

Chapter 6:

Using Sample Data to Make Estimations About Population Parameters

August 2024

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)

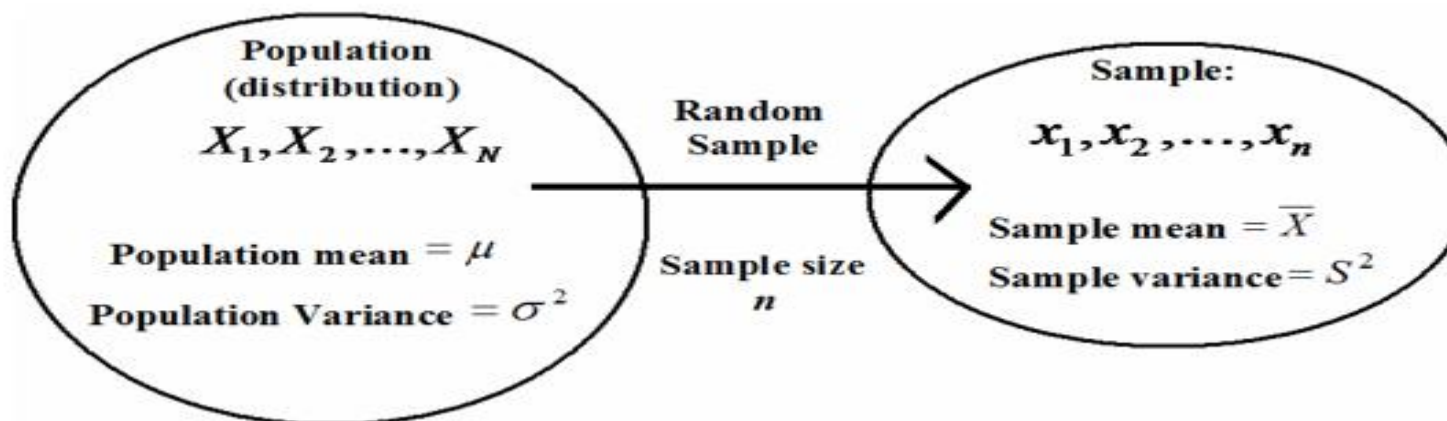
Answering research questions about the unknown parameters of the population (confirming or denying some conjectures or statements about the unknown parameters).

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, \dots, X_N

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

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

Sample:

Sample Size = n

Sample values: x_1, x_2, \dots, x_n

Sample Mean: $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$

Sample Variance: $S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

(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:

$$x_1, x_2, \dots, x_n$$

- Compute the sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

Result:

The sample mean
mean (μ).

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

is a "good" point estimator of the population

(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- α) % confident level

- How to get α when confidence level (1- α) % known

Example1 :

If we are 95% confident ,find α ?

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

Example2 :

If we are 99% confident ,find α ?

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

Example3 :

If we are 80% confident ,find α ?

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

Example4 :

If we are 92% confident ,find α ?

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

Result: (For the case when σ is known)

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

A $(1 - \alpha)$ 100% confidence interval for μ is:

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \sigma_{\bar{X}}$$

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

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

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

(b) If x_1, x_2, \dots, 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 **$(1 - \alpha)$ 100% confidence interval for μ** is:

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \sigma_{\bar{X}}$$

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

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

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

Note that :

1. We are $(1 - \alpha)$ 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 $= \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
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. $Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ = margin of error = precision of the estimate.
6. In general the interval estimate (confidence interval) may be expressed as follows:

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \sigma_{\bar{X}}$$

estimator \pm (reliability coefficient) \times (standard Error)

estimator \pm margin of error

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)

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

A $(1-\alpha)$ 100% confidence interval for μ is:

$$\begin{aligned} & \bar{X} \pm t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\bar{X}} \\ & \bar{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \\ & \left(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right) \end{aligned}$$

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

Note that:

1. We are $(1-\alpha)$ 100% confidence that the true value of μ belongs to the interval

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

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:

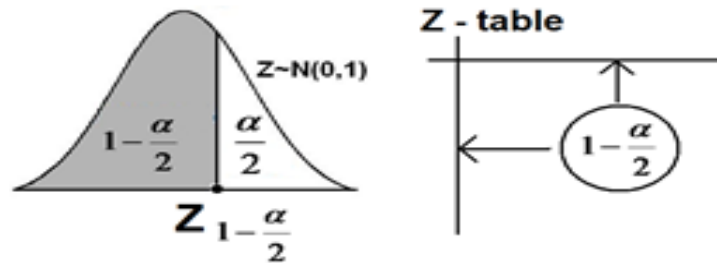
Estimator \pm (Reliability Coefficient) \times (Estimate of the Standard Error)

$$\bar{X} \pm t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\bar{X}}$$

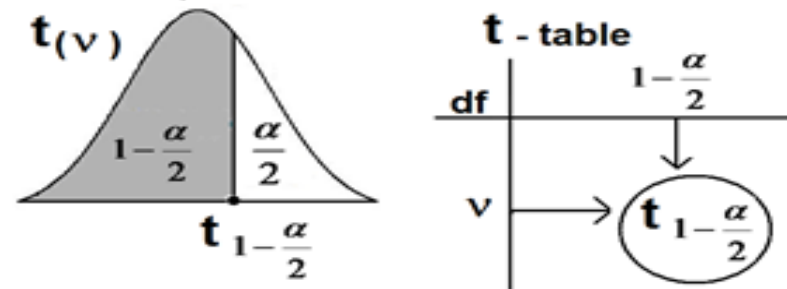
...

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: ($\text{df} = v = n-1$)



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 $\alpha = 0.1$:

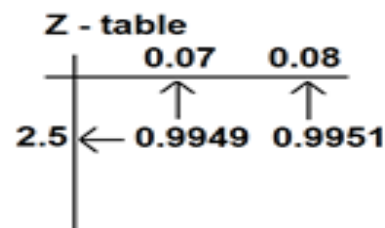
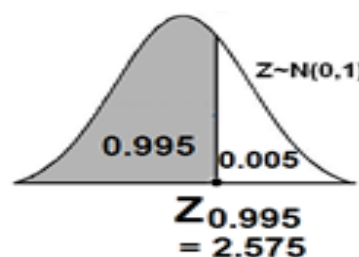
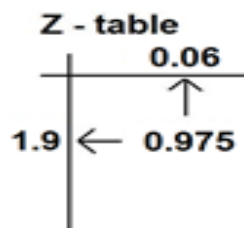
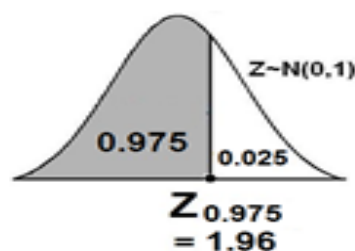
$$1 - \frac{\alpha}{2} = 1 - \frac{0.1}{2} = 0.95 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$$

(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 $\alpha = 0.01$:

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



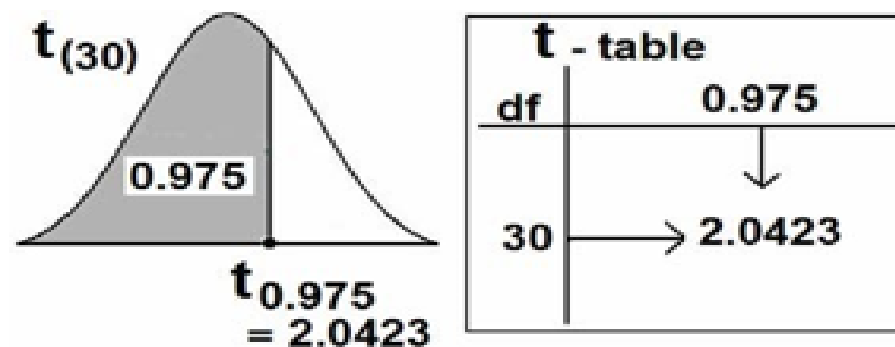
Example:

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

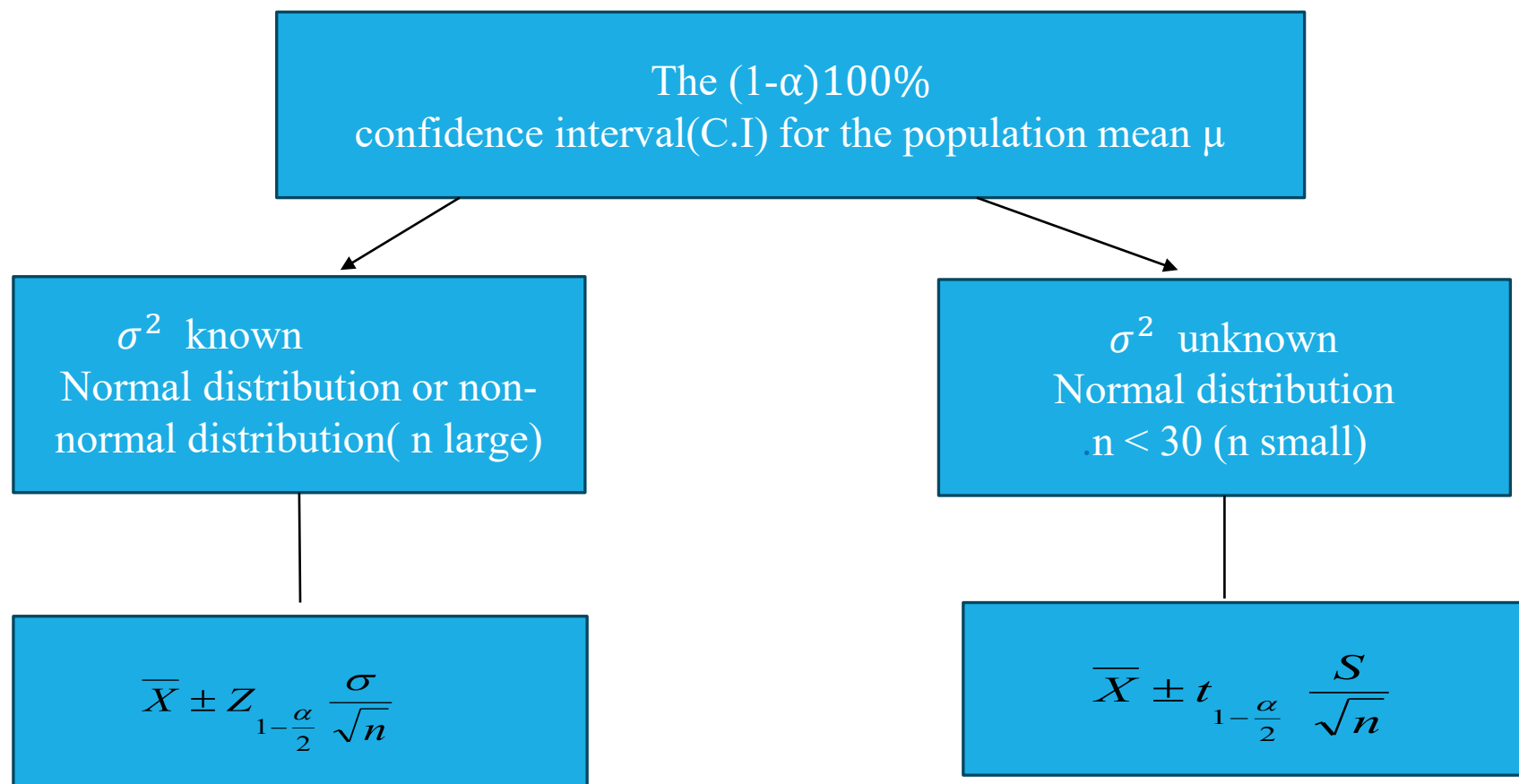
Solution:

$$df = v = 30$$

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



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



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

Example: (The case where σ^2 is 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 ($\sigma^2=10.89$)

$X \sim \text{Normal}(\mu, 10.89)$

$\mu = ??$ (unknown- we need to estimate μ)

Sample size: $n = 123$ (large)

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

...

(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 .

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

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

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:

$$25.71 < \mu < 26.69$$

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 σ^2 is 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.

$X \sim \text{Normal}(\mu, \sigma^2)$

$\mu = ??$ (unknown- we need to estimate μ)

$\sigma^2 = ??$ (unknown)

Sample size: $n = 21$

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

Sample standard deviation: $S = 10$

Degrees of freedom:

$$df = v = n-1 = 21 - 1 = 20$$

...

(a) Point Estimation:

We need to find a point estimate for μ .

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

$\mu \approx 37$ years

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

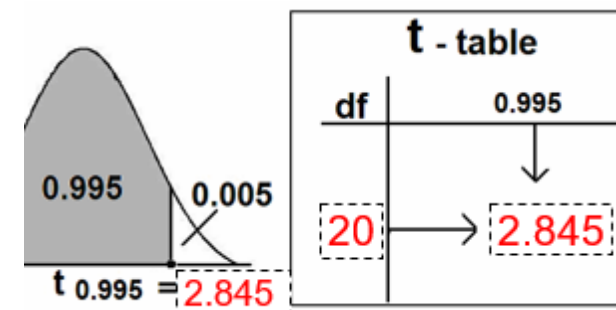
We need to find 99% C. I. for μ .

$$99\% = (1 - \alpha) 100\%$$

$$\alpha = 1/100 \Leftrightarrow \alpha = 0.01 \Leftrightarrow \alpha/2 = 0.01/2 = 0.005 \Leftrightarrow 1 - \alpha/2 = 0.995$$

$$v = df = n - 1 = 21 - 1 = 20$$

The reliability coefficient is: $t_{1-\frac{\alpha}{2}} = t_{0.995} = 2.845$



99% confidence interval for μ is:

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

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

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

$$(30.792, 43.208)$$

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

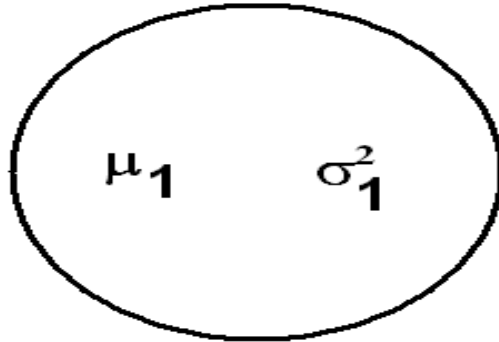
$$30.792 < \mu < 43.208$$

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

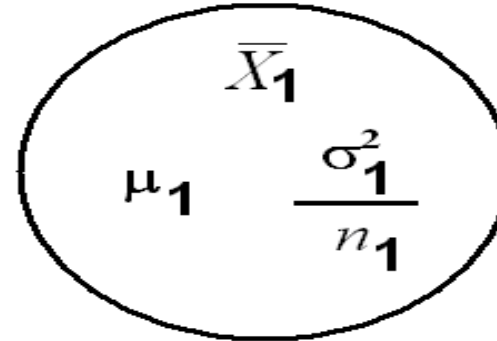
Suppose that we have two populations:

- 1-st population with mean μ_1 and variance σ_1^2
- 2-nd population with mean μ_2 and variance σ_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 \bar{X}_1 and S_1^2 be the sample mean and the sample variance of the 1-st sample.
- Let \bar{X}_2 and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.

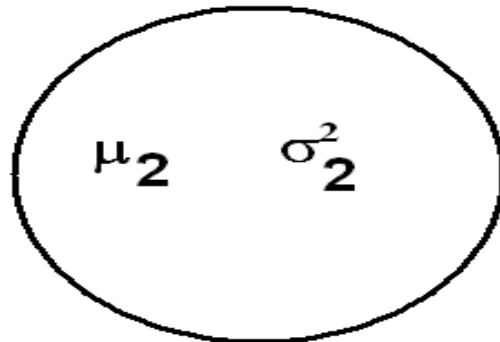
1-st Population



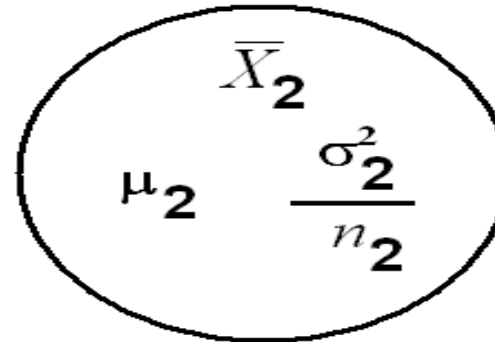
1-st Sample



2-nd Population



2-nd Sample



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:

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

3. Standard error of $\bar{X}_1 - \bar{X}_2$ is:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

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:

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

Point Estimation of $\mu_1 - \mu_2$:

Result:

$X_1 - X_2$ 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 $(1-\alpha)$ 100% confidence interval for $\mu_1 - \mu_2$ is:

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

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

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

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

Estimator \pm (Reliability Coefficient) \times (Standard Error)

(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

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

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

$$df = v = n_1 + n_2 - 2 .$$

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.

Result:

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

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

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

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 Confidence Interval (C.I) for the Difference between two Population Means $\mu_1 - \mu_2$:

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

σ_1^2, σ_2^2 known
Normal distribution or
non-normal distribution(n_1, n_2 large)

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

Normal distribution
 $\sigma_1^2 = \sigma_2^2 = \sigma^2$
unknown but equal
 $n_1 < 30, n_2 < 30$ (n_1, n_2 small)

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

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

$$df = v = n_1 + n_2 - 2$$

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 $\mu_A - \mu_B$, 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 $\mu_A - \mu_B$. Assume that the population standard deviations are 8 and 6 for type (A) and (B), respectively.

...

Solution:

Surgery	Type (A)	Type (B)
Sample Size	$n_A = 75$	$n_B = 50$
Sample Mean	$\bar{X}_A = 42$	$\bar{X}_B = 36$
Population Standard Deviation	$\sigma_A = 8$	$\sigma_B = 6$

(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 $\mu_A - \mu_B$:

$$\alpha = ??$$

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

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

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

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

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

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

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

$$6 \pm 2.578$$

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

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

...

Note:

Since the confidence interval does not include zero,

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

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.

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 $\mu_A - \mu_B$.

...

Solution:

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	$S^2_A = 7.87$	$S^2_B = 7.10$

(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 $\mu_A - \mu_B$:

$$95\% = (1-\alpha) 100\%$$

$$0.95 = (1-\alpha) \Leftrightarrow \alpha = 0.05 \Leftrightarrow \alpha/2 = 0.025 \Leftrightarrow 1 - \alpha/2 = 0.975$$

$$df = v = n_A + n_B - 2 = 10 .$$

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

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:

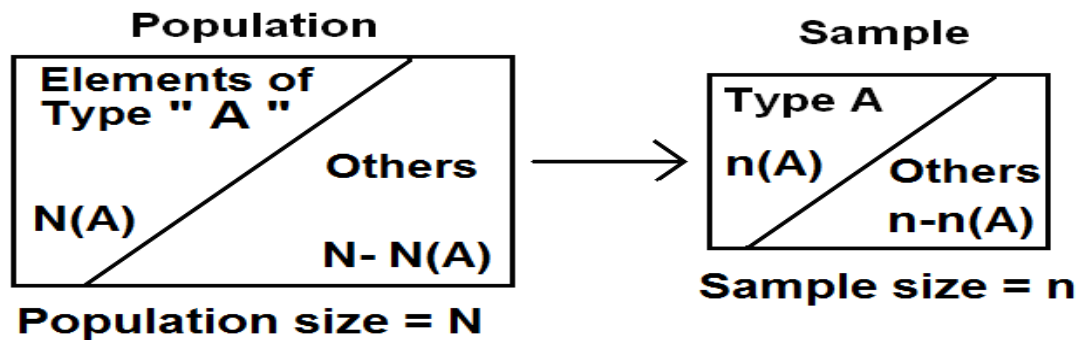
Since the confidence interval include zero, we conclude that the two population means may be equal ($\mu_A - \mu_B = 0 \Leftrightarrow \mu_A = \mu_B$).

Since , $0 \in (-1.35, 5.69)$

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

...

6.5 Confidence Interval for a Population Proportion (p):



Recall:

1. For the population:

$N(A)$ = number of elements in the population with aspecified characteristic "A"

N = total number of elements in the population(population size) The population proportion is:

$$p = \frac{N(A)}{N} \quad (p \text{ is a parameter})$$

2. For the sample:

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

n = sample size

The sample proportion is:

$$\hat{p} = \frac{n(A)}{n} \quad (\hat{p} \text{ is a statistic})$$

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: $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$

6. The standard error (standard deviation) of (\hat{p}) is:

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 $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$ that is:

$$\sigma_{\hat{p}}^2 = p(1-p)/n$$

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

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

(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

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

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

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:

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

(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:

$$Z_{1-\frac{\alpha}{2}} = Z_{0.975} = 1.96$$

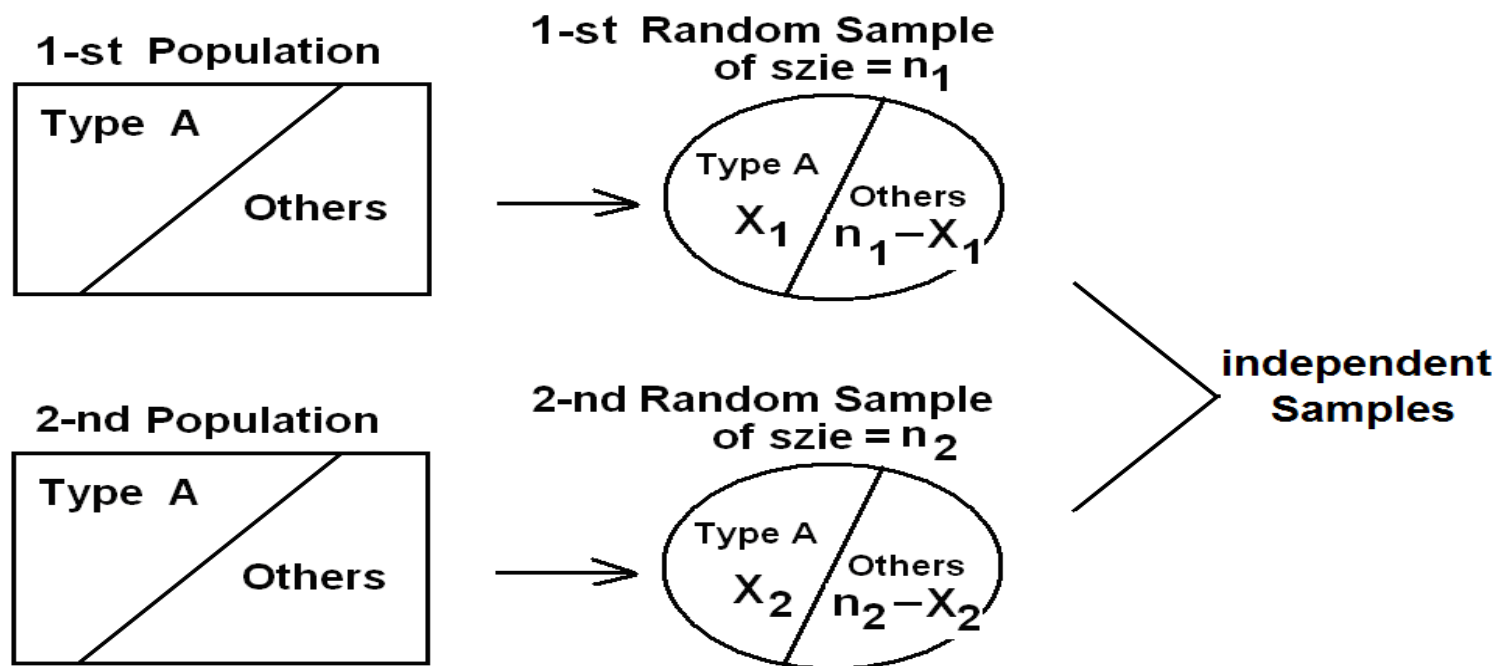
A 95% C.I. for the proportion (p) is:

$$\begin{aligned} & \hat{p} \pm Z_{1-\alpha} \sqrt{\frac{\hat{p}\hat{q}}{n}} \\ & 0.643 \pm (1.96) \sqrt{\frac{(0.643)(1-0.643)}{950}} \\ & 0.643 \pm (1.96)(0.01554) \\ & 0.643 \pm 0.0305 \quad \longrightarrow \quad (0.6127, 0.6735) \end{aligned}$$

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: $\sigma_{\hat{p}_1 - \hat{p}_2}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

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

4. For large samples sizes ($n_1 \geq 30, n_2 \geq 30, n_1 p_1 > 5, n_1 q_1 > 5, n_2 p_2 > 5, n_2 q_2 > 5$), 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 $\sigma_{\hat{p}_1 - \hat{p}_2}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$, that is :

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

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

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 $(1 - \alpha)100\%$ confidence interval for $p_1 - p_2$ is:

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

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

Estimator \pm (Reliability Coefficient) \times (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:

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

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

For the second sample we have:

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

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

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

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

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04$$

$$= 0.01$$

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

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

The reliability coefficient:

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

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

,

$$\begin{aligned}
 & (\hat{p}_1 - \hat{p}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \\
 & (\hat{p}_1 - \hat{p}_2) \pm Z_{0.95} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \\
 & 0.01 \pm 1.645 \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}} \\
 & 0.01 \pm 0.01173 \\
 & -0.0017 < p_1 - p_2 < 0.0217
 \end{aligned}$$

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

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

College of Science
Department of Statistics & OR

STAT 109
Biostatistics

Chapter 7

**Using Sample Statistics To
Test Hypotheses About Population Parameters**

August 2024

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

7.1 Introduction

7.2 Hypothesis Testing: A Single Population Mean (μ)

7.3 Hypothesis Testing: The Difference Between Two Population Mean independent ($\mu_1 - \mu_2$)

7.4 Paired Comparisons

7.5 Hypothesis Testing: A Single Population Proportion (P)

7.6 Hypothesis Testing: The Difference Between Two Population Proportions ($P_1 - P_2$)

7.1 Introduction

Consider a population with some unknown parameter θ . We are interested in testing (confirming or denying) some conjectures about θ . For example, we might be interested in testing the conjecture that $\theta > \theta_0$, where θ_0 is a given value.

- A hypothesis is a statement about one or more populations.
- A research hypothesis is the conjecture or supposition that motivates the research.
- A statistical hypothesis is a conjecture (or a statement) concerning the population which can be evaluated by appropriate statistical technique.
- For example, if θ is an unknown parameter of the population, we might be interested in testing the conjecture stating that $\theta \geq \theta_0$ against $\theta < \theta_0$ (for some specific value θ_0).

The hypothesis

- We usually test the null hypothesis (H_0) against the alternative (or the research) hypothesis (H_1 or H_A) by choosing one of the following situations:

(i)	$H_0: \theta = \theta_0$	against	$H_A: \theta \neq \theta_0$
(ii)	$H_0: \theta \geq \theta_0$	against	$H_A: \theta < \theta_0$
(iii)	$H_0: \theta \leq \theta_0$	against	$H_A: \theta > \theta_0$

- Equality sign must appear in the null hypothesis.
- H_0 is the null hypothesis and H_A is the alternative hypothesis.
(H_0 and H_A are **complement** of each other)
- The null hypothesis (H_0) is also called "the hypothesis of no difference".
- The alternative hypothesis (H_A) is also called the **research hypothesis**.

There are 4 possible situations in testing a statistical hypothesis:

		Condition of Null Hypothesis H_0 (Nature/reality)	
		Ho is true	Ho is false
Possible Action (Decision)	Accepting H_0	Correct Decision	Type II error (β)
	Rejecting H_0	Type I error (α)	Correct Decision

- There are two types of Errors:
 - Type I error = Rejecting H_0 when H_0 is true
 - $P(\text{Type I error}) = P(\text{Rejecting } H_0 \mid H_0 \text{ is true}) = \alpha$
 - Type II error = Accepting H_0 when H_0 is false
 - $P(\text{Type II error}) = P(\text{Accepting } H_0 \mid H_0 \text{ is false}) = \beta$

- The level of significance of the test is the probability of rejecting true H_0 :

$$\alpha = P(\text{Rejecting } H_0 \mid H_0 \text{ is true}) = P(\text{Type I error})$$

- There are 2 types of alternative hypothesis:

- One-sided alternative hypothesis:

- $H_0: \theta \geq \theta_0$ against $H_A: \theta < \theta_0$

- $H_0: \theta \leq \theta_0$ against $H_A: \theta > \theta_0$

- Two-sided alternative hypothesis:

- $H_0: \theta = \theta_0$ against $H_A: \theta \neq \theta_0$

- We will use the terms "accepting" and "not rejecting" interchangeably. Also, we will use the terms "acceptance" and "nonrejection" interchangeably.

We will use the terms "accept" and "fail to reject" interchangeably

The Procedure of Testing H_0 (against H_A):

The test procedure for rejecting H_0 (accepting H_A) or accepting H_0 (rejecting H_A) involves the following steps:

1. Determine the hypothesis : Null hypothesis (H_0) and Alternative hypothesis (H_A) .
2. Determining a test statistic (T.S.)

We choose the appropriate test statistic based on the point estimator of the parameter.

The test statistic has the following form:

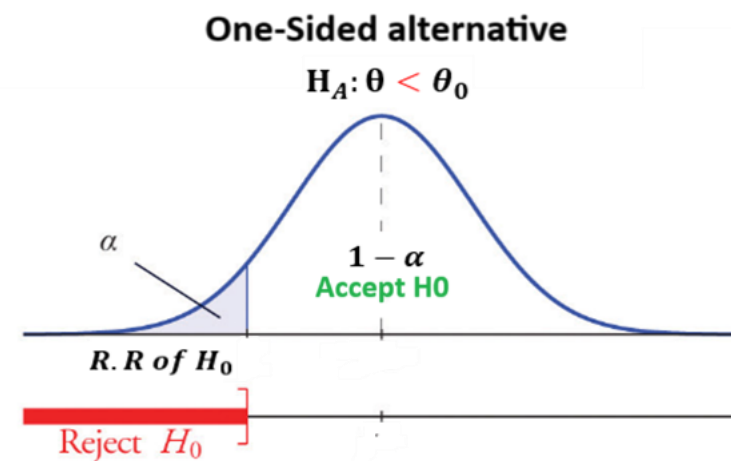
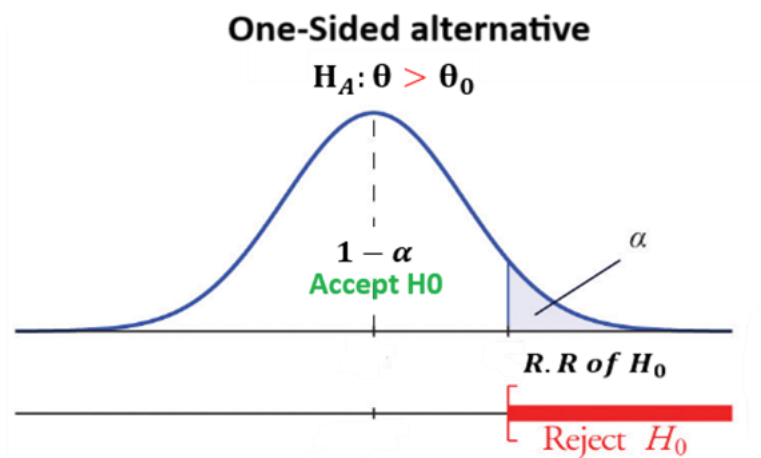
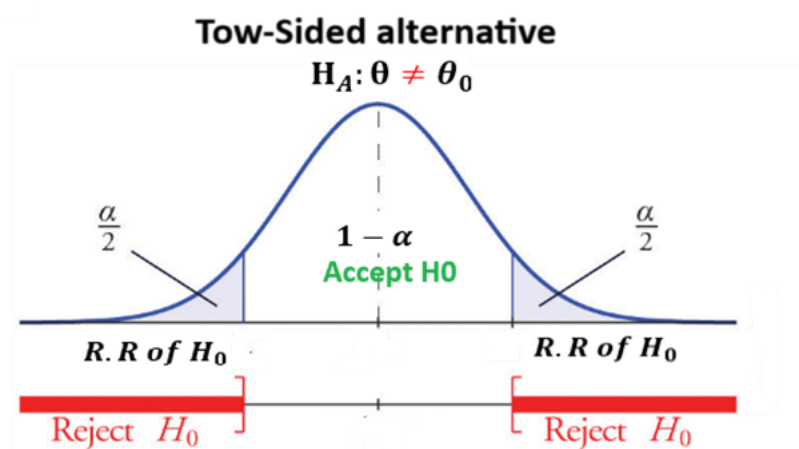
$$\text{Test statistic} = \frac{\text{Estimate} - \text{hypothesized parameter}}{\text{Standard error of the estimate}}$$

3. Determining the level of significance (α):
 $\alpha = 0.01, 0.025, 0.05, 0.10$
4. Determining the **rejection region of H_0** (R.R.) and the **acceptance region of H_0** (A.R.).

The R.R. of H_0 depends on H_A and α

- H_A determines the direction of the R.R. of H_0 .
- α determines the size of the R.R. of H_0 . (α = the size of the R.R. of H_0 = shaded area)

The rejection region(R.R) depends on the sign of H_A and α



5. Decision:

We reject H_0 (and accept H_A) if the value of the test statistic (T.S.) belongs to the R.R. of H_0 , and vice versa.

Notes:

1. The rejection region of H_0 (R.R.) is sometimes called "the critical region".
2. The values which separate the rejection region (R.R.) and the acceptance region (A.R.) are called the critical values or "Reliability coefficient".

7.2 Hypothesis Testing: A Single Population Mean (μ)

Suppose that X_1, X_2, \dots, X_n is a random sample of size n from a distribution (or population) with mean μ and variance σ^2 .

We need to test some hypotheses (make some statistical inference) about the mean (μ).

The Procedure for hypotheses testing about the mean (μ): Let μ_0 be a given known value

(1) First case:

Assumptions:

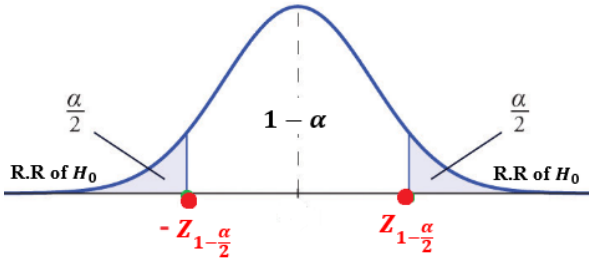
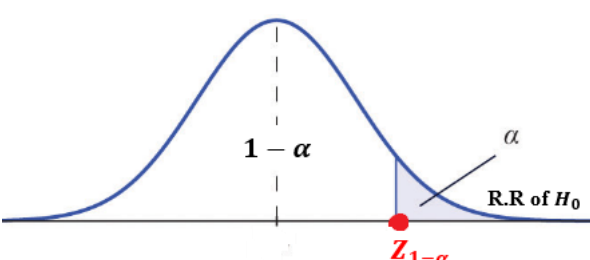
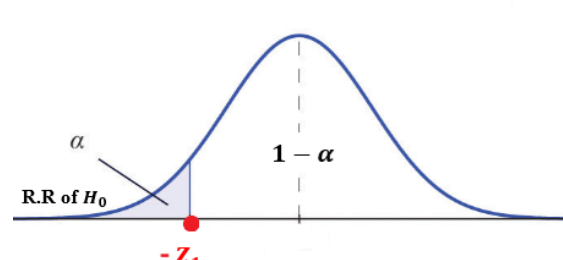
- The variance σ^2 is known .
- Normal distribution with any sample size or
Non-normal distribution with large sample size

(2) Second case:

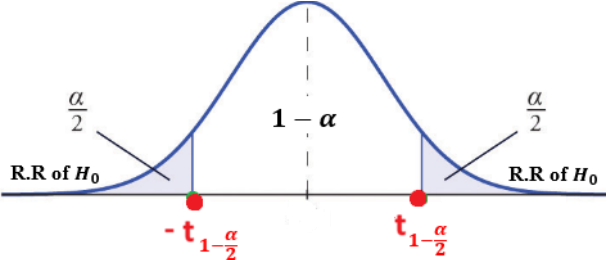
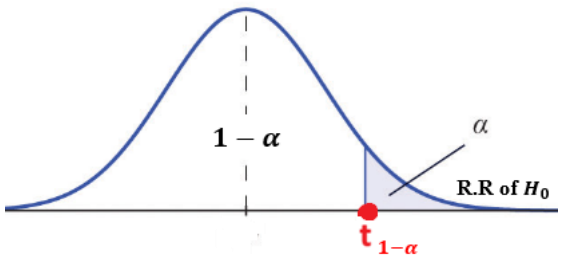
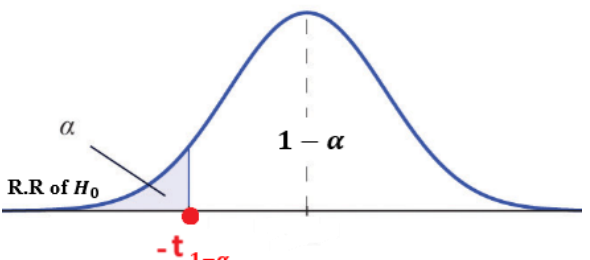
Assumptions:

- The variance σ^2 is unknown .
- Normal distribution .
- Any sample size .

Test procedures:

Hypotheses	$H_0 : \mu = \mu_0$ $H_A : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_A : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_A : \mu < \mu_0$
First case	σ is known , Normal distribution or Non-normal distribution (n large = $n \geq 30$)		
Test Statistic (T.S.)	$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$		
R.R. & A.R. of H_0			
Critical value(s)	$Z_{1-\frac{\alpha}{2}}$ and $Z_{1-\frac{\alpha}{2}}$	$Z_{1-\alpha}$	$-Z_{1-\alpha}$
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$Z_{T.S} > Z_{1-\frac{\alpha}{2}}$ or $Z_{T.S} < -Z_{1-\frac{\alpha}{2}}$	$Z_{T.S} > Z_{1-\alpha}$	$Z_{T.S} < -Z_{1-\alpha}$
	T.S. \in R.R. Two - Sided Test	T.S. \in R.R. One - Sided Test	T.S. \in R.R. One - Sided Test

Test procedures:

Hypotheses	$H_0 : \mu = \mu_0$ $H_A : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_A : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_A : \mu < \mu_0$
Second case	σ is unknown, Normal distribution and n small (n < 30)		
Test Statistic (T.S.)	$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1} \quad df = n - 1$		
R.R. & A.R. of H_0			
Critical value(s)	$t_{1-\frac{\alpha}{2}}$ and $-t_{1-\frac{\alpha}{2}}$	$t_{1-\alpha}$	$-t_{1-\alpha}$
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$T_{T.S} > t_{1-\frac{\alpha}{2}}$ or $T_{T.S} < -t_{1-\frac{\alpha}{2}}$	$T_{T.S} > t_{1-\alpha}$	$T_{T.S} < -t_{1-\alpha}$
	T.S. \in R.R. Two - Sided Test	T.S. \in R.R. One - Sided Test	T.S. \in R.R. One - Sided Test

Example : First case (Variance σ^2 known)

A random sample of 100 recorded deaths in the United States during the past year showed an average of 71.8 years. Assuming a population standard deviation of 8.9 year .

Does this seem to indicate that the mean life span today is greater than 70 years?

Use a 0.05 level of significance.

Solution:

$$n = 100 \text{ (large } n \geq 30 \text{)}$$

$$\sigma = 8.9 \text{ (} \sigma \text{ known)}$$

$$\bar{X} = 71.8$$

$$\alpha = 0.05 \text{ (level of significance)}$$

$$\mu = \text{average (mean)life span}$$

$$\mu_0 = 70$$

First case:

Assumptions:

- The variance σ^2 is known .
- Non-normal distribution with large sample size

Hypotheses:

$$H_0 : \mu \leq 70 \quad (\mu_0 = 70)$$
$$H_A : \mu > 70 \quad (\text{research hypothesis})$$

Test statistics (T.S.)

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{71.8 - 70}{8.9 / \sqrt{100}} = 2.02$$

Rejection Region of H_0 (R.R.): (critical region)

$$\alpha = 0.05 \gg \text{Critical Value : } Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.645$$

Decision :

Reject $H_0: \mu \leq 70$ if :

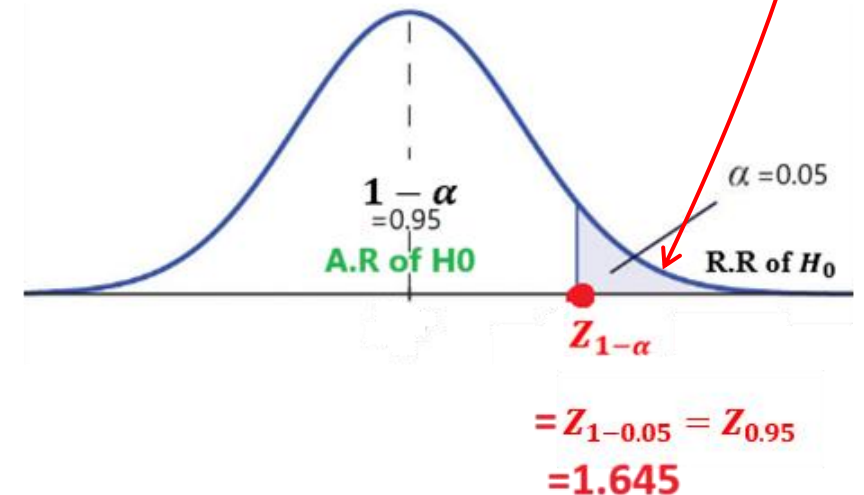
$$Z_{T.S} > Z_{1-\alpha}$$

$$2.02 > 1.645 \quad \text{condition satisfied}$$

Decision is Reject $H_0: \mu \leq 70$ and Accept $H_A: \mu > 70$

Therefore, we conclude that the mean life span today is greater than 70 years.

Another way to take decision is by graph :
Determine test statistics value on the graph



Third way we can use to take decision by Using P- Value as a decision tool:

P-value: is the smallest value of α for which we can reject the null hypothesis H_0 .

Calculate P-value :

- Calculating P-value depends on the alternative hypothesis H_A .
- Suppose that $Z_c = Z(T.S) = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ is the computed value of the test Statistic.
- The following table illustrates how to compute P-value, and how to use P-value for testing the null hypothesis:

Alternative Hypothesis (H_A)	$H_A : \mu \neq \mu_0$	$H_A : \mu > \mu_0$	$H_A : \mu < \mu_0$
P-Value =	$2 \times P(Z > Z_c)$	$P(Z > Z_c)$	$P(Z > -Z_c)$ $= P(Z < Z_c)$
Significance Level =	α		
Decision:	Reject H_0 if $P\text{-value} < \alpha$		

Example : For the previous example

Hypotheses:

$$H_0 : \mu \leq 70 \quad (\mu_0 = 70)$$

$$H_A : \mu \geq 70 \quad (\text{research hypothesis})$$

Test statistics (T.S.)

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{71.8 - 70}{8.9 / \sqrt{100}} = 2.02$$

To calculate P- value The alternative hypothesis was $H_A: \mu \geq 70$, $\alpha = 0.05$

$$\text{P-value} = P(Z > Z_c) = P(Z > 2.02) = 1 - P(Z < 2.02) = 1 - 0.9783 = 0.0217$$

Decision :

Reject H_0 if P-value $< \alpha$

$$0.0217 < 0.05 \quad (\text{Decision satisfied})$$

Then , we Reject H_0 .

Example: Second case (Variance σ^2 unknown)

The manager of a private clinic claims that the mean time of the patient-doctor visit in his clinic is 8 minutes. Test the hypothesis that $\mu=8$ minutes against the alternative that $\mu \neq 8$ minutes if a random sample of 25 patient-doctor visits yielded a mean time of 7.8 minutes with a standard deviation of 0.5 minutes. It is assumed that the distribution of the time of this type of visits is normal. Use a 0.01 level of significance.

Solution:

Normal distribution ,

$$\bar{X} = 7.8 ,$$

$\alpha=0.01$ (level of significance)

$S = 0.5$ Sample standard deviation (σ is unknown)

$$n = 25 \text{ (} n \text{ small } n < 30 \text{)}$$

μ = mean time of the visit

Second case:

Assumptions:

- The variance σ^2 is unknown .
- Normal distribution
- Any sample size .

Hypotheses:

$$H_0 : \mu = 8 \quad (\mu_0 = 8)$$
$$H_A : \mu \neq 8 \quad (\text{research hypothesis})$$

Test statistics (T.S.)

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{7.8 - 8}{0.5/\sqrt{21}} = -2$$

Rejection Region of H_0 (R.R.): (critical region)

$$\alpha = 0.01 \gg \text{Critical Value : } \pm t_{1-\frac{\alpha}{2}} = \pm t_{1-\frac{0.01}{2}} = \pm t_{0.995} = \pm 2.797$$

$$\text{df} = n-1 = 21-1 = 20$$

Decision :

$$\text{Reject } H_0: \mu = 8 \text{ if : } T_{T.S} < -T_{1-\frac{\alpha}{2}} \text{ or } T_{T.S} > T_{1-\frac{\alpha}{2}}$$

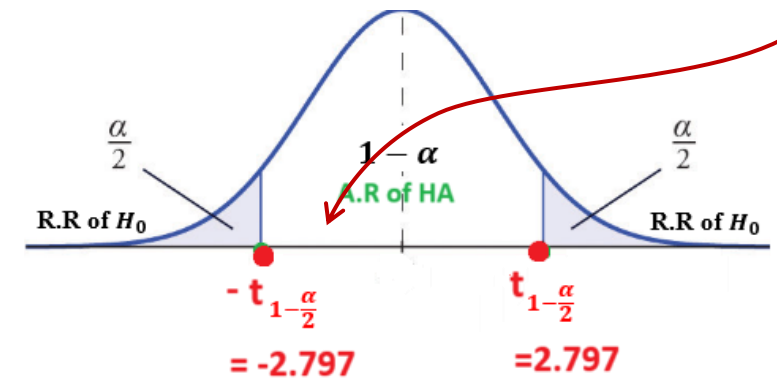
$$-2 \not< -2.797 \text{ (First condition not satisfied, then try the second condition)}$$

$$-2 \not> 2.797 \text{ (Second condition not satisfied)}$$

Decision is Accept $H_0: \mu = 8$ and Reject $H_0: \mu \neq 8$

Therefore, we conclude that the claim is correct.

Another way to take decision is by graph :
Determine test statistics value on the graph



Special case :

For the case of **non-normal population with unknown variance**, and when the sample size is large ($n \geq 30$), we may use the following test statistic:

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim N(0,1)$$

That is, we replace the population standard deviation (σ) by the sample standard deviation (S), and we conduct the test as described for the first case.

7.3 Hypothesis Testing: The Difference Between Two Population Means: (Independent Populations)

Suppose that we have two (independent) populations:

- 1-st population with mean μ_1 and variance σ_1^2 .
- 2-nd population with mean μ_2 and variance σ_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 \bar{X}_1 and S_1^2 be the sample mean and the sample variance of the 1-st sample.
- Let \bar{X}_2 and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.

We wish to test some hypotheses comparing the population means.

Hypotheses:

We choose one of the following situations:

- (i) $H_0: \mu_1 = \mu_2$ against $H_A: \mu_1 \neq \mu_2$
- (ii) $H_0: \mu_1 \geq \mu_2$ against $H_A: \mu_1 < \mu_2$
- (iii) $H_0: \mu_1 \leq \mu_2$ against $H_A: \mu_1 > \mu_2$

or equivalently,

- (i) $H_0: \mu_1 - \mu_2 = 0$ against $H_A: \mu_1 - \mu_2 \neq 0$
- (ii) $H_0: \mu_1 - \mu_2 \geq 0$ against $H_A: \mu_1 - \mu_2 < 0$
- (iii) $H_0: \mu_1 - \mu_2 \leq 0$ against $H_A: \mu_1 - \mu_2 > 0$

Test Statistic (T.S):

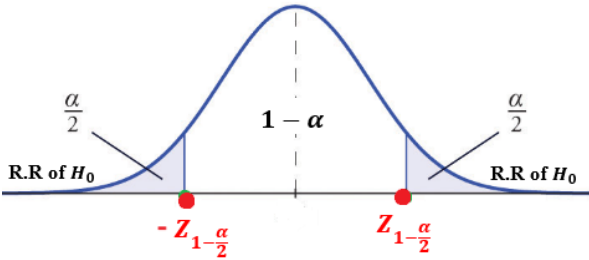
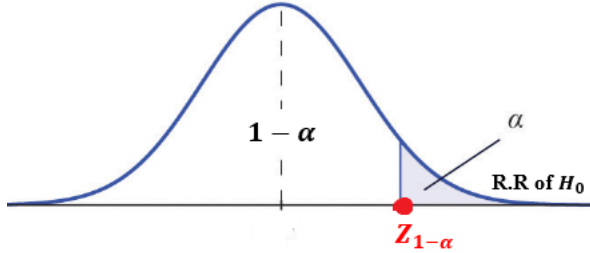
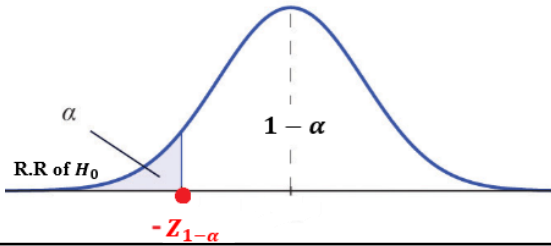
(1) First Case:

Assumptions:

- Normal populations
- non-normal populations with large sample sizes
- σ_1^2 and σ_2^2 are known, then the **test statistic** is:

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

Test procedures:

Hypotheses	$H_0 : \mu_1 - \mu_2 = \mu_0$ $H_A : \mu_1 - \mu_2 \neq \mu_0$	$H_0 : \mu_1 - \mu_2 \leq \mu_0$ $H_A : \mu_1 - \mu_2 > \mu_0$	$H_0 : \mu_1 - \mu_2 \geq \mu_0$ $H_A : \mu_1 - \mu_2 < \mu_0$
First case	if σ_1^2 and σ_2^2 are known , Normal populations or non-normal populations with large sample sizes		
Test Statistic (T.S.)	$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$		
R.R. & A.R. of H_0			
Critical value(s)	$Z_{1-\frac{\alpha}{2}}$ and $-Z_{1-\frac{\alpha}{2}}$	$Z_{1-\alpha}$	$-Z_{1-\alpha}$
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$Z_{T.S} > Z_{1-\frac{\alpha}{2}}$ or $Z_{T.S} < -Z_{1-\frac{\alpha}{2}}$	$Z_{T.S} > Z_{1-\alpha}$	$Z_{T.S} < -Z_{1-\alpha}$
	T.S. \in R.R. Two - Sided Test	T.S. \in R.R. One - Sided Test	T.S. \in R.R. One - Sided Test

(2) Second Case:

Assumptions:

- Normal populations .
- σ_1^2 and σ_2^2 are unknown but equal $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

then the **test statistic** is:

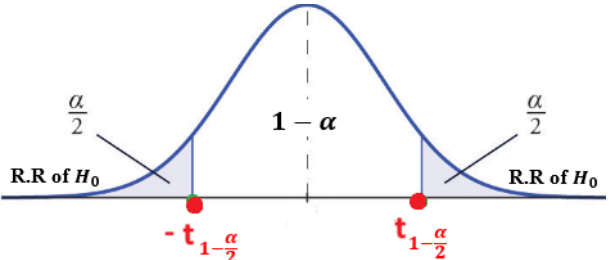
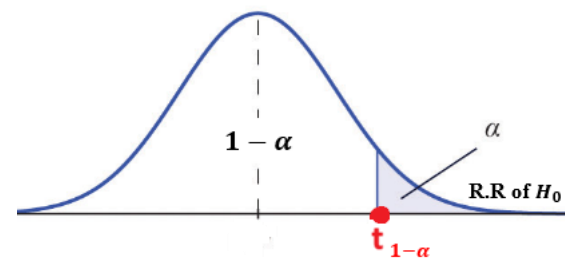
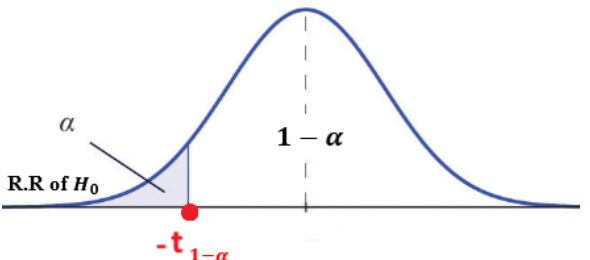
$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

where the pooled variance estimate of σ^2 is

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

and the degrees of freedom of is $df = v = n_1 + n_2 - 2$.

Test procedures:

Hypotheses	$H_0 : \mu_1 - \mu_2 = \mu_0$ $H_A : \mu_1 - \mu_2 \neq \mu_0$	$H_0 : \mu_1 - \mu_2 \leq \mu_0$ $H_A : \mu_1 - \mu_2 > \mu_0$	$H_0 : \mu_1 - \mu_2 \geq \mu_0$ $H_A : \mu_1 - \mu_2 < \mu_0$
Second case	if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown + Normal populations		
Test Statistic (T.S.)	$T = \frac{\bar{X} - \bar{X}_2}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}} \sim t_{n_1+n_2-2}, \quad S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}, \quad df = n_1 + n_2 - 2$		
R.R. & A.R. of H_0			
Critical value(s)	$t_{1-\frac{\alpha}{2}}$ and $-t_{1-\frac{\alpha}{2}}$	$t_{1-\alpha}$	$-t_{1-\alpha}$
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$T_{T.S} > t_{1-\frac{\alpha}{2}}$ or $T_{T.S} < -t_{1-\frac{\alpha}{2}}$	$T_{T.S} > t_{1-\alpha}$	$T_{T.S} < -t_{1-\alpha}$
	T.S. \in R.R. Two - Sided Test	T.S. \in R.R. One - Sided Test	T.S. \in R.R. One - Sided Test

Example: (σ_1^2 and σ_2^2 are known)

Researchers wish to know if the data they have collected provide sufficient evidence to indicate the difference in mean serum uric acid levels between individuals with Down's syndrome and normal individuals. The data consist of serum uric acid on 12 individuals with Down's syndrome and 15 normal individuals.

The sample means are $\bar{X}_1 = 4.5$ mg/100 ml and $\bar{X}_2 = 3.4$ mg/100 ml. Assume the populations are normal with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1.5$. Use significance level $\alpha = 0.05$.

Solution:

μ_1 = mean serum uric acid levels for the individuals with Down's syndrome.

μ_2 = mean serum uric acid levels for the normal individuals.

$n_1 = 12$, $\bar{X}_1 = 4.5$, $\sigma_1^2 = 1$ (σ_1^2 known)

$n_2 = 15$, $\bar{X}_2 = 3.4$, $\sigma_2^2 = 1.5$ (σ_2^2 known)

Normal populations

Hypotheses:

$$\begin{array}{l} H_0 : \mu_1 = \mu_2 \quad \text{OR} \quad (H_0 : \mu_1 - \mu_2 = 0) \\ H_A : \mu_1 \neq \mu_2 \quad (H_0 : \mu_1 - \mu_2 \neq 0) \end{array}$$

Test statistics (T.S.)

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{4.5 - 3.4}{\sqrt{\frac{1}{12} + \frac{1.5}{15}}} = 2.569$$

Rejection Region of H_0 (R.R.): (critical region : H_A Two sided)

$$\alpha = 0.05 \gg \text{Critical Value : } \pm Z_{1-\frac{\alpha}{2}} = \pm Z_{1-\frac{0.05}{2}} = \pm Z_{0.975} = \pm 1.96$$

Decision :

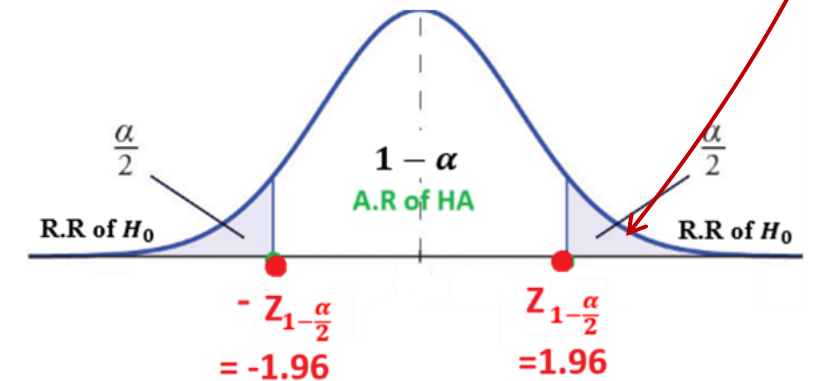
$$\text{Reject } H_0 : \mu_1 = \mu_2 \text{ if : } Z_{T.S} < -Z_{1-\frac{\alpha}{2}} \text{ or } Z_{T.S} > Z_{1-\frac{\alpha}{2}}$$

$$2.569 < -1.96 \text{ (First condition not satisfied, then try the second condition)}$$

$$2.569 > 1.96 \text{ (Second condition satisfied)}$$

Decision is Reject $H_0 : \mu_1 = \mu_2$ and Accept $H_A : \mu_1 \neq \mu_2$
we conclude that the two populations means are not equal.

Another way to take decision is by graph :
Determine test statistics value on the graph



2- Calculate the P-value

$$\text{P-value} = 2 \times P(Z > |Z_c|)$$

$$= 2 \times P(Z > |2.569|)$$

$$= 2 \times P(Z > 2.57)$$

$$= 2 \times [1 - P(Z < 2.57)]$$

$$= 2 \times [1 - 0.99492]$$

$$= 2 \times 0.00508$$

$$= 0.01016$$

$$(\text{where } Z_c = Z(\text{Test statistic}) = 2.569)$$

$$\text{P-value} = 0.01016$$

Decision : Reject H_0 if $\text{P-value} < \alpha$

$$\text{Since } 0.0106 < 0.05$$

We reject H_0 and accept H_A

Example: ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown)

An experiment was performed to compare the abrasive wear of two different materials used in making artificial teeth. 12 pieces of material(1) were tested by exposing each piece to a machine measuring wear. 10 pieces of material(2) were similarly tested. In each case, the depth of wear was observed. The samples of material(1) gave an average wear of 85 units with a sample standard deviation of 4, while the samples of materials(2) gave an average wear of 81 and a sample standard deviation of 5.

Can we conclude at the 0.05 level of significance that the mean abrasive wear of material(1) is greater than that of material(2)? Assume normal populations with equal variances.

Solution:

Normal populations

$\alpha=0.05$ (level of significance)

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown

Material(1)	Material(2)
$n_1 = 12$	$n_2 = 10$
$\bar{X}_1 = 85$	$\bar{X}_2 = 81$
$S_1 = 4$	$S_2 = 5$

Hypotheses:

$$\begin{array}{ll} H_0 : \mu_1 \leq \mu_2 & \text{OR} \quad (H_0: \mu_1 - \mu_2 \leq 0) \\ H_A : \mu_1 > \mu_2 & (H_0: \mu_1 - \mu_2 > 0) \end{array}$$

$$\text{Test statistics (T.S.) } T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{85 - 81}{\sqrt{\frac{20.05}{12} + \frac{20.05}{10}}} = \mathbf{2.09}$$

$$\text{pooled variance } (S_p^2) : S_p^2 = \frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1) \times (4^2) + (10 - 1) \times (5^2)}{12 + 10 - 2} = 20.05$$

Rejection Region of H_0 (R.R.): (critical region : H_A Two sided)

$$\alpha = 0.05 \gg \text{Critical Value : } t_{1-\alpha} = t_{1-0.05} = t_{0.95} = 1.725$$

$$\text{df} = n_1 + n_2 - 2 = 12 + 10 - 2 = 20$$

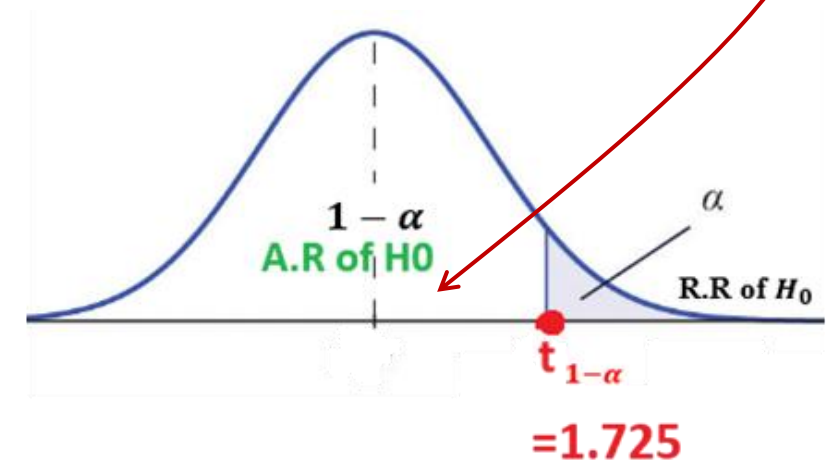
Decision :

$$\text{Reject } H_0: \mu_1 \leq \mu_2 \text{ if : } T_{T.S} > t_{1-\alpha}$$
$$2.09 > 1.725 \text{ (condition not satisfied)}$$

Decision is Accept $H_0: \mu_1 \leq \mu_2$ and Reject $H_A: \mu_1 > \mu_2$

Therefore, we conclude that the mean abrasive wear of material (1) is not greater than that of material (2).

Another way to take decision is by graph :
Determine test statistics value on the graph



7.4 Paired Comparisons:

- In this section, we are interested in comparing the means of two related (non-independent/dependent) normal populations.
- In other words, we wish to make statistical inference for the difference between the means of two related normal populations.
- Paired t-Test concerns about testing the equality of the means of two related normal populations.

Examples of related (dependent) populations are:

1. Height of the father and height of his son.
2. Mark of the student in MATH and his mark in STAT.
3. Pulse rate of the patient before and after the medical treatment.
4. Hemoglobin level of the patient before and after the medical treatment.

Test procedure

Let

X : the values of the first population .

Y : the values of the first population .

$D = \text{Values of } X - \text{Values of } Y$

Means

μ_1 : Mean of the first population .

μ_2 : Mean of the second population .

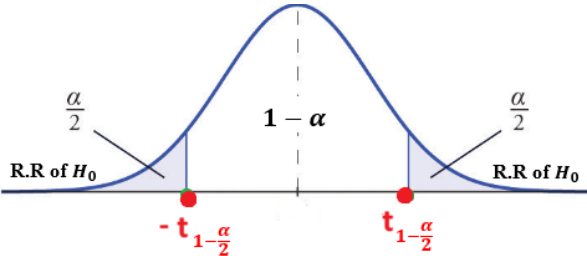
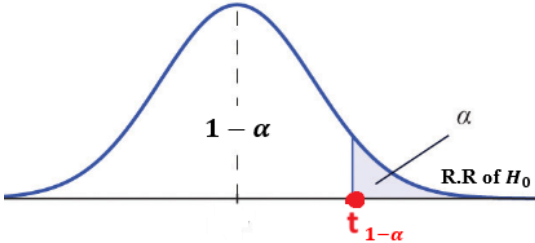
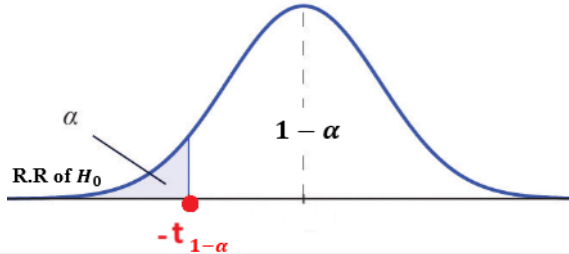
$\mu_D = \text{Mean of } X - \text{Mean of } Y$

$$(\mu_D = \mu_1 - \mu_2)$$

Confident interval about a difference of two population means ($\mu_D = \mu_1 - \mu_2$) :
(Dependent , Related populations)

calculate the following quantities: Or (Can use the calculator)	<p>-The differences (D-observations): $D_i = X_i - Y_i$ (i=1, 2, ..., n)</p> <p>-Sample mean of the D-observations: $\bar{D} = \frac{\sum_{i=1}^n X_i}{n}$</p> <p>-Sample variance of the D-observations: $S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}$</p> <p>-Sample standard deviation of the D-observations: $S_D = \sqrt{S_D^2}$</p>
Confident interval for $\mu_D = \mu_1 - \mu_2$	
A 100% confidence interval for μ_D	$\bar{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}, \quad df = n-1$

Testing hypothesis about a difference of two population means ($\mu_D = \mu_1 - \mu_2$) : (Dependent , Related populations)

Hypothesis	$H_0: \mu_1 - \mu_2 = 0$ vs $H_A: \mu_1 - \mu_2 \neq 0$ Or $H_0: \mu_D = 0$ vs $H_A: \mu_D \neq 0$	$H_0: \mu_1 - \mu_2 \leq 0$ vs $H_A: \mu_1 - \mu_2 > 0$ Or $H_0: \mu_D \leq 0$ vs $H_A: \mu_D > 0$	$H_0: \mu_1 - \mu_2 \geq 0$ vs $H_A: \mu_1 - \mu_2 < 0$ Or $H_0: \mu_D \geq 0$ vs $H_A: \mu_D < 0$
Test statistic(T.S)	$T = \frac{\bar{D}}{S/\sqrt{n}}, df = n-1$		
R.R. and A.R. of H_0			
Critical value (s)	$t_{(1-\frac{\alpha}{2})}$ and $-t_{(1-\frac{\alpha}{2})}$	$t_{1-\alpha}$	$-t_{1-\alpha}$
Decision :	Reject H_0 (and accept H_A) at the significance level α if:		
Reject H_0 if the following condition satisfied.	$T_{T.S} > t_{1-\frac{\alpha}{2}}$ or $T_{T.S} < -t_{1-\frac{\alpha}{2}}$	$T_{T.S} > t_{1-\alpha}$	$T_{T.S} < -t_{1-\alpha}$
	T.S. \in R.R. Two-Sided Test	T.S. \in R.R. One-Sided Test	T.S. \in R.R. One-Sided Test

Example (effectiveness of a diet program)

Suppose that we are interested in study the effectiveness of a certain diet program on ten individual . Let the random variable X and Y given as following table :

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X_i)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y_i)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7

1. Find a 95% confidence interval for the difference between the mean of weights before the diet program (μ_1) and the mean of weights after the diet program (μ_2) [$\mu_D = \mu_1 - \mu_2$].
2. Does these data provide sufficient evidence to allow us to conclude that the diet program is effective ? Use $\alpha=0.05$ and assume that the populations are normal.

Solution:

Let the random variables X and Y are as follows:

X = the weight of the individual **before** the diet program

Y = the weight of the same individual **after** the diet program.

We assume that the distributions of these random variables are normal with means μ_1 and μ_2 , respectively .

Populations:

1-st population (X): weights **before** a diet program mean

2-nd population (Y): weights **after** the diet program mean

These two variables are related (dependent/non-independent) because they are measured on the same individual.

We select a random sample of n individuals. At the beginning of the study, we record the individuals' weights before the diet program (X). At the end of the diet program, we record the individuals' weights after the program (Y). We end up with the following information and calculations:

Individual I	Weight before X_i	Weight after Y_i	Difference D_i = X_i - Y_i
1	X₁	Y₁	D₁ = X₁ - Y₁
2	X₂	Y₂	D₂ = X₂ - Y₂
.	.	.	.
.	.	.	.
.	.	.	.
n	X_n	Y_n	D_n = X_n - Y_n

Find the following measures by calculator or rules

- The sample mean of the D-observations: $\bar{D} = \frac{\sum_{i=1}^n X_i}{n} = \frac{54.5}{10} = 5.45$
- Sample variance of the D-observations: $S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1} = \frac{(6.9-5.45)^2 + \dots + (12.4-5.45)^2}{10-1} = 50.328$
- Sample standard deviation of the D-observations: $S_D = \sqrt{S_D^2} = \sqrt{50.328} = 7.09$

First :

Calculate difference $D_i = X_i - Y_i$
(Last column)

Then, From calculator,

-Enter the data in the last column
($D_i = X_i - Y_i$)

- Find the sample mean of the
D-observations ($\bar{D} = 5.45$)

-Find the sample standard deviation of the
D-observations ($S_D = 7.09$)

i	X_i	Y_i	$D_i = X_i - Y_i$
1	86.6	79.7	6.9
2	80.2	85.9	-5.7
3	91.5	81.7	9.8
4	80.6	82.5	-1.9
5	82.3	77.9	4.4
6	81.9	85.8	-3.9
7	88.4	81.3	7.1
8	85.3	74.7	10.6
9	83.1	68.3	14.8
10	82.1	69.7	12.4
Sum			$\sum_{i=1}^{10} D_i = 54.5$

1. We need to find a 95% confidence interval for $\mu_D = \mu_1 - \mu_2$

Reliability coefficient $t_{1-\alpha/2}$:

$$\alpha = 0.05, \quad t_{1-\frac{\alpha}{2}} = t_{1-\frac{0.05}{2}} = t_{0.975} = 2.262 \quad (\text{df} = n - 1 = 10 - 1 = 9)$$

The 95% C.I for $\mu_D = \mu_1 - \mu_2$

$$\bar{D} \pm t_{1-\alpha/2} \frac{s_D}{\sqrt{n}}$$

$$5.45 \pm 2.262 \times \frac{7.09}{\sqrt{10}}$$

$$5.45 \pm 5.0715$$

$$(5.45 - 5.0715, 5.45 + 5.0715)$$

$$(0.38, 10.52)$$

$$0.38 < \mu_D < 10.52$$

2. Does these data provide sufficient evidence to allow us to conclude that the diet program is effective?

Use $\alpha=0.05$ and assume that the populations are normal.

μ_1 =Mean of weights before a diet program mean

μ_2 =Mean of weights after the diet program mean

μ_D =Mean of X – Mean of Y ($\mu_D = \mu_1 - \mu_2$)

Hypotheses:

H_0 : the diet program has no effect on weight

H_A : the diet program has an effect on weight

Equivalently,

$H_0: \mu_1 = \mu_2$ (no effect) VS $H_A: \mu_1 \neq \mu_2$ (There is effect)

$H_0: \mu_1 - \mu_2 = 0$ VS $H_A: \mu_1 - \mu_2 \neq 0$

$H_0: \mu_D = 0$ VS $H_A: \mu_D \neq 0$ ($\mu_D = \mu_1 - \mu_2$)

Hypotheses:

$$\begin{array}{l} H_0 : \mu_1 = \mu_2 \\ H_A : \mu_1 = \mu_2 \end{array} \quad \text{OR} \quad \begin{array}{l} (H_0 : \mu_1 - \mu_2 \neq 0) \\ (H_A : \mu_1 - \mu_2 \neq 0) \end{array} \quad \text{OR} \quad \begin{array}{l} H_0 : \mu_D \neq 0 \\ H_A : \mu_D \neq 0 \end{array}$$

$$\text{Test statistics (T.S.) } T = \frac{\bar{D}}{s_D/\sqrt{n}} = \frac{5.45}{7.09/\sqrt{10}} = \mathbf{2.43}$$

Rejection Region of H_0 (R.R.): (critical region : H_A Two sided)

$$\alpha = 0.05 \gg \text{Critical Value : } \pm t_{1-\frac{\alpha}{2}} = \pm t_{1-\frac{0.05}{2}} = \pm t_{0.975} = \pm 2.262$$

$$\text{df} = n - 1 = 10 - 1 = 9$$

Decision :

Reject $H_0 : \mu_1 = \mu_2$ if :

$$T_{T.S} < -t_{1-\frac{\alpha}{2}} \quad \text{or} \quad T_{T.S} > t_{1-\frac{\alpha}{2}}$$

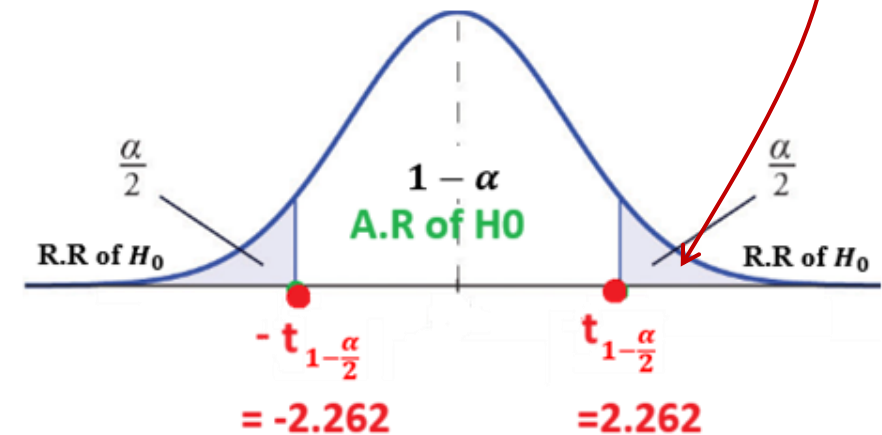
$2.43 < -2.262$ (First condition not satisfied, then try the second condition)

$2.43 > 2.262$ (Second condition satisfied)

Decision is Reject $H_0 : \mu_1 = \mu_2$ (no effect) and Accept $H_A : \mu_1 \neq \mu_2$ (effect)

We conclude that there is effect of the diet program.

Another way to take decision is by graph :
Determine test statistics value on the graph



Note:

The sample mean of the weights before the program is ($\bar{X} = 84.2$)

The sample mean of the weights after the program is ($\bar{Y} = 78.75$)

Since the diet program is effective and since

$$\bar{X} > \bar{Y}$$

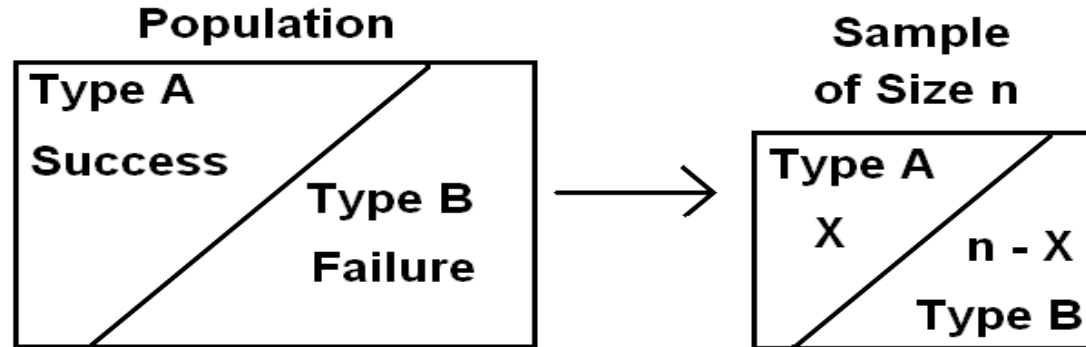
$$84.2 > 78.75$$

Weight before $>$ weight after

we can conclude that the program is effective in reducing the weight (the diet is good).

7.5 Hypothesis Testing: A Single Population Proportion (p):

In this section, we are interested in testing some hypotheses about the population proportion (p).



Recall:

p = Population proportion of elements of Type A in the population.

$$P = \frac{\text{No. of elements of type A in the population}}{\text{Total number of elements in the population}} = \frac{A}{N} \quad (N = \text{Population size})$$

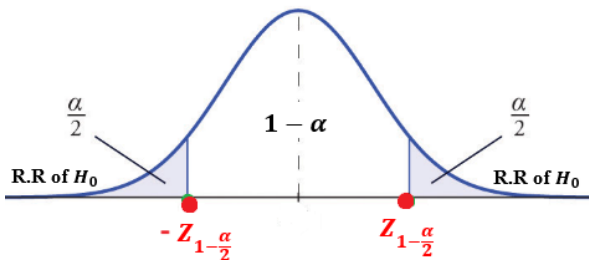
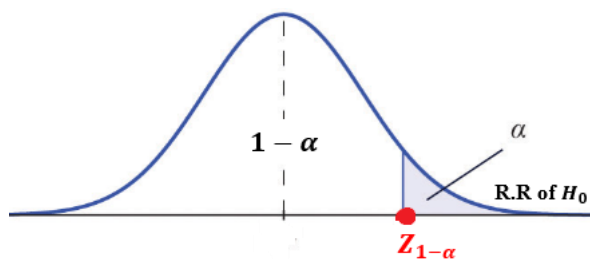
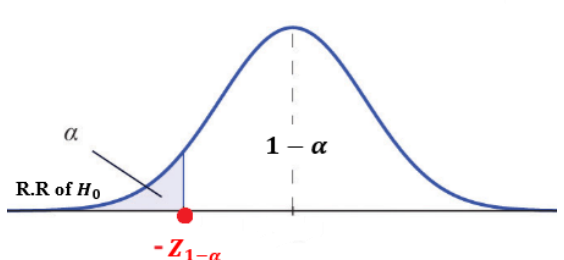
$$\hat{P} = \frac{\text{No. of elements of type A in the sample}}{\text{Total number of elements in the sample}} = \frac{X}{n} \quad (n = \text{Sample size})$$

\hat{P} is a "good" point estimate for p .

For large n , ($n \geq 30$, $np > 5$, $nq > 5$)

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

Test procedures:

Hypotheses	$H_0 : P = P_0$ $H_A : P \neq P_0$	$H_0 : P \leq P_0$ $H_A : P > P_0$	$H_0 : P \geq P_0$ $H_A : P < P_0$
Test Statistic (T.S.)	$Z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0 q_0}{n}}} \sim N(0, 1) \quad P_0 : \text{be a given known value and } q_0 = 1 - P_0$		
R.R. & A.R. of H_0			
Critical value(s)	$Z_{1-\frac{\alpha}{2}}$ and $Z_{1-\frac{\alpha}{2}}$	$Z_{1-\alpha}$	$-Z_{1-\alpha}$
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$Z_{T.S} > Z_{1-\frac{\alpha}{2}}$ or $Z_{T.S} < -Z_{1-\frac{\alpha}{2}}$	$Z_{T.S} > Z_{1-\alpha}$	$Z_{T.S} < -Z_{1-\alpha}$
	T.S. \in R.R. Two - Sided Test	T.S. \in R.R. One - Sided Test	T.S. \in R.R. One - Sided Test

Example

A researcher was interested in the proportion of females in the population of all patients visiting a certain clinic. The researcher claims that 70% of all patients in this population are females. Would you agree with this claim if a random survey shows that 24 out of 45 patients are females? Use a 0.10 level of significance.

Solution:

p = Proportion of female in the population.

$n=45$ (large)

X = no. of female in the sample = 24

\hat{P} = proportion of females in the sample

$$\hat{P} = \frac{X}{n} = \frac{24}{45} = 0.533$$

$\alpha=0.10$ (level of significance)

Hypotheses:

$$H_0 : P = 0.7 \quad (P_0 = 0.7) \text{ (} H_0 \text{ is the researcher claim)}$$

$$H_A : P \neq 0.7$$

Test statistics (T.S.)

$$q_o = 1 - P_0 = 1 - 0.7 = 0.3$$

$$Z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0 q_0}{n}}} = \frac{0.533 - 0.7}{\sqrt{\frac{0.7 \times 0.3}{45}}} = -2.44$$

Rejection Region of H_0 (R.R.): (critical region)

$$\alpha = 0.10 \gg \text{Critical Value : } Z_{1-\frac{\alpha}{2}} = Z_{1-\frac{0.10}{2}} = Z_{0.95} = 1.645$$

Decision :

$$\text{Reject } H_0 : P = 0.7 \text{ if : } Z_{T.S} < -Z_{1-\frac{\alpha}{2}} \text{ or } Z_{T.S} > Z_{1-\frac{\alpha}{2}}$$

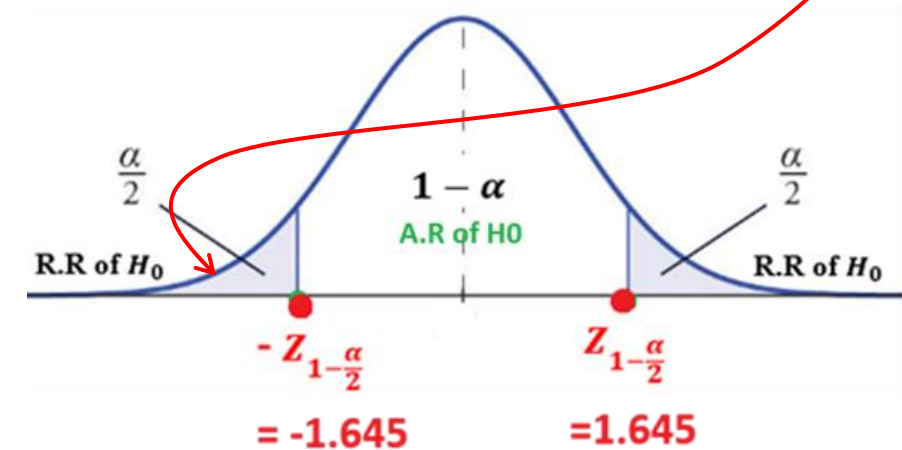
$$-2.44 < -1.645$$

(First condition satisfied, not need to try the second condition)

Decision is Reject $H_0 : P = 0.7$ and Accept $H_A : P \neq 0.7$

Therefore, we do not agree with the claim stating that 70% of the patients in this population are females.

Another way to take decision is by graph :
Determine test statistics value on the graph



Example (Reading)

In a study on the fear of dental care in a certain city, a survey showed that 60 out of 200 adults said that they would hesitate to take a dental appointment due to fear. Test whether the proportion of adults in this city who hesitate to take dental appointment is less than 0.25. Use a level of significance of 0.025.

Solution:

p = Proportion of adults in the city who hesitate to take a dental appointment.

$n = 200$ (large)

X = no. of adults who hesitate in the sample = 60

\hat{P} = proportion of adults who hesitate in the sample.

$$\hat{P} = \frac{X}{n} = \frac{60}{200} = 0.3$$

$\alpha = 0.025$ (level of significance)

Hypotheses:

$$H_0 : P \geq 0.25 \quad (P_0 = 0.25)$$

$$H_A : P \leq 0.25$$

Test statistics (T.S.)

$$q_0 = 1 - P_0 = 1 - 0.25 = 0.75$$

$$Z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0 q_0}{n}}} = \frac{0.3 - 0.25}{\sqrt{\frac{0.25 \times 0.75}{200}}} = 1.633$$

Rejection Region of H_0 (R.R.): (critical region)

$$\alpha = 0.025 \gg \text{Critical Value : } Z_{1-\alpha} = Z_{1-0.025} = Z_{0.975} = 1.96$$

Decision :

Reject $H_0 : P = 0.7$ if :

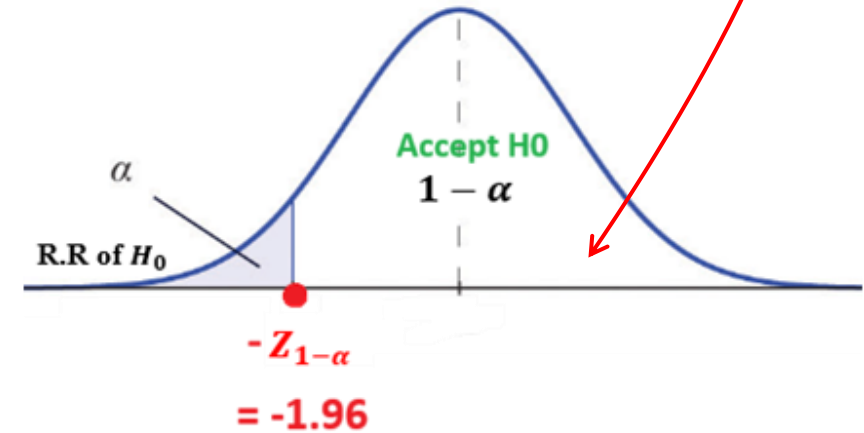
$$Z_{T.S} < -Z_{1-\alpha}$$

$$1.633 < -1.96$$

condition not satisfied

Decision is Accept $H_0 : P \geq 0.25$ and Reject $H_A : P < 0.25$

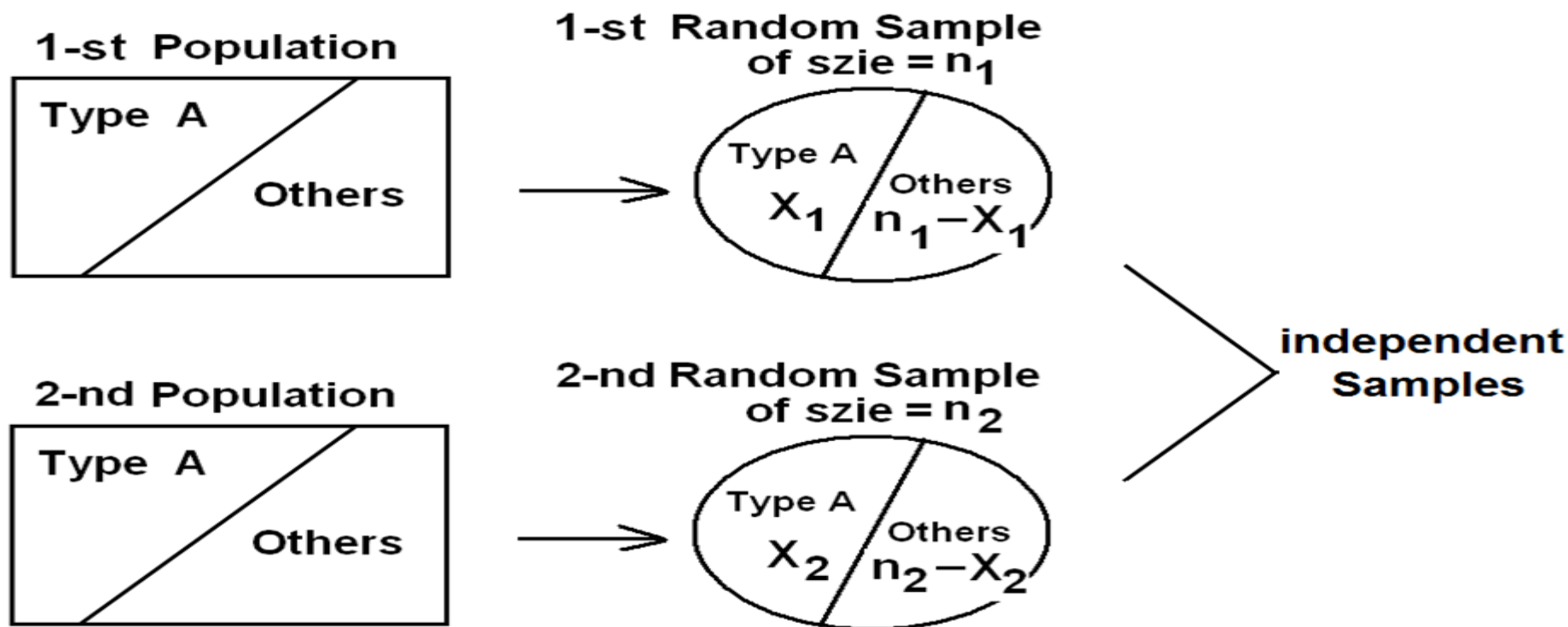
Another way to take decision is by graph :
Determine test statistics value on the graph



7.6 Hypothesis Testing:

The Difference Between Two Population Proportions ($p_1 - p_2$):

In this section, we are interested in testing some hypotheses about the difference between two population proportions ($P_1 - P_2$)



Suppose that we have two populations:

- P_1 = population proportion of the 1-st population.
- P_2 = population proportion of 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 1-st sample.

Hypotheses:

We choose one of the following situations:

- (i) $H_0: p_1 = p_2$ against $H_A: p_1 \neq p_2$
- (ii) $H_0: p_1 \geq p_2$ against $H_A: p_1 < p_2$
- (iii) $H_0: p_1 \leq p_2$ against $H_A: p_1 > p_2$

or equivalently,

- (i) $H_0: p_1 - p_2 = 0$ against $H_A: p_1 - p_2 \neq 0$
- (ii) $H_0: p_1 - p_2 \geq 0$ against $H_A: p_1 - p_2 < 0$
- (iii) $H_0: p_1 - p_2 \leq 0$ against $H_A: p_1 - p_2 > 0$

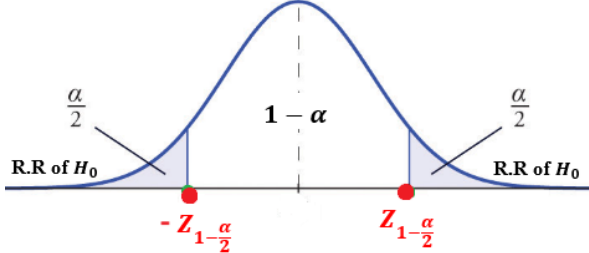
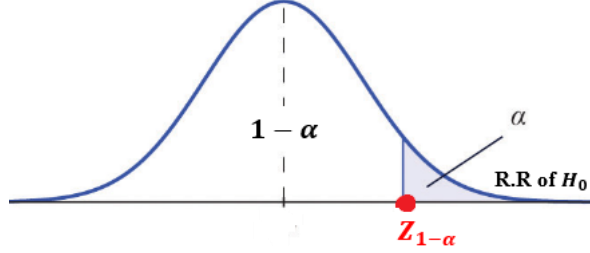
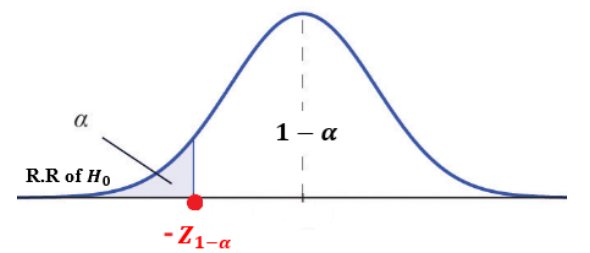
Note, under the assumption of the equality of the two population proportions ($H_0: P_1 = P_2 = P$), the pooled estimate of the common proportion p (**pooled proportion**) is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} \quad (\bar{q} = 1 - \bar{p})$$

The test statistic (T.S.) is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \sim N(0,1)$$

Test procedures:

Hypotheses	$H_0 : P_1 - P_2 = 0$ $H_A : P_1 - P_2 \neq 0$	$H_0 : P_1 - P_2 \leq 0$ $H_A : P_1 - P_2 > 0$	$H_0 : P_1 - P_2 \geq 0$ $H_A : P_1 - P_2 < 0$
Test Statistic (T.S.)	$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{\bar{P}\bar{q}}{n_1} + \frac{\bar{P}\bar{q}}{n_2}}} \sim N(0,1)$		
R.R. & A.R. of H_0			
Critical value(s)	$Z_{1-\frac{\alpha}{2}}$ and $Z_{1-\frac{\alpha}{2}}$	$Z_{1-\alpha}$	$-Z_{1-\alpha}$
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$Z_{T.S} > Z_{1-\frac{\alpha}{2}}$ or $Z_{T.S} < -Z_{1-\frac{\alpha}{2}}$	$Z_{T.S} > Z_{1-\alpha}$	$Z_{T.S} < -Z_{1-\alpha}$
	T.S. \in R.R. Two - Sided Test	T.S. \in R.R. One - Sided Test	T.S. \in R.R. One - Sided Test

Example:

In a study about the obesity (overweight), a researcher was interested in comparing the proportion of obesity between males and females. The researcher has obtained a random sample of 150 males and another independent random sample of 200 females. The following results were obtained from this study.

	n	Number of obese people (X)
Males	150	21
Females	200	48

Can we conclude from these data that there is a difference between the proportion of obese males and proportion of obese females? Use $\alpha = 0.05$.

Solution :

P_1 = population proportion of obese males.

P_2 = population proportion of obese females .

\hat{P}_1 = sample proportion of obese males

\hat{P}_2 = sample proportion of obese females

$\alpha=0.05$ (level of significance)

Male

$$X_1 = 21$$

$$X_2 = 150$$

$$\hat{P}_1 = \frac{X_1}{n_1} = \frac{21}{150} = 0.14$$

Female

$$X_2 = 48$$

$$n_2 = 200$$

$$\hat{P}_2 = \frac{X_2}{n_2} = \frac{48}{200} = 0.24$$

The pooled estimate of the common proportion p is:

$$\bar{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{21 + 48}{150 + 200} = 0.197 \quad (\bar{q} = 1 - \bar{P} = 1 - 0.197 = 0.803)$$

Hypotheses:

$$H_0: p_1 = p_2 \quad \text{vs} \quad H_A: p_1 \neq p_2$$

or

$$H_0: p_1 - p_2 = 0 \quad \text{vs} \quad H_A: p_1 - p_2 \neq 0$$

Hypotheses:

$$\begin{aligned} H_0 : P_1 &= P_2 & \text{OR} & & (H_0: P_1 - P_2 = 0) \\ H_A : P_1 &\neq P_2 & & & (H_A: P_1 - P_2 \neq 0) \end{aligned}$$

Test statistics (T.S.)

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{\bar{P}\bar{q}}{n_1} + \frac{\bar{P}\bar{q}}{n_2}}} = \frac{0.14 - 0.24}{\sqrt{\frac{0.196 \times 0.803}{150} + \frac{0.196 \times 0.803}{200}}} = -2.328$$

Rejection Region of H_0 (R.R.): (critical region)

$$\alpha = 0.05 \gg \text{Critical Value : } Z_{1-\frac{\alpha}{2}} = Z_{1-\frac{0.05}{2}} = Z_{(0.975)} = 1.96$$

Decision :

Reject $H_0: P_1 = P_2$ if :

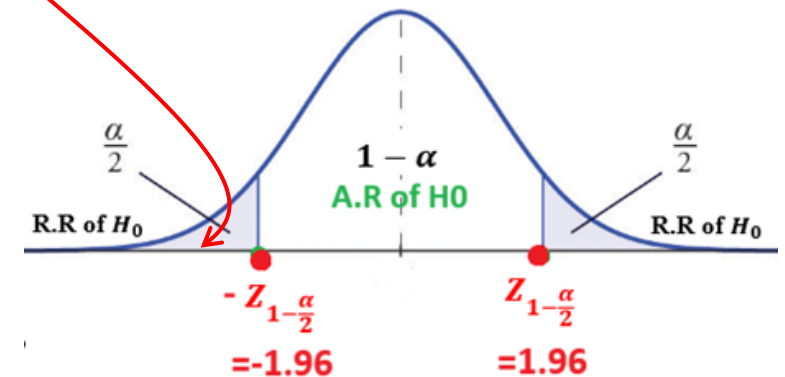
$$Z_{T.S} < -Z_{1-\frac{\alpha}{2}} \text{ or } Z_{T.S} > Z_{1-\frac{\alpha}{2}}$$

$-2.328 < -1.96$ (First condition satisfied, then do not try the second condition)

Decision is Reject $H_0: P_1 = P_2$ and Accept $H_A: P_1 \neq P_2$

Therefore, we conclude that there is a difference between the proportion of obese males and the proportion of obese females

Another way to take decision is by graph :
Determine test statistics value on the graph



Since $Z = -2.328 \in R.R.$, we reject $H_0: P_1 = P_0$ and accept $H_A: P_1 \neq P_2$ at $\alpha = 0.05$. Therefore, we conclude that there is a difference between the proportion of obese males and the proportion of obese females. Additionally, since,

$$\hat{P}_1 = 0.14 < \hat{P}_2 = 0.24$$

We may conclude that the proportion of obesity for females is larger than that for males.

TABLE OF CONTENTS

Subject	Page
Outline of the course	
Marks Distribution and Schedule of Assessment Tasks	
CHAPTER 1: Organizing and Displaying Data	
Introduction	
Statistics	
Biostatistics	
Populations	
Population Size	
Samples	
Sample Size	
Variables	
- Types of Variables	
-Types of Quantitative Variables	
Organizing the Data	
Simple frequency distribution or ungrouped frequency distribution	
Grouped Frequency Distributions	
Width of a class interval	
Displaying Grouped Frequency Distributions	

CHAPTER 2: Basic Summary Statistics	
Measures of Central Tendency	
Mean	
-Population Mean	
-Sample Mean	
-Advantages and Disadvantages of the Mean	
Median	
-Advantages and Disadvantages of the Median	
Mode	
-Advantages and Disadvantages of the Mode	
Measures of Dispersion (Variation)	
Range	
Variance	
Deviations of Sample Values from the Sample Mean	
Population Variance	
Sample Variance	
Calculating Formula for the Sample Variance	
Standard Deviation	
Coefficient of Variation	

CHAPTER 3: Basic Probability Concepts	
General Definitions and Concepts	
Probability	
An Experiment	
Sample Space	
Events	
Equally Likely Outcomes	
Probability of an Event	
Some Operations on Events	
Union of Two events	
Intersection of Two Events	
Complement of an Event	
General Probability Rules	
Applications	
Conditional Probability	
Independent Events	
Bayes' Theorem	

CHAPTER 4: Probability Distributions	
Introduction	
Probability Distributions of Discrete R.V.'s	
Graphical Presentation	
Population Mean of a Discrete Random Variable	
Cumulative Distributions	
Binomial Distribution	
Poisson Distribution	
Probability Distributions of Continuous R.V.	
Normal Distribution	
Standard Normal Distribution	
Calculating Probabilities of Standard Normal Distribution	
Calculating Probabilities of Normal Distribution	
t-distribution	

CHAPTER 5: Sampling Distribution	
Results about Sampling Distribution of the Sample Mean	
-Sampling Distribution of the single Sample Mean	
-Sampling Distribution of the Difference between Two Means	
Results about Sampling Distribution of the Sample Proportion	
-Sampling Distribution of Proportion	
-Sampling Distribution of the Difference between Two Proportions	

CHAPTER 6 & 7: Statistical Inferences (Estimation and Hypotheses Testing)	
CHAPTER 6 : Estimation	
Estimation of the Population Mean:	
- Point Estimation of the population mean	
- Interval Estimation of the population mean	
Estimation for the Population Proportion :	
- Point Estimate for the Population Proportion	
- Interval Estimation for the Population Proportion	
CHAPTER 7: Tests of Hypotheses	
Tests of Hypotheses	
Single Sample: Tests Concerning a Single Mean	
Two Samples: Tests on Two Means	
Two-sample: Paired t-test	
Single Sample: Tests on a Single Proportion:	
Two Samples: Tests on Two Proportions	