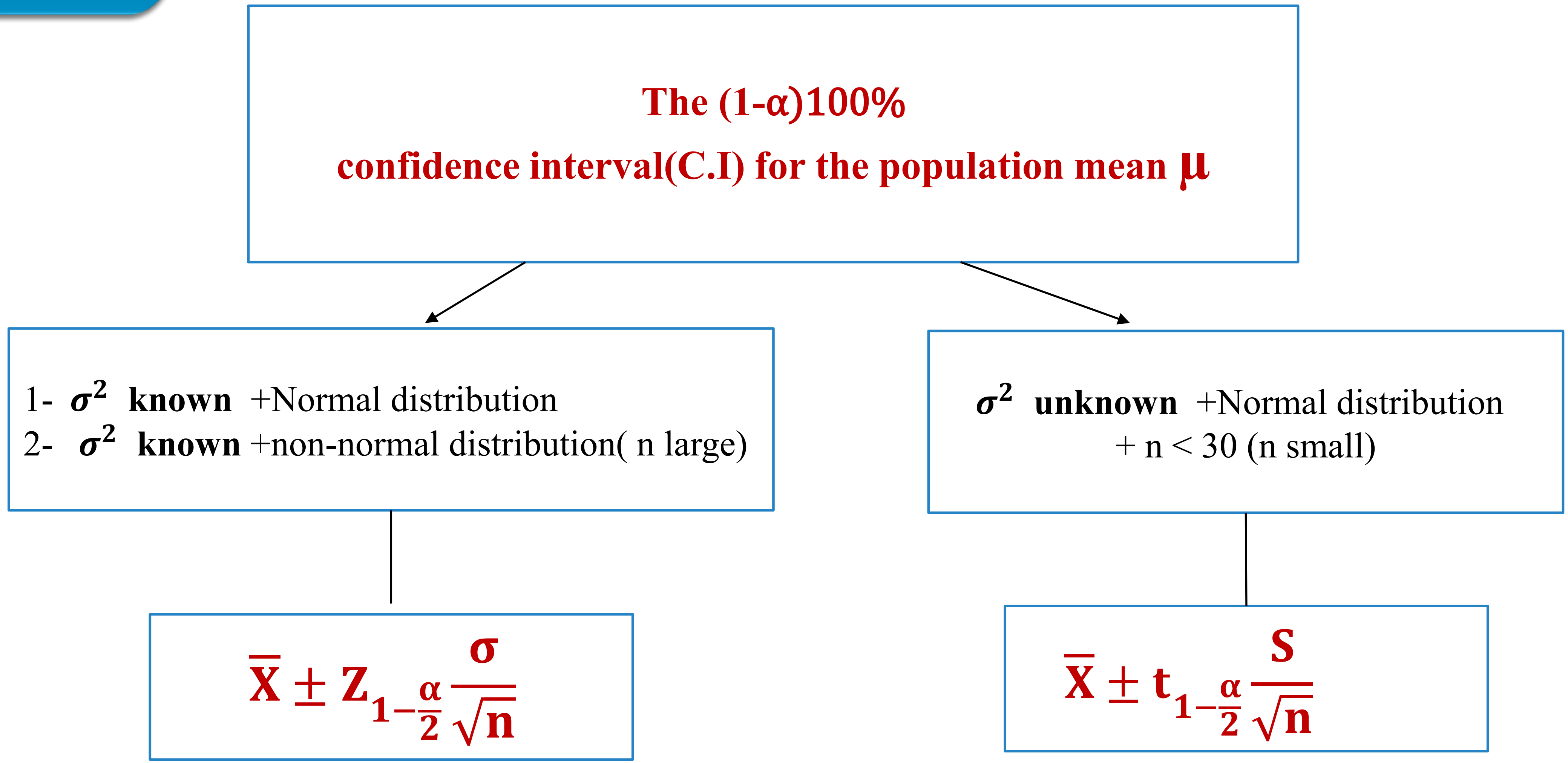
df = ν = 30

$$1-\frac{\alpha}{2} = 1-\frac{0.05}{2} = 0.975 \quad \Longrightarrow \quad t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.0423$$



t - table

df	0.975
30	2.0423



The (1-α)100%

confidence interval(C.I) for the population mean μ

1- σ^2 **known** +Normal distribution
2- σ^2 **known** +non-normal distribution(n large)

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

σ^2 **unknown** +Normal distribution
+ n < 30 (n small)

$$\bar{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

How to know if σ Known or Unknown :

σ known

- The population variance ………(σ^2)
- The population standard deviation…….. (σ)
- It is normal distribution with variance …..(σ^2)
- It is normal distribution with standard deviation …..(σ)

σ unknown: (Use S instead)

- Sample variance….. (S^2)
- Sample standard deviation ….. (S)
- If we have a sample of size ..(n),has mean ($\overline{X}$) with variance …(S^2)
- If we have a sample of size ..(n),has mean ($\overline{X}$) with standard deviation …(S)

Example: (The case where is σ^2 known)

Diabetic ketoacidosis is a potential fatal complication of diabetes mellitus throughout the world and is characterized in part by very high blood glucose levels. In a study on 123 patients living in Saudi Arabia of age 15 or more who were admitted for diabetic ketoacidosis, the mean blood glucose level was 26.2 mmol/l. Suppose that the blood glucose levels for such patients have a normal distribution with a standard deviation of 3.3 mmol/l.

(1) Find a point estimate for the **mean** blood glucose level of such diabetic ketoacidosis patients.

(2) Find a 90% confidence interval for the **mean** blood glucose level of such diabetic ketoacidosis patients.

Solution:

Variable = X = blood glucose level (quantitative variable).

Population = diabetic ketoacidosis patients in Saudi Arabia of age 15 or more.

Parameter of interest is: μ the mean blood glucose level.

Distribution is normal with standard deviation $\sigma = 3.3$.

σ^2 is known (σ^2=10.89)

X ~ Normal(μ , 10.89)

μ = ?? (unknown- we need to estimate μ)

Sample size: $n = 123$ (large)

Sample mean: $\bar{X} = 26.2$

Note :

$\boldsymbol{\sigma^2}$ **known** +Normal distribution

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

(1) Point Estimation:

We need to find a point estimate for μ .

$\bar{X} = 26.2$ is a point estimate for μ .

$\mu \approx 26.2$

(2) Interval Estimation (Confidence Interval = C. I.):

We need to find 90% C. I. for μ .

$90\% = (1-\alpha)\ 100\%$

$\alpha = \frac{10}{100} = 0.1 \Leftrightarrow \frac{\alpha}{2} = 0.05 \Leftrightarrow 1\text{-}\frac{\alpha}{2} = 0.95$

The reliability coefficient is: $Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$

90% confidence interval for μ is: $\left(\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$

$$\left(26.2 - (1.645)\frac{3.3}{\sqrt{123}}, 26.2 + (1.645)\frac{3.3}{\sqrt{123}}\right)$$

$$(26.2 - 0.4894714 , 26.2 + 0.4894714)$$

$$(25.710529 , 26.689471)$$

We are 90% confident that the true value of the mean μ lies in the interval $(25.71, 26.69)$, that is:

$$25.71 < \mu < 26.69$$

The length of the 90% confidence interval for μ :

length= Upper limit – Lower limit = 26.69 -25.71 = 0.98

Note: for this example, even if the distribution is not normal, we may use the same solution because the sample size n =123 is large.

Example: (The case where is σ^2 unknown)

A study was conducted to study the age characteristics of Saudi women having breast lump. A sample of 21 Saudi women gave a mean of 37 years with a standard deviation of 10 years. Assume that the ages of Saudi women having breast lumps are normally distributed.

(a) Find a point estimate for the **mean** age of Saudi women having breast lumps.

(b) Construct a 99% confidence interval for the **mean** age of Saudi women having breast lumps.

Solution:

X = Variable = age of Saudi women having breast lumps (quantitative variable).

Population = All Saudi women having breast lumps.

Parameter of interest is μ = the age mean of Saudi women having breast lumps.

X ~ Normal(μ , σ^2)

μ = ?? (unknown - we need to estimate μ)

σ^2 = ?? (unknown)

Sample size: $n = 21$

Sample mean: $\bar{X} = 37$

Sample standard deviation: S = 10

Degrees of freedom:

df = ν = n-1 = 21 – 1= 20

Note :

$\boldsymbol{\sigma^2}$ **Unknown** +Normal distribution+ n small

$$\bar{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

(a) Point Estimation:

We need to find a point estimate for μ.

$\bar{X} = 37$ is a "good" point estimate for μ.

μ ≈ 37 years

(b) Interval Estimation (Confidence Interval = C. I.):

We need to find 99% C. I. for μ .

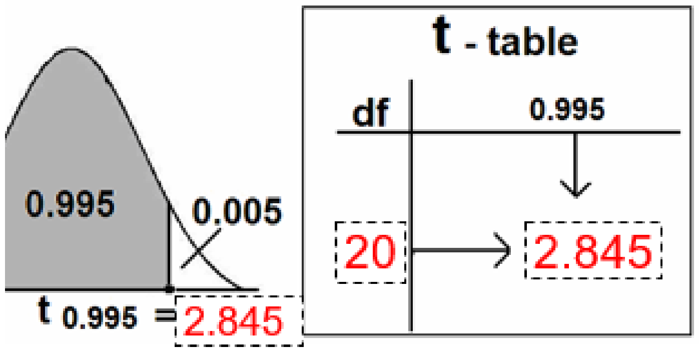
99% = (1− α) 100%

$$\alpha = \frac{1}{100} \Leftrightarrow \alpha = 0.01 \Leftrightarrow \frac{\alpha}{2} = \frac{0.01}{2} = 0.005 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.995$$

$$\mathrm{v} = \mathrm{df} = \mathrm{n} - 1 = 21 - 1 = 20$$

The reliability coefficient is:

$$t_{1-\frac{\alpha}{2}} = t_{0.995} = 2.845$$



99% confidence interval for μ is:

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$$

$$\left(37 - (2.845)\frac{10}{\sqrt{21}}, 37 + (2.845)\frac{10}{\sqrt{21}}\right)$$

$$(37 - 6.208, 37 + 6.208)$$

$$(30.792, 43.208)$$

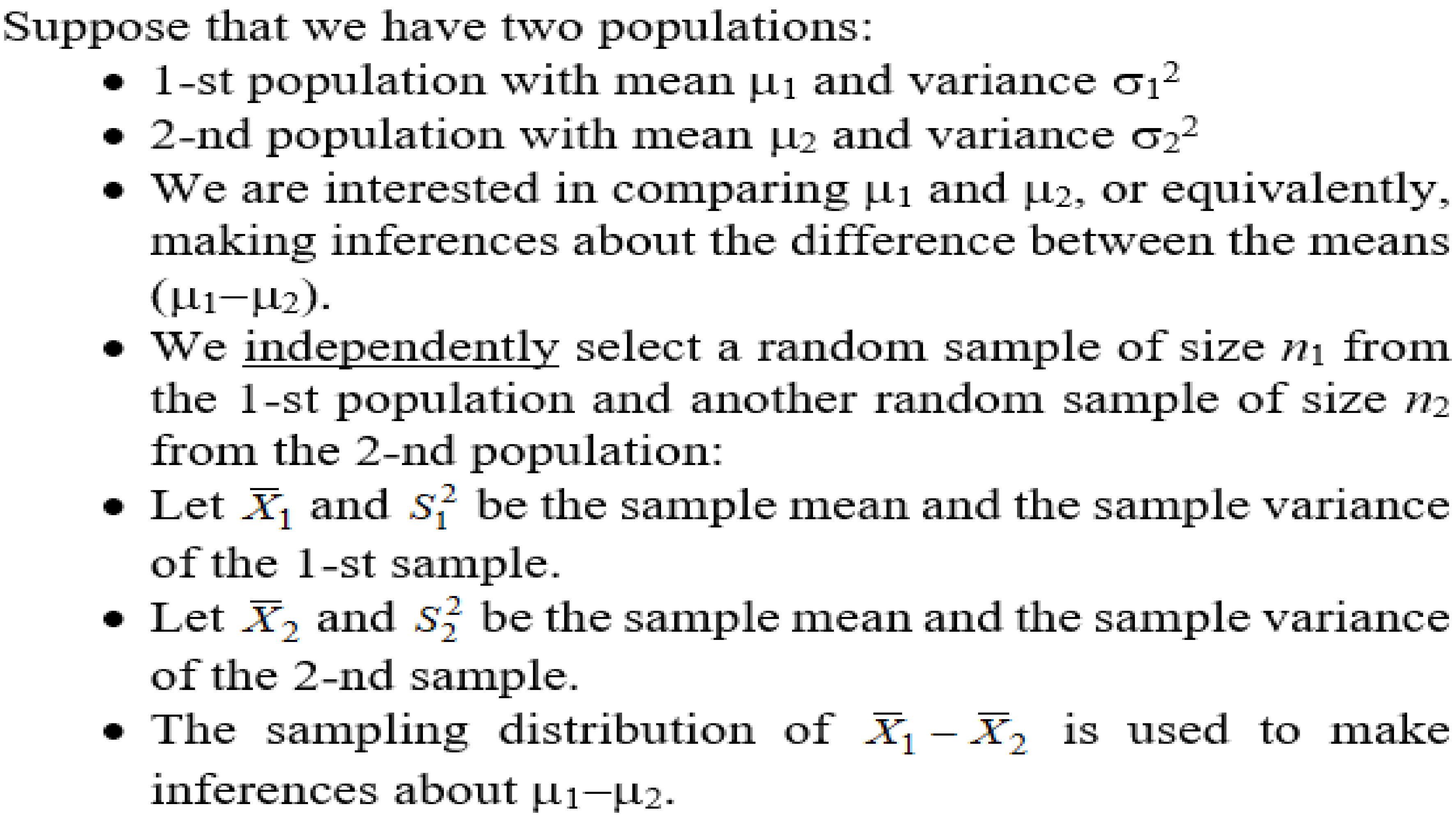
We are 99% confident that the true value of the mean μ lies in the interval $(30.792, 43.208)$, that is:

$$30.792 < \mu < 43.208$$

The length of the 99% confidence interval for μ :

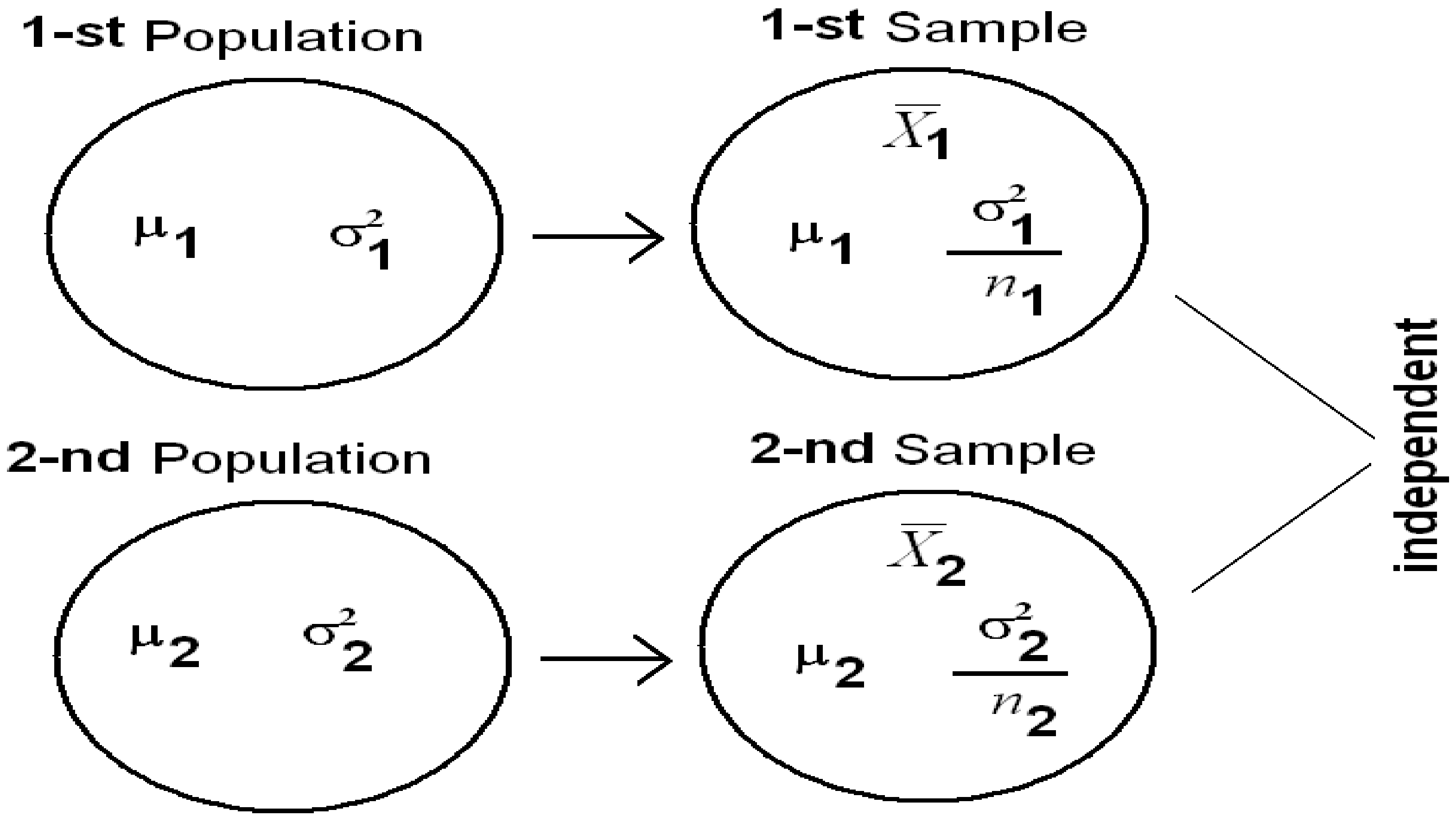
length= Upper limit – Lower limit = 43.208 -30.792 = 12.416

6.4 Confidence Interval for the Difference between Two Population Means ($\mu_1 - \mu_2$):



Suppose that we have two populations:

- 1-st population with mean μ_1 and variance $\sigma_1{}^2$
- 2-nd population with mean μ_2 and variance $\sigma_2{}^2$
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means ($\mu_1 - \mu_2$).
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let $\overline{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1-st sample.
- Let $\overline{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\overline{X}_1 - \overline{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.


1-st Population
μ_1 σ_1^2
1-st Sample
$\bar{X}_1$
μ_1 $\frac{\sigma_1^2}{n_1}$
2-nd Population
μ_2 σ_2^2
2-nd Sample
$\bar{X}_2$
μ_2 $\frac{\sigma_2^2}{n_2}$
independent

Recall:

1. Mean of $\bar{X}_1 - \bar{X}_2$ is: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

2. Variance of $\bar{X}_1 - \bar{X}_2$ is: $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

3. Standard error of $\bar{X}_1 - \bar{X}_2$ is: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

4. If the two random samples were selected from normal distributions (or non-normal distributions with large sample sizes) with known variances σ_1^2 and σ_2^2, then the difference between the sample means $(\bar{X}_1 - \bar{X}_2)$ has a normal distribution with mean $(\mu_1 - \mu_2)$ and variance $((\sigma_1^2 / n_1) + (\sigma_2^2 / n_2))$, that is:

- $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2 \, , \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$

- $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$

Point Estimation of $\mu_1 - \mu_2$:

Result:

$\bar{X}_1 \_ \bar{X}_2$ is a "good" point estimate for $\mu_1 - \mu_2$

Interval Estimation (Confidence Interval) of $\mu_1 - \mu_2$:

We will consider two cases

(i) First Case: σ_1^2 and σ_2^2 are known:

If σ_1^2 and σ_2^2 are known, we use the following result to find an interval estimate for $\mu_1 - \mu_2$.

Result:

A (1−α) 100% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{1-\frac{\alpha}{2}} \sigma_{\bar{X}_1 - \bar{X}_2}$$

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Estimator ± (Reliability Coefficient) × (Standard Error)

$$\left((\bar{X}_1 - \bar{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \, , (\bar{X}_1 - \bar{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

$$(\bar{X}_1 - \bar{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(ii) Second Case:

Unknown equal Variances: ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ **is unknown**):

If σ_1^2 and σ_2^2 are equal but unknown ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then the pooled estimate of the common variance σ^2 is

$$\mathbf{S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

where S_1^2 is the variance of the 1-st sample and S_2^2 is the variance of the 2-nd sample. The degrees of freedom of S_P^2 is

$$\text{df} = v = n_1 + n_2 - 2 \ .$$

We use the following result to find an interval estimate for $\mu_1 - \mu_2$ when we have normal populations with unknown and equal variances + n_1 and n_2 small .

Result:

A (1−α) 100% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

Estimator ± (Reliability Coefficient) × (Standard Error)

$$\left((\bar{X}_1 - \bar{X}_2) - t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} , (\bar{X}_1 - \bar{X}_2) + t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right)$$

where reliability coefficient $t_{1-\frac{\alpha}{2}}$ is the t-value with df = $v = n_1 + n_2 - 2$ degrees of freedom.

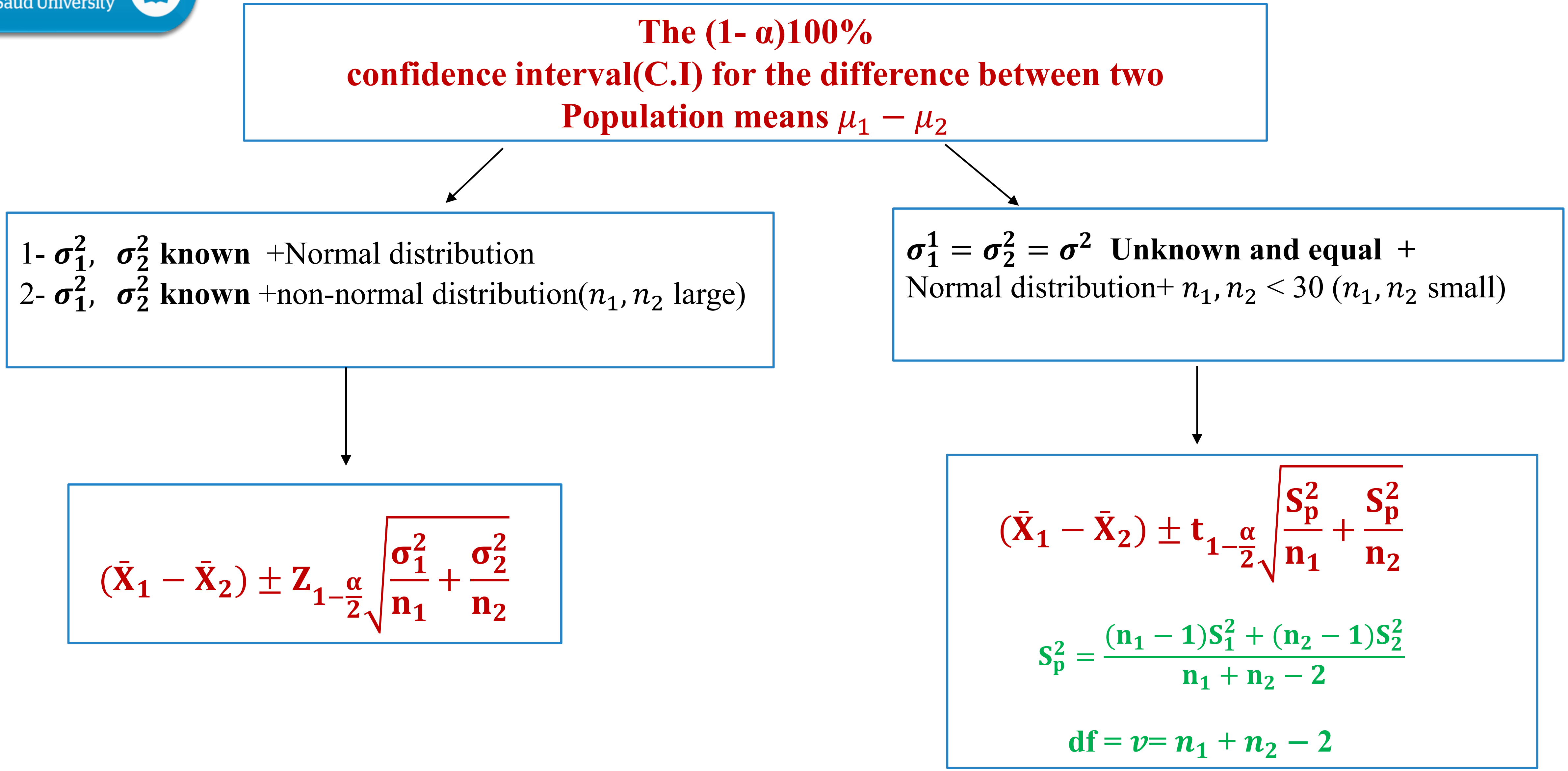
Note:

- If $0 \in (L, U) \Longrightarrow (\mu_A - \mu_B = 0 \Leftrightarrow \mu_A = \mu_B)$.

We conclude that the two population means may be equal .

- If $0 \notin (L, U) \Longrightarrow (\mu_A - \mu_B \neq 0 \Leftrightarrow \mu_A \neq \mu_B)$.

We conclude that the two population means not be equal .



The (1- α)100% confidence interval(C.I) for the difference between two Population means $\mu_1 - \mu_2$

1- $\boldsymbol{\sigma_1^2}$, $\boldsymbol{\sigma_2^2}$ **known** +Normal distribution

2- $\boldsymbol{\sigma_1^2}$, $\boldsymbol{\sigma_2^2}$ **known** +non-normal distribution(n_1, n_2 large)

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\boldsymbol{\sigma_1^1 = \sigma_2^2 = \sigma^2}$ **Unknown and equal** + Normal distribution+ $n_1, n_2 < 30$ (n_1, n_2 small)

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$\mathbf{df} = v = n_1 + n_2 - 2$$

Example: (1st Case: σ_1^2 and σ_2^2 are known)

An experiment was conducted to compare time length (duration time) of two types of surgeries (A) and (B). 75 surgeries of type (A) and 50 surgeries of type (B) were performed. The average time length for (A) was 42 minutes and the average for (B) was 36 minutes.

(1) Find a point estimate for $\boldsymbol{\mu_A - \mu_B}$, where μ_A and μ_B are population means of the time length of surgeries of type (A) and (B), respectively.

(2) Find a 96% confidence interval for $\boldsymbol{\mu_A - \mu_B}$. Assume that the population standard deviations are 8 and 6 for type (A) and (B), respectively.

Surgery	Type (A)	Type (B)
Sample size	$n_A = 75$	$n_B = 75$
Sample mean	$\bar{X}_A = 42$	$\bar{X}_B = 42$
Population Standard Deviation	$\sigma_A = 8$	$\sigma_B = 8$

Solution:

(1) A point estimate for $\mu_A - \mu_B$ is:

$$\bar{X}_A - \bar{X}_B = 42 - 36 = 6.$$

(2) Finding a 96% confidence interval for $\mu_A - \mu_B$:

$\alpha = ??$

$96\% = (1-\alpha)\ 100\%$

$0.96 = (1-\alpha) \Leftrightarrow \alpha = 0.04 \Leftrightarrow \alpha/2 = 0.02 \Leftrightarrow 1 - \alpha/2 = 0.98$

The reliability coefficient is: : $Z_{1-\frac{\alpha}{2}} = Z_{0.98} = 2.055$

A 96% C.I. for $\mu_A - \mu_B$ is:

$$(\bar{X}_B - \bar{X}_A) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}}$$

$$6 \pm Z_{0.98} \sqrt{\frac{8^2}{75} + \frac{6^2}{50}}$$

$$6 \pm (2.055) \sqrt{\frac{64}{75} + \frac{36}{50}}$$

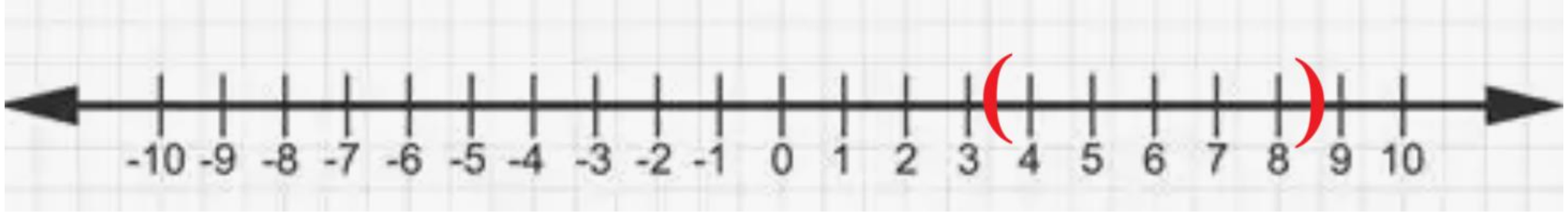
$$6 \pm 2.578$$

$$3.422 < \mu_A - \mu_B < 8.58$$

We are 96% confident that $\mu_A - \mu_B \in (3.42, 8.58)$.

Note:

Since the confidence interval does not include zero,



$0 \notin (3.42\ ,\ 8.58)$ ⟹ $(\mu_A - \mu_B \neq 0 \Leftrightarrow \mu_A \neq \mu_B)$.

we conclude that the two populations means are not equal .

Therefore, we may conclude that the **mean time length is not the same for the two types of surgeries.**

The length of the 96% confidence interval for $\mu_A - \mu_B$:

length= Upper limit – Lower limit = 8.58 -3.42 =5.16

Example: (2nd Case: $\sigma_1^2 = \sigma_2^2$ unknown)

To compare the time length (duration time) of two types of surgeries (A) and (B), an experiment shows the following results based on two independent samples:

Type A:	140, 138, 143, 142, 144, 137
Type B:	135, 140, 136, 142, 138, 140

(1) Find a point estimate for $\mu_A - \mu_B$, where μ_A (μ_B) is the mean time length of type A (B).

(2) Assuming normal populations with **equal variances**, find a 95% confidence interval for $\mu_A - \mu_B$.

Surgery	Type (A)	Type (B)
Sample size	$n_A = 6$	$n_B = 6$
Sample mean	$\bar{X}_A = 140.67$	$\bar{X}_B = 138.50$
Sample Variance	$S_A^2 = 7.87$	$S_B^2 = 7.10$

Solution:

(1) A point estimate for $\mu_A - \mu_B$ is:

$\bar{X}_A - \bar{X}_B$= 140.67 - 138.50 = 2.17

(2) Finding a 95% confidence interval for $\mu_A - \mu_B$:

95% = (1−α) 100%

$$0.95 = (1-\alpha) \Leftrightarrow \boldsymbol{\alpha = 0.05} \Leftrightarrow \frac{\alpha}{2} = 0.025 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.975$$

$$\mathrm{df} = \mathrm{v} = \mathrm{n_A} + \mathrm{n_B} - 2 = 10 .$$

The reliability coefficient is : $t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.228$

The pooled estimate of the common variance is:

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} = \frac{(6-1)(7.87)+(6-1)(7.1)}{6+6-2} = 7.485$$

A 95% C.I. for $\mu_A - \mu_B$ is:

$$(\bar{X}_A - \bar{X}_B) \pm t_{1-\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_A} + \frac{S_p^2}{n_B}}$$

$$2.17 \pm (2.228)\sqrt{\frac{7.485}{6} + \frac{7.485}{6}}$$

$$2.17 \pm 3.519$$

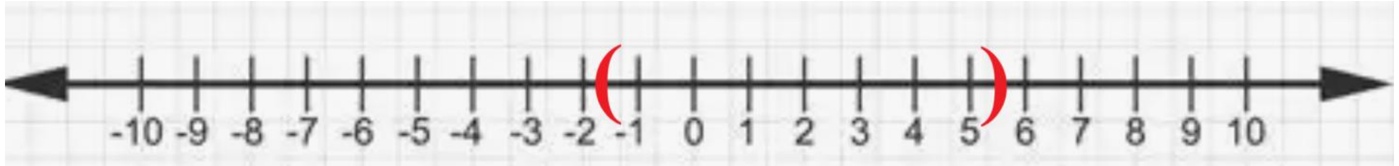
$$-1.35 < \mu_A - \mu_B < 5.69$$

We are 95% confident that $\mu_A - \mu_B \in (-1.35, 5.69)$.

Note:

Since the confidence interval include zero, we conclude that the two population means may be equal.

$$0 \in (-1.35, 5.69) \Longrightarrow (\mu_A - \mu_B = 0 \Leftrightarrow \mu_A = \mu_B).$$

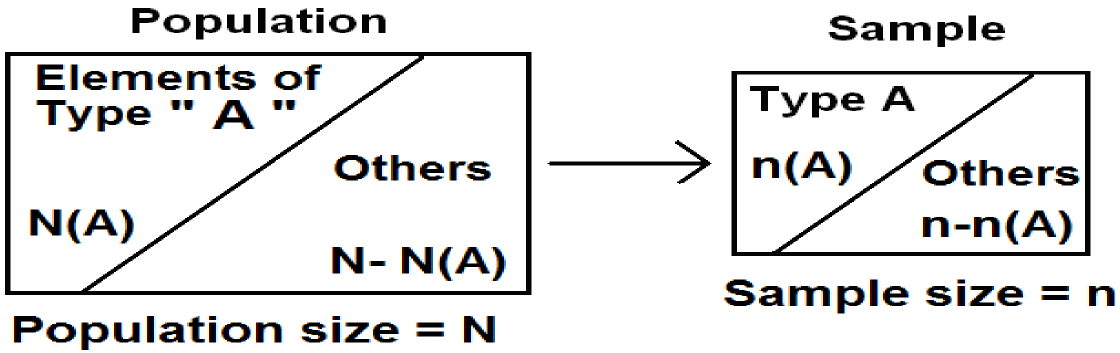


Therefore, **we may conclude that the mean time length is the same for the both types of surgeries.**

The length of the 95% confidence interval for $\mu_A - \mu_B$:

length= Upper limit – Lower limit = 5.69 – (-1.35) = 7.04

6.5 Confidence Interval for a Population Proportion (p):



Population

Elements of Type " A "

N(A)

Others

N- N(A)

Population size = N

Sample

Type A

n(A)

Others

n-n(A)

Sample size = n

Recall:

1-For the population:

N (A) = number of elements in the population with a specified characteristic “A”

N = total number of elements in the population (population size)

The population proportion is: $p = \dfrac{N(A)}{N}$ (p is a parameter)

2- For the sample:

n(A) = number of elements in the sample with the same characteristic “A”

n = sample size

The sample proportion is: $\hat{p} = \dfrac{n(A)}{n}$ ($\hat{p}$ is a statistic)

3. The sampling distribution of the sample proportion ($\hat{p}$) is used to make inferences about the population proportion (p).

4. The mean of ($\hat{p}$) is: $\mu_{\hat{p}} = p$

5. The variance of ($\hat{p}$) is: $\sigma^2_{\hat{p}} = p(1-p)/n$

6. The standard error (standard deviation) of ($\hat{p}$) is: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

7. For large sample size ($n \geq 30, np > 5, n(1-p) > 5$), the sample proportion ($\hat{p}$) has approximately a normal distribution with mean $\mu_{\hat{p}} = p$ and a variance $\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$ that is:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$ (approximately)

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$ (approximately)

(i) Point Estimate for (p):

A good point estimate for the population proportion (p) is the sample proportion ($\hat{p}$).

(ii) Interval Estimation (Confidence Interval) for (p):

For large sample size ($n \geq 30$, $np > 5$, $n(1 - p) > 5$),an approximate $(1-\alpha)$ 100% confidence interval for (p) is

$$\hat{p} \pm Z_{1-\alpha}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$\left(\hat{p} - Z_{1-\alpha}\sqrt{\frac{\hat{p}\hat{q}}{n}}\ ,\ \hat{p} + Z_{1-\alpha}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

(where $\hat{q}$=1- $\hat{p}$)

Estimator ± (Reliability Coefficient) × (Standard Error)

Example:

In a study on the obesity of Saudi women, a random sample of 950 Saudi women was taken. It was found that 611 of these women were obese (overweight by a certain percentage).

(1) Find a point estimate for the **true proportion** of Saudi women who are obese.

(2) Find a 95% confidence interval for the **true proportion** of Saudi women who are obese.

Solution:

Variable: whether or not a women is obese (qualitative variable)

Population: all Saudi women

Parameter: p = the proportion of women who are obese.

Sample:

n = 950 (950 women in the sample)

n(A) = 611 (611 women in the sample who are obese)

The sample proportion (the proportion of women who are obese in the sample.) is:

$$\hat{P} = \frac{n(A)}{n} = \frac{611}{950} = 0.643 \qquad (\hat{q} = 1 - \hat{p} = 1 - 0.643 = 0.357)$$

(1) A point estimate for p is: $\hat{p} = 0.643$

(2) We need to construct 95% C.I. for the proportion (p).

$$95\% = (1-\alpha)100\% \Leftrightarrow 0.95 = 1-\alpha \Leftrightarrow \alpha = 0.05 \Leftrightarrow \frac{\alpha}{2} = 0.025 \Leftrightarrow 1-\frac{\alpha}{2} = 0.975$$

The reliability coefficient: $Z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$

A 95% C.I. for the proportion (p) is:

$$\hat{p} \pm Z_{1-\alpha}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.643 \pm (1.96)\sqrt{\frac{(0.643)(1-0.643)}{950}}$$

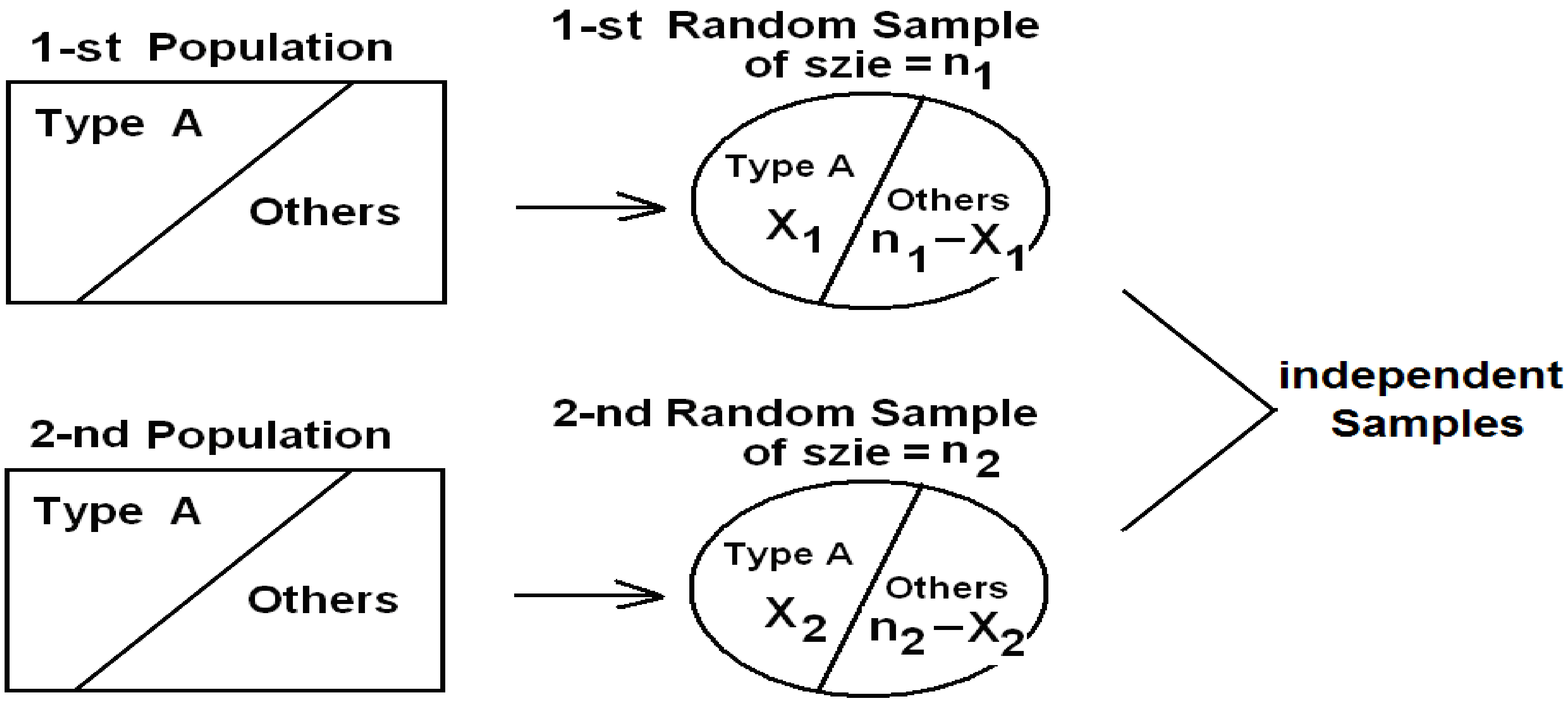
$$0.643 \pm (1.96)(0.01554)$$

$$0.643 \pm 0.0305 \quad \Rightarrow \quad (0.6127\ ,\ 0.6735)$$

We are 95% confident that the true value of the population proportion of obese women, p, lies in the interval $(0.61,\ 0.67)$, that is:

$$0.61 < p < 0.67$$

6.6 Confidence Interval for the Difference Between Two Population Proportions (p_1- p_2):



Suppose that we have two populations with:

- p_1 = population proportion of elements of type (A) in the 1-st population.
- p_2 = population proportion of elements of type (A) in the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about p_1- p_2.
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:

- Let X_1 = no. of elements of type (A) in the 1-st sample.
- Let X_2 = no. of elements of type (A) in the 2-nd sample.

- $\hat{p}_1 = \frac{X_1}{n_1}$ = the sample proportion of the 1-st sample.

- $\hat{p}_2 = \frac{X_2}{n_2}$ = the sample proportion of the 2-nd sample.

- The sampling distribution of $\hat{p}_1$ - $\hat{p}_2$ is used to make inferences about p_1- p_2 .

Recall:

1. Mean of $\hat{p}_1 - \hat{p}_2$ is: $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$

2. Variance of $\hat{p}_1 - \hat{p}_2$ is: $\sigma^2_{\hat{p}_1-\hat{p}_2} = \dfrac{p_1\,q_1}{n_1} + \dfrac{p_2\,q_2}{n_2}$

3. Standard error (standard deviation) of $\hat{p}_1 - \hat{p}_2$ is: $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{p_1\,q_1}{n_1} + \dfrac{p_2\,q_2}{n_2}}$

4. For large samples sizes ($n_1 \ge 30, n_2 \ge 30, n_1 p_1 > 5, n_1 q_1 > 5, n_2 p_2 > 5, n_2 q_2 > 5$), we have that $\hat{p}_1 - \hat{p}_2$ has approximately normal distribution with mean $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{p}_1-\hat{p}_2} = \dfrac{p_1\,q_1}{n_1} + \dfrac{p_2\,q_2}{n_2}$, that is :

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}\right) \qquad \text{(Approximately)}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1\,q_1}{n_1} + \dfrac{p_2\,q_2}{n_2}}} \sim \mathrm{N}(0,1) \qquad \text{(Approximately)}$$

Note: $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Point Estimation for $p_1 - p_2$:

Result:

A good point estimator for the difference between the two proportions, $p_1 - p_2$, is: $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$

Interval Estimation (Confidence Interval) for $p_1 - p_2$:

Result:

For large n_1 and n_2, an approximate (1- α)100% confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$\left((\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \quad , \quad (\hat{p}_1 - \hat{p}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right)$$

Estimator ± (Reliability Coefficient) × (Standard Error)

Example

A researcher was interested in comparing the proportion of people having cancer disease in two cities (A) and (B). A random sample of 1500 people was taken from the first city (A), and another independent random sample of 2000 people was taken from the second city (B). It was found that 75 people in the first sample and 80 people in the second sample have cancer disease.

(1) Find a point estimate for the **difference between the proportions** of people having cancer disease in the two cities.

(2) Find a 90% confidence interval for the **difference between the two proportions**.

Solution:

p_1 = population proportion of people having cancer disease in the first city (A)

p_2 = population proportion of people having cancer disease in the second city (B)

$\hat{p}_1$ = sample proportion of the first sample

$\hat{p}_2$ = sample proportion of the second sample

X_1 = number of people with cancer in the first sample

X_2 = number of people with cancer in the second sample

For the first sample we have:

$$n_1 = 1500\,,\ \ X_1 = 75$$

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{75}{1500} = 0.05 \qquad \hat{q}_1 = 1 - \hat{p}_1 = 1 - 0.05 = 0.95$$

For the second sample we have:

$$n_2 = 2000\ ,\ X_2 = 80$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{80}{2000} = 0.04 \qquad \hat{q}_2 = 1 - \hat{p}_2 = 1 - 0.04 = 0.96$$

(1) Point Estimation for $p_1 - p_2$:

A good point estimate for the difference between the two proportions $p_1 - p_2$, is:

$$\begin{aligned}\hat{p}_1 - \hat{p}_2 &= 0.05 - 0.04 \\ &= 0.01\end{aligned}$$

2)Finding 90% Confidence Interval for $p_1 - p_2$:

$$90\% = (1-\alpha)100\% \Leftrightarrow 0.90 = (1-\alpha) \Leftrightarrow \alpha = 0.1 \Leftrightarrow \frac{\alpha}{2} = 0.05$$

The reliability coefficient:

$$Z_{1-\frac{\alpha}{2}} = z_{0.95} = 1.645$$

A 90% confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$0.01 \pm 1.645\sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}}$$

$$0.01 \pm 0.01173$$

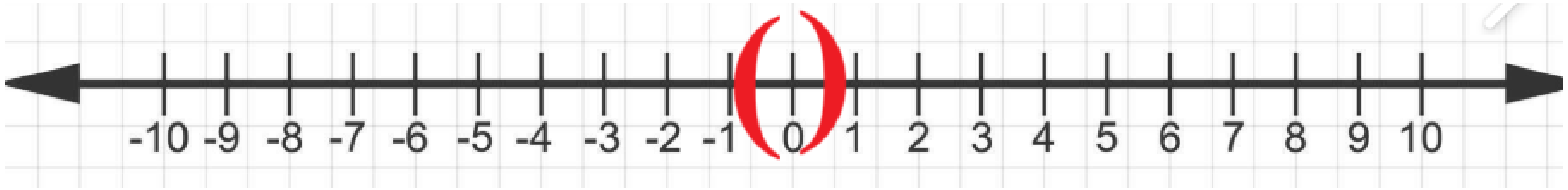
$$-0.0017 < \boldsymbol{p_1 - p_2} < 0.0217$$

We are 90% confident that $p_1 - p_2 \in (-0.0017, 0.0217)$.

Note:

Since the confidence interval includes zero, we may conclude that the two population proportions are equal

Since $0 \in (-0.0017, 0.0217)$ ⟶ $(p_1 - p_2 = 0 \Leftrightarrow p_1 = p_2)$.



Therefore, **we may conclude that the proportion of people having cancer is the same in both cities.**

The length of the 90% confidence interval for $p_1 - p_2$:

length= Upper limit – Lower limit = 0.0217 - (- 0.0017) =0.0234