CHAPTER 6: Using Sample Data to Make Estimations About Population Parameters

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.

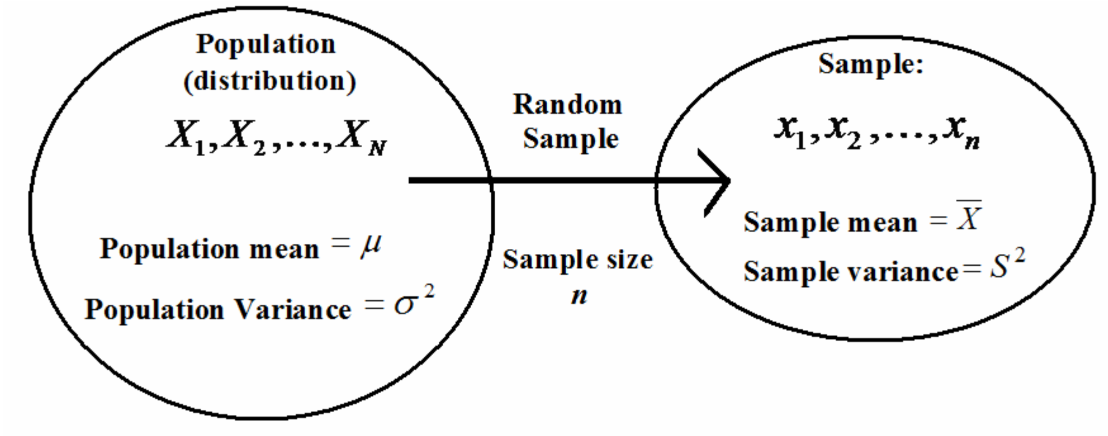
(2) Hypothesis Testing: (chapter 7)
Answering research questions about the unknown parameters of the population (confirming or denying some conjectures or statements about the unknown parameters).

6.1: 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	$\mu_1 - \mu_2$	$\overline{X_1} - \overline{X_2}$
The Difference between Two Proportion	$P_1 - P_2$	$\widehat{P_1} - \widehat{P_2}$

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:	Sample:
Population Size = N	Sample Size = n
Population Values: $X_1, X_2, \dots, X_N$	Sample values: $x_1, x_2, \dots, x_n$
Population Mean: $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$	Sample Mean: $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$	Sample Variance: $S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$

(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: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Result:

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

(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 ?

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

(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 $1-\alpha$".

(confidence level), degree of confidence

* $1-\alpha$ = is called the confidence coefficient (level)
* L = lower limit of the confidence interval
* U = upper limit of the confidence interval

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

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

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

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

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

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

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

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

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

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

Note that:

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

2. Upper limit of the confidence interval = $\overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

3. Lower limit of the confidence interval = $\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

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:

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

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

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

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

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

where the degrees of freedom is:

$$df = \nu = n - 1.$$

Note that:

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

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

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

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

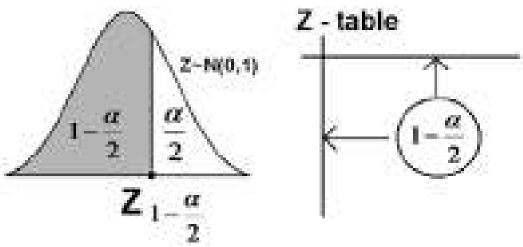
5. In general the interval estimate (confidence interval) may be expressed as follows:

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

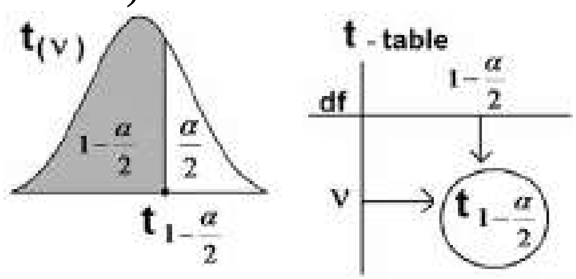
$$\overline{X} \pm t_{1-\frac{\alpha}{2}}\ \hat{\sigma}_{\overline{X}}$$

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: $(df = \nu = n - 1)$



Example:

Suppose that Z ~ 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 $\alpha = 0.1$:

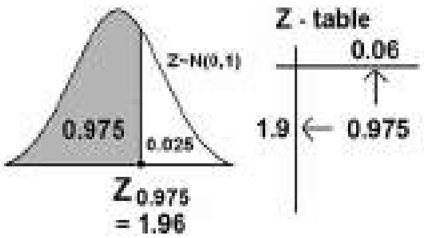
$$1-\frac{\alpha}{2}=1-\frac{0.1}{2}=0.95 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$$

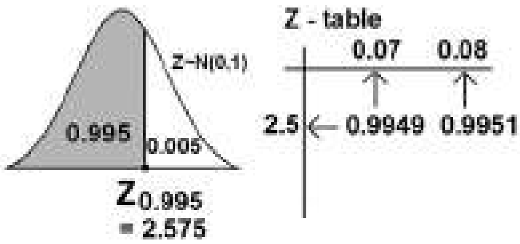
(2) For $\alpha = 0.05$:

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

(3) For $\alpha = 0.01$:

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





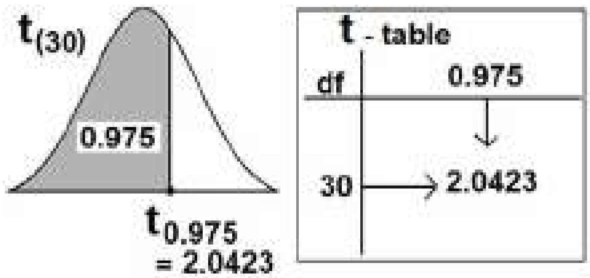
Example:

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

Solution:

df = ν = 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:

The $(1-\alpha)100\%$ Confidence Interval for the Population Mean μ

1. Normal Distribution + σ^2 is known.
2. Non-Normal Distribution + $n \geq 30$ + σ^2 is known.

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

Normal Distribution + σ^2 is unknown + $n < 30$.

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

Example: (The case where ^{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 (Continuous quantitative variable).

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

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

Distribution is normal with standard deviation $= 3.3$.

^{2} is known ($\sigma^2 = 10.89$)

X ~ Normal(, 10.89)

= ?? (unknown- we need to estimate)

Sample size: $n = 123$ (large)

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

(1) Point Estimation:

We need to find a point estimate for .

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

≈ 26.2

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

We need to find 90% C. I. for .

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

$$1- \quad = 0.9 \Leftrightarrow \quad = 0.1 \Leftrightarrow \frac{\ }{2} = 0.05 \quad \Leftrightarrow \quad 1-\frac{\ }{2} = 0.95$$

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

90% confidence interval for is:

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

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

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

$$(25.710529\ ,\ 26.689471)$$

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: $\mu =$ 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: $n = \mathbf{21}$

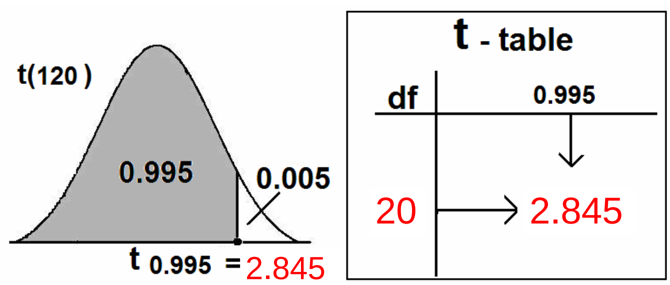
Sample mean: $\overline{X} = 37$

Sample standard deviation: $S = 10$
Degrees of freedom: df = ν = 21 - 1 = 20
(a) Point Estimation: We need to find a point estimate for μ.
$\overline{X} = 37$ 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.

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

$$1-\alpha = 0.99 \Leftrightarrow \alpha = 0.01 \quad \Leftrightarrow \quad \frac{\alpha}{2} = 0.005 \quad \Leftrightarrow \quad 1-\frac{\alpha}{2} = 0.995$$

ν = df = 21-1=20
The reliability coefficient is: $t_{1-\frac{\alpha}{2}} = t_{0.995} = 2.845$



99% confidence interval for μ is:

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

$$37 \pm (2.845)\frac{10}{\sqrt{21}}$$

$$37 \pm 6.208$$

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

$$(\,30.792\,,\,43.208\,)$$

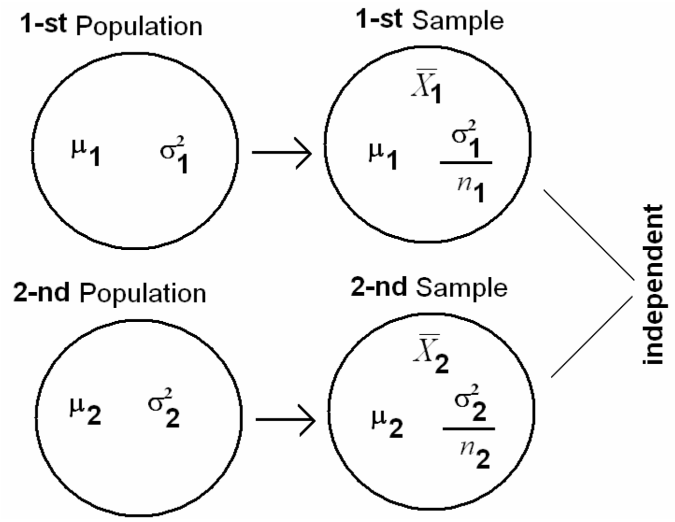
$$30.792 < \mu < 43.208$$

We are 99% confident that the true value of the mean μ lies in the interval (30.792 , 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 $\mu_1 - \mu_2$.



Recall:

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

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

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

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 $(\overline{X}_1 - \overline{X}_2)$ has a normal distribution with mean $(\mu_1 - \mu_2)$ and variance $((\sigma_1^2/n_1) + (\sigma_2^2/n_2))$, that is:

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

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

Point Estimation of $\mu_1 - \mu_2$:

Result:

$\overline{X}_1 - \overline{X}_2$ 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:

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

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

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

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

Estimator ± (Reliability Coefficient) × (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

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

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=$\nu=n_1+n_2-2$ degrees of freedom.

~~**Example:** (1st Case:~~ $\cancel{\sigma_1^2}$ ~~and~~ $\cancel{\sigma_2^2}$ ~~are known)~~

~~An experiment was conducted to compare time length~~

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

The $(1-\alpha)100\%$ Confidence Interval for the Difference between two Population Means $\mu_1 - \mu_2$

1. Normal Distribution + σ_1^2 and σ_2^2 are known
2. Non-Normal Distribution + $n_1, n_2 \geq 30 + \sigma_1^2$ and σ_2^2 are known.

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

Normal Distribution + $n_1, n_2 < 30$ + $\sigma_1^2 = \sigma_2^2 = \sigma^2$ are unknown but equal

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

pooled variance:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}, \text{d.f} = n_1 + n_2 - 2$$

Example:

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 $\mu_A - \mu_B$:

$\alpha = ??$

$96\% = (1-\alpha)100\% \Leftrightarrow 0.96 = (1-\alpha) \Leftrightarrow \alpha = 0.04 \Leftrightarrow \alpha/2 = 0.02$

Reliability Coefficient: $Z_{1-\frac{\alpha}{2}} = Z_{0.98} = 2.055$

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, we conclude that the two population means are not equal ($\mu_A - \mu_B \neq 0 \Leftrightarrow \mu_A \neq \mu_B$). 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:

First we calculate the mean and the variances of the two samples, and we get:

Surgery	Type (A)	Type (B)
Sample Size	$n_A = 6$	$n_B = 6$
Sample Mean	$\overline{X}_A = 140.67$	$\overline{X}_B = 138.50$
Sample Variance	$S^2_A = 7.87$	$S^2_B = 7.10$

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

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

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

$95\% = (1-\alpha)100\% \Leftrightarrow 0.95 = (1-\alpha) \Leftrightarrow \alpha = 0.05 \Leftrightarrow \alpha/2 = 0.025$
.df $= \nu = n_A + n_B - 2 = 10$
Reliability Coefficient: $t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.228$

The pooled estimate of the common variance is:

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

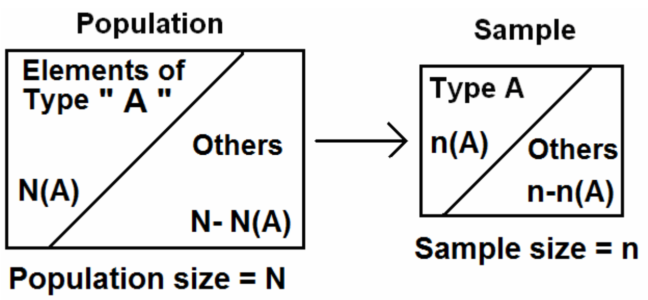
A 95% C.I. for $\mu_A - \mu_B$ is:

$$(\bar{X}_A - \bar{X}_B) \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_A} + \frac{S_p^2}{n_B}}$$
$$2.17 \pm (2.228)\sqrt{\frac{7.485}{6} + \frac{7.485}{6}}$$
$$2.17 \pm 3.519$$
$$-1.35 < \mu_A - \mu_B < 5.69$$

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

Note: Since the confidence interval includes zero, we conclude that the two population means may be equal ($\mu_A - \mu_B = 0 \Leftrightarrow \mu_A = \mu_B$). Therefore, we may conclude that the mean time length is the same for 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 a specified characteristic "A"

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

The population proportion is:

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

2. For the sample:

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

n = sample size

The sample proportion is:

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

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

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

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

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

(i) Point Estimate for (p):

Result:

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

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

Result:

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

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \; , \quad \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

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:

$$\hat{p} = \frac{n(A)}{n} = \frac{611}{950} = 0.643$$

(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:

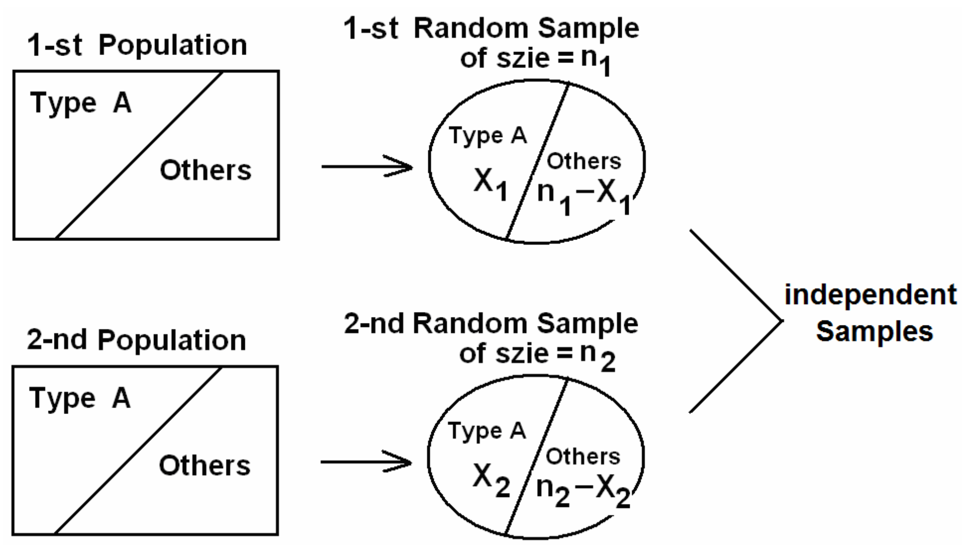
$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.643 \pm (1.96)\sqrt{\frac{(0.643)(1-0.643)}{950}}$$
$$0.643 \pm (1.96)(0.01554)$$
$$0.643 \pm 0.0305$$
$$(0.6127\ ,\ 0.6735)$$

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 <u>independently</u> 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: $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$
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 $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$, that is:

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

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

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

<u>Point Estimation for $p_1 - p_2$:</u>

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 ± (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:

$n_1 = 1500$, $X_1=75$

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

For the second sample we have:

$n_2 = 2000$, $X_2 = 80$

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

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

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

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

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

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

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

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

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

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{0.95} \sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\,\hat{q}_2}{n_2}}$$

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

$$0.01 \pm 0.01173$$

$$-0.0017 < p_1 - p_2 < 0.0217$$

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 ($p_1 - p_2 = 0 \Leftrightarrow p_1 = p_2$). Therefore, we may conclude that the proportion of people having cancer is the same in both cities.

CHAPTER 7: Using Sample Statistics To Test Hypotheses About Population Parameters:

In this chapter, we are interested in testing some hypotheses about the unknown population parameters.

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 $\theta > \theta_o$, where θ_o 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 sating that $\theta \geq \theta_o$ against $\theta < \theta_o$ (for some specific value θ_o).
- We usually test the null hypothesis (H_o) against the alternative (or the research) hypothesis (H_1 or H_A) by choosing one of the following situations:
  - (i) $H_o: \theta = \theta_o$ against $H_A: \theta \neq \theta_o$
  - (ii) $H_o: \theta \geq \theta_o$ against $H_A: \theta < \theta_o$
  - (iii) $H_o: \theta \leq \theta_o$ against $H_A: \theta > \theta_o$
- Equality sign must appear in the null hypothesis.
- H_o is the null hypothesis and H_A is the alternative hypothesis. (H_o and H_A are complement of each other)
- The null hypothesis (H_o) 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_o (Nature/reality)	
		H_o is true	H_o is false
Possible Action (Decision)	Accepting H_o	Correct Decision	Type II error (β)
	Rejecting H_o	Type I error (α)	Correct Decision

- There are two types of Errors:
  - Type I error = Rejecting H_o when H_o is true
    P(Type I error) = P(Rejecting Ho | Ho is true) = α
  - Type II error = Accepting Ho when Ho is false
    P(Type II error) = P(Accepting Ho | Ho is false) = β

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

  α = P(Rejecting H_o | H_o is true) = P(Type I error)

- There are 2 types of alternative hypothesis:
  - One-sided alternative hypothesis:
    - $H_O: \theta \geq \theta_o$ against $H_A: \theta < \theta_o$
    - $H_O: \theta \leq \theta_o$ against $H_A: \theta > \theta_o$
  - Two-sided alternative hypothesis:
    - $H_O: \theta = \theta_o$ against $H_A: \theta \neq \theta_o$

- 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_o (against H_A):

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

1- Determining Hypothesis

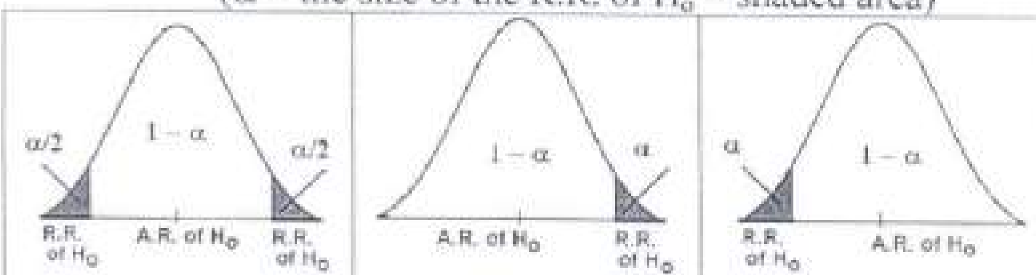
2. Determining a test statistic (T.S.)
We choose the appropriate test statistic based on the point estimator of the parameter.
The test statistic has the following form:

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

3. Determining the level of significance (α):

$$\alpha = 0.01,\ 0.025,\ 0.05,\ 0.10$$

4. Determining the rejection region of H_o (R.R.) and the acceptance region of H_o (A.R.).
The R.R. of H_o depends on H_A and α:
   - H_A determines the direction of the R.R. of H_o
   - α determines the size of the R.R. of H_o
     (α = the size of the R.R. of H_o = shaded area)



$H_A: \theta \neq \theta_o$	$H_A: \theta > \theta_o$	$H_A: \theta < \theta_o$
Two-sided alternative	One-sided alternative	One-sided alternative

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

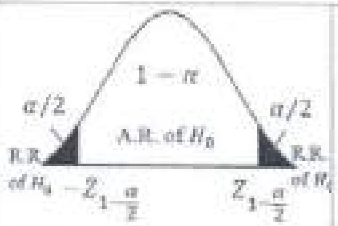
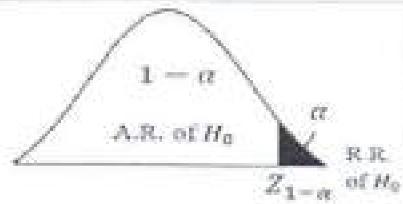
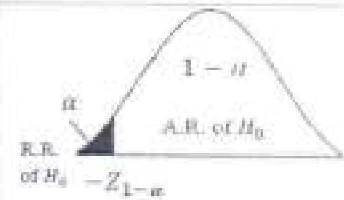
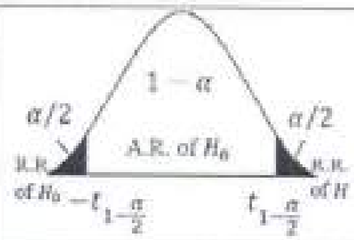
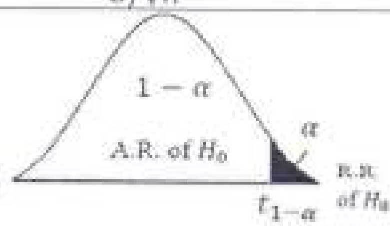
Notes:
1. The rejection region of H_o (R.R.) is sometimes called "the critical region".
2. The values which separate the rejection region (R.R.) and the acceptance region (A.R.) are called "the critical values" **or Relibility Cofficient.**

7.2 Hypothesis Testing: A Single Population Mean (μ):

Suppose that $X_1, X_2, \ldots, X_n$ 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).

Chapter 7 : Testing Hypothesis about population mean(μ):

Hypothesis	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_A: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_A: \mu < \mu_0$
First Case	σ is known; Normal or [Non-normal Distribution($n \geq 30$)]		
Test Statistic (T.S.)	$Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$		
Rejection Region(R.R) & Acceptance Region(A.R)			
Reliability Coefficient critical values	$-Z_{1-\alpha/2}$ or $Z_{1-\alpha/2}$	$Z_{1-\alpha}$	$-Z_{1-\alpha}$
Decision : Reject H_0 if the following condition satisfies	Reject H_0 (Accept H_A) at the significant level α if :		
	$Z > Z_{1-\alpha/2}$ Or $Z < -Z_{1-\alpha/2}$	$Z > Z_{1-\alpha}$ (one – Sided Test)	$Z < -Z_{1-\alpha}$ (one – Sided Test)
Second Case	σ is unknown; Normal, $n<30$ (small)		
Test Statistic (T.S.)	$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}, \quad df = n-1$		
Rejection Region(R.R) & Acceptance Region(A.R)			
Reliability Coefficient	$-t_{1-\alpha/2}$ or $t_{1-\alpha/2}$	$t_{1-\alpha}$	$-t_{1-\alpha}$
Decision : **Reject H_0 if the Following condition satisfies**	Reject H_0 (Accept H_A) at the significant level α if :		
	$T > t_{1-\alpha/2}$ Or $T < -t_{1-\alpha/2}$ (Two – Sided Test)	$T > t_{1-\alpha}$ (one – Sided Test)	$T < -t_{1-\alpha}$ (one – Sided Test)
Special Case	σ is unknown; Non-Normal, $n \geq 30$ (Large)		
Test Statistic (T.S.)	$Z = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}},$		
Rejection Region	Use the same R.R & A.R as in First Case(Z Case)		

Example: (first case: variance σ^2 is 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:

$n=100$ (large),

$\sigma = 8.9$ (σ known)

$\bar{X}=71.8$, $\sigma=8.9$ (σ is known)

μ =average (mean) life span

$\mu_0= 70$

$\alpha=0.05$

1) Hypotheses:

H₀: $\mu \leq 70$ ($\mu_0=70$)

H_A: $\mu > 70$ (research hypothesis)

H_o: $\mu \leq 70$ $(\mu_o = 70)$

H_A: $\mu > 70$ (research hypothesis)

Test statistics (T.S.) :

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

Level of significance:

$\alpha = 0.05$ $1-\alpha = 0.95$

Rejection Region of H_o (R.R.): (critical region) is $Z_{1-\alpha} = Z_{0.95} = 1.645$

We should reject H_o if: $Z(\text{test}) > Z_{1-\alpha}$

$2.02 > 1.645$ (condition satisfied)

✓

Decision : we reject H_o

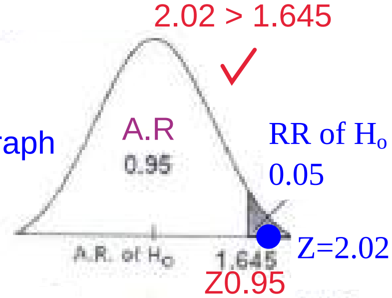
accept H_A

Another solution :

From curve:

1)determine test value on graph

2) Z(test)=2.02 in R.R,then reject H_o



Note: Using P- Value as a decision tool:

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

Calculating P-value:

* Calculating P-value depends on the alternative hypothesis H_A.
* Suppose that $z_c = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}}$ 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:

Z_c = Z (test)

Alternative Hypothesis:	HA: $\mu \neq \mu_o$	HA: $\mu > \mu_o$	HA: $\mu < \mu_o$
P - Value	$2\times P(Z > \vert Z_c \vert)$	$P(Z > Z_c)$	$P(Z > -Z_c)$
Significance Level =	α		
Decision	Reject Ho if P-value $\leq \alpha$.		

Example:

For the previous example, we have found that:

$$Z_C = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = 2.02$$

The alternative hypothesis was HA: $\mu > 70$.

P - Value = $P(Z > Z_C)$

$= P(Z > 2.02) = 1 - P(Z < 2.02) = 1 - 0.9783 = 0.0217$

The level of significance was $\alpha = 0.05$.

Since P-value $\leq \alpha$, we reject H_o.

Decision :Reject H_0 If :

P-value < α
0.0217<0.05

then
reject H0

Example: (second case: variance is 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:

The distribution is normal.

$n = 25$ (small)

$\bar{X} = 7.8$

S=0.5 (sample standard deviation): σ is unknown

μ = mean time of the visit , $\alpha = 0.01$

Hypotheses:

H_o: $\mu = 8$ (μ_o=8)

H_A: $\mu \neq 8$ (research hypothesis)

Test statistics (T.S.):

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{7.8 - 8}{0.5/\sqrt{25}} = -2$$

df = ν = n-1 = 25 -1= 24

Level of significance:

$\alpha = 0.01$, $\alpha/2 = 0.005$, $1 - \alpha/2 = 0.995$

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

$t_{1-\alpha/2} = t_{0.995} = 2.797$

We should reject Ho if:

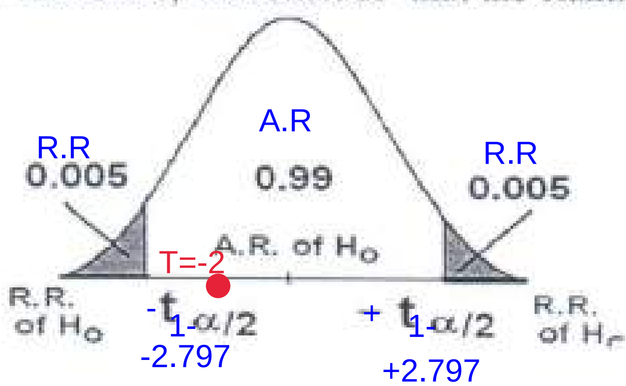
$T < -t_{1-\alpha/2}$ or $T > +t_{1-\alpha/2}$

-2 < -2.797 or -2 > 2.797

(both conditions not satisfied)
Decision: Accept H0

Decision:

Since T = −2 ∈ A.R., we accept H_o: $\mu = 8$ at $\alpha = 0.01$ and reject H_A: $\mu \neq 8$. Therefore, we conclude that the claim is correct.



Another solution:
From curve:

t (test)= -2 (In A.R)

Decision :Accept H_0

Note:

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:

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

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) $H_o: \mu_1 = \mu_2$ against $H_A: \mu_1 \neq \mu_2$

(ii) $H_o: \mu_1 \geq \mu_2$ against $H_A: \mu_1 < \mu_2$

(iii) $H_o: \mu_1 \leq \mu_2$ against $H_A: \mu_1 > \mu_2$

or equivalently,

(i) $H_o: \mu_1-\mu_2 = M_o$ against $H_A: \mu_1 - \mu_2 \neq M_o$

(ii) $H_o: \mu_1-\mu_2 \geq M_o$ against $H_A: \mu_1 - \mu_2 < M_o$

(iii) $H_o: \mu_1-\mu_2 \leq M_o$ against $H_A: \mu_1 - \mu_2 > M_o$

Test Statistic:

(1) First Case:

For normal populations (or non-normal populations with large sample sizes), and if σ_1^2 and σ_2^2 are known, then the test statistic is:

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

(2) Second Case:

For normal populations, and if σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then the test statistic is:

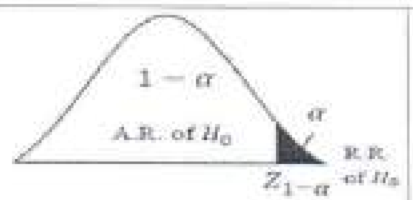
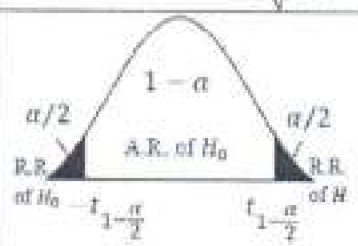
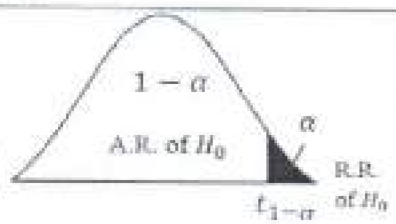
$$T = \frac{\overline{X}_1 - \overline{X}_2 - M_o}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

where the pooled 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 S_p^2 is df= $\nu = n_1 + n_2 - 2$.

Testing Hypothesis about difference between two population means ($\mu_1 - \mu_2$) : (Independent population)

Hypothesis	$H_0: \mu_1 - \mu_2 = M_0$ $H_A: \mu_1 - \mu_2 \neq M_0$	$H_0: \mu_1 - \mu_2 \leq M_0$ $H_A: \mu_1 - \mu_2 > M_0$	$H_0: \mu_1 - \mu_2 \geq M_0$ $H_A: \mu_1 - \mu_2 < M_0$
First Case	σ_1^2 , σ_2^2 are known + Normal or Non-Normal with large samples		
Test Statistic (T.S.)	$Z = \dfrac{\bar{X}_1 - \bar{X}_2 - M_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$		
Rejection Region(R.R) & Acceptance Region(A.R)			
Reliability Coefficient	$-Z_{1-\alpha/2}$ or $Z_{1-\alpha/2}$	$Z_{1-\alpha}$	$-Z_{1-\alpha}$
Decision : Reject H_0 if the following condition satisfies	Reject H_0 (Accept H_A) at the significant level α if :		
	$Z > Z_{1-\alpha/2}$ Or $Z < -Z_{1-\alpha/2}$	$Z > Z_{1-\alpha}$ (one – Sided Test)	$Z < -Z_{1-\alpha}$ (one – Sided Test)
Second Case	σ_1^2 , σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) + Normal		
Test Statistic (T.S.)	$T = \dfrac{\bar{X}_1 - \bar{X}_2 - M_0}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}}$, $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ $df = n_1 + n_2 - 2$		
Rejection Region(R.R) & Acceptance Region(A.R)			
Reliability Coefficient	$-t_{1-\alpha/2}$ or $t_{1-\alpha/2}$	$t_{1-\alpha}$	$-t_{1-\alpha}$
Decision : Reject H_0 if the Following condition satisfies	Reject H_0 (Accept H_A) at the significant level α if :		
	$T > t_{1-\alpha/2}$ Or $T < -t_{1-\alpha/2}$ (Two –sided test)	$T > t_{1-\alpha}$ (one – Sided Test)	$T < -t_{1-\alpha}$ (one – Sided Test)

Example: (σ_1^2 , 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/100ml
$\bar{X}_2 = 3.4$ mg/100ml

Assume the populations are normal with variances

$\sigma_1^2 = 1$

$\sigma_2^2 = 1.5$

. 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.

$n_1 = 12 \qquad \bar{X}_1 = 4.5 \qquad \sigma_1^2 = 1$

$n_2 = 15 \qquad \bar{X}_2 = 3.4 \qquad \sigma_2^2 = 1.5.$

Hypotheses:

H_o: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$

or

H_o: $\mu_1 - \mu_2 = 0$ against H_A: $\mu_1 - \mu_2 \neq 0$

Calculation:

1) $\alpha = 0.05$ 2) $1-\alpha/2 = 1-0.05/2 = 0.975$

3) $Z_{1-\alpha/2} = Z_{0.975} = 1.96$

Test Statistic (T.S.):

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - M_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{4.5 - 3.4 - 0}{\sqrt{\frac{1}{12} + \frac{1.5}{15}}} = 2.569$$

Decision : Reject H_0 If

Z(test)> $Z_{0.975}$ or Z < $-Z_{0.975}$

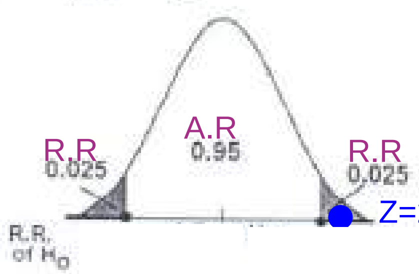
2.569 >1.96

✓

first condition satisfied
no need to check the second one

Decision : Reject H_0

Another solution:
from curve
Since
Z(test)=2.569 in R.R
Decision:
Reject H_0



Decision:

Since Z=2.569 ∈R.R. we reject H_o: $\mu_1=\mu_2$ and we accept (do not reject) H_A: $\mu_1 \neq \mu_2$ at α=0.05. Therefore, we conclude that the two population means are not equal.

Notes:

1. We can easily show that a 95% confidence interval for ($\mu_1 - \mu_2$) is (0.26, 1.94), that is:

$$0.26 < \mu_1 - \mu_2 < 1.94$$

Since this interval does not include 0, we say that 0 is not a candidate for the difference between the population means ($\mu_1-\mu_2$), and we conclude that $\mu_1-\mu_2\neq0$, i.e., $\mu_1\neq\mu_2$. Thus we arrive at the same conclusion by means of a confidence interval.

Another solution: by p-value:

2. $P-Value = 2\times P(Z>|Z_c|)$

$$= 2P(Z>2.57) = 2[1-P(Z<2.57)] = 2(1-0.9949) = 0.0102$$

The level of significance was $\alpha = 0.05$.
Since P-value < α, we reject H_o.

Reject H_o If p-value < α

0.0102< 0.05 ✓

Decision: Reject H_o

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:

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

Hypotheses:

H_o: $\mu_1 \le \mu_2$

H_A: $\mu_1 > \mu_2$

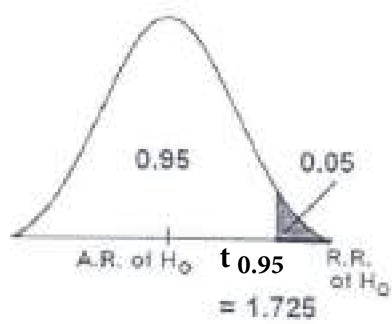
Or equivalently,

H_o: $\mu_1 - \mu_2 \le 0$

H_A: $\mu_1 - \mu_2 > 0$

Calculation:

$\alpha=0.05$



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

Reliability Coefficient:

$$df = v = 12 + 10 - 2 = 20$$

$\alpha = 0.05$ ------- $1-\alpha = 0.95$ -------- $t_{1-\alpha} = t_{0.95} = 1.725$

Test Statistic (T.S.):

$$T = \frac{\bar{X}_1 - \bar{X}_2 - M_o}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{85-81-0}{\sqrt{\frac{20.05}{12} + \frac{20.05}{10}}} = 2.09$$

Reject H_o

If T(test)> $t_{1-\alpha}$

2.09 > 1.725

Decision : Reject H_o

Decision:

Since T= 2.09 $\in$ R.R. (T= 2.09 > $t_{0.95}$ = 1.725), we reject H_o and we accept **H_A: $\mu_1 - \mu_2 > 0$ (H_A: $\mu_1 > \mu_2$)** at α =0.05. Therefore, we conclude that the mean abrasive wear of material 1 is greater than that of material 2.

7.4 Paired Comparisons:

Paired T-Test :

- 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 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: Values of the first population

Y: Values of the Second population

D: Values of X – Values of Y

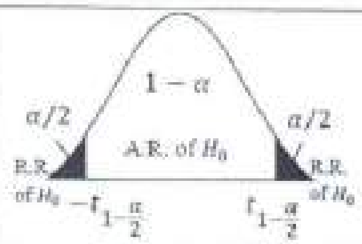
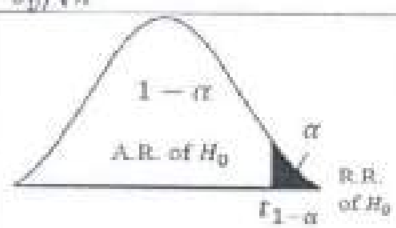
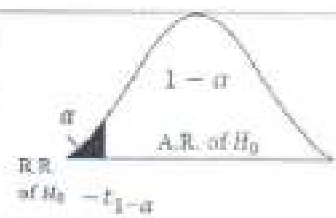
Means :

μ_1= Mean of the first population

μ_2= Mean of the Second population

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

Confident Interval and Testing Hypothesis about difference between two population means ($\mu_D = \mu_1 - \mu_2$) : (Dependent/Related population)

Calculate the following Quantities	• The difference (D-observation): $D_i = X_i - Y_i$, i=1,2,3,4,……,n • Sample mean of the D-Observations : $\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$ • Sample Variance $S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$ • Sample Standard Deviation $S_D = \sqrt{S_D^2}$		
Confident Interval for $\mu_D = \mu_1 - \mu_2$			
100(1-α)% Confident Interval for μ_D	$\bar{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$, $df = n - 1$		
Testing Hypothesis for $\mu_D = \mu_1 - \mu_2$			
Hypothesis	$H_0: \mu_1 - \mu_2 = M_O$ $H_A: \mu_1 - \mu_2 \neq M_O$ Or $H_0: \mu_D = M_O$ vs $H_A: \mu_D \neq M_O$	$H_0: \mu_1 - \mu_2 \leq M_O$ $H_A: \mu_1 - \mu_2 > M_O$ Or $H_0: \mu_D \leq M_O$ vs $H_A: \mu_D > M_O$	$H_0: \mu_1 - \mu_2 \geq M_O$ $H_A: \mu_1 - \mu_2 < M_O$ Or $H_0: \mu_D \geq M_O$ vs $H_A: \mu_D < M_O$
Test Statistic (T.S.)	$T = \frac{\bar{D} - M_O}{S_D / \sqrt{n}}$, $df = \nu = n - 1$		
Rejection Region(R.R) & Acceptance Region(A.R)			
Reliability Coefficient	$-t_{1-\alpha/2}$ or $t_{1-\alpha/2}$	$t_{1-\alpha}$	$-t_{1-\alpha}$
Decision : Reject H_0 if the Following condition satisfies	Reject H_0 (Accept H_A) at the significant level α if :		
	$T > t_{1-\alpha/2}$ Or $T < -t_{1-\alpha/2}$ (Two –sided test)	$T > t_{1-\alpha}$ (one – Sided Test)	$T < -t_{1-\alpha}$ (one – Sided Test)

Example:

Suppose that we are interested in studying the effectiveness of a certain diet program on ten individual . Let the random variables 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

Find :

1) A 95% Confident 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 the data provide sufficient evidence to allow us to conclude that the diet is good? Use $\alpha = 0.05$ and assume population is normal .

Solution :

1-st population (X) = the weight of the individual before the diet program.
2-nd population (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.

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

Calculate mean, standard deviation by calculator

$\bar{D}$, S_D

i	Xi	Yi	$D_i = Xi - Y_i$
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	$\Sigma x = 842$	$\Sigma y = 787.5$	$\Sigma D = 54.5$

<u>**First, we need to calculate :**</u>(By calculator)

Sample Mean:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{54.5}{10} = 5.45$$

Sample Variance :

$$S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1} = \frac{(6.9-5.45)^2 + (-5.7-5.45)^2 + \cdots\ldots\ldots + (12.4-5.45)^2}{10-1} = 50.33$$

Sample Standard Deviation : $S_D = \sqrt{S_D^2} = \sqrt{50.33} = 7.09$

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

α=0.05 ------ 1 – 0.05/2 =1-0.025= 0.975 (df =10 -1=9)

$t_{1-\alpha/2} = t_{0.975}$ =2.262

Then 95% Confident Interval for $\mu_D = \mu_1 - \mu_2$

$$\overline{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

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

$$5.45 \pm 5.0715$$

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

$$(0.38\,,\,10.52)$$

$$0.38 < \mu_D < 10.52$$

2)Does the data provide sufficient evidence to allow us to conclude that the diet is good? Use α =0.05 and assume population is normal .

Diet is good means --- weight after will be less than weight befor.

Solution:

μ_1= Mean of the first population

μ_2= Mean of the second population

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

Hypothesis :

$H_0: \mu_1 \leq \mu_2$ vs $H_A: \mu_1 > \mu_2$

or $H_0: \mu_1 - \mu_2 \leq 0$ vs $H_A: \mu_1 - \mu_2 > 0$

or $H_0: \mu_D \leq 0$ vs $H_A: \mu_D > 0$

Test Statistic:

$$\overline{D} = 5.45,\ S_D = 7.09\ , n = 10$$

$$T = \frac{\overline{D} - M_O}{\frac{S_D}{\sqrt{n}}} = \frac{5.45 - 0}{\frac{7.09}{\sqrt{10}}} = 2.43$$

Rejection Region(R.R):

α=0.05 ------ $1-\alpha = 0.95$ ----- $t_{1-\alpha} = t_{0.95} = 1.833$ (df =n-1 =9)

Reject H_0 if $T > t_{1-\alpha}$

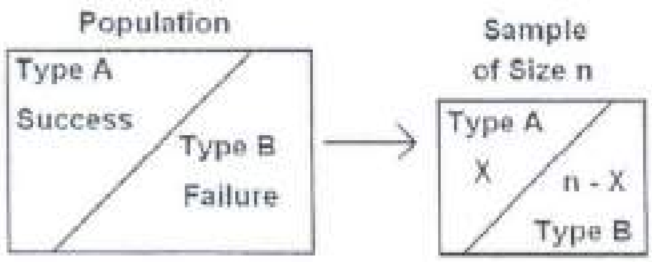
2.45 > 1.833 (condition satisfied)

Then reject H_0 and accept $H_A: \mu_1 > \mu_2$

So, we have a good diet program .

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

$$p = \frac{\textit{no. of elements of type A in the population}}{\textit{Total no. of elements in the population}}$$

$$p = \frac{A}{N} \qquad (N = \textit{population size})$$

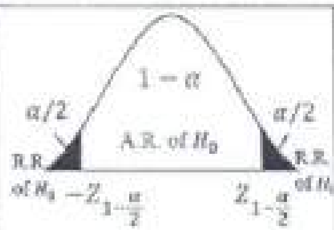
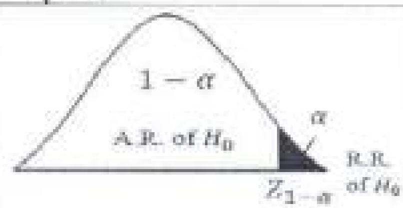
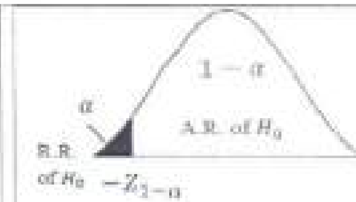
- n = sample size
- X = no. of elements of type A in the sample of size n.
- $\hat{p}$ = Sample proportion elements of Type A in the sample

$$\hat{p} = \frac{\textit{no. of elements of type A in the sample}}{\textit{no. of elements in the sample}}$$

$$\hat{p} = \frac{X}{n} \qquad \text{(n=sample size=no. of elements in the sample)}$$

- $\hat{p}$ is a "good" point estimate for p.
- For large n, ($n \geq 30$, $np > 5$), we have

Test Procedure:(P_0 is known number)

<table>
<tr><td>Hypothesis</td><td>$H_0: P = P_0$
$H_A: P \neq P_0$</td><td>$H_0: P \leq P_0$
$H_A: P > P_0$</td><td>$H_0: P \geq P_0$
$H_A: P < P_0$</td></tr>
<tr><td>Test Statistic (T.S.)</td><td colspan="3">$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \quad , \quad q_0 = 1 - p_0$$</td></tr>
<tr><td>Rejection Region(R.R) & Acceptance Region(A.R)</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Reliability Coefficient</td><td>$-Z_{1-\alpha/2}$ or $Z_{1-\alpha/2}$</td><td>$Z_{1-\alpha}$</td><td>$-Z_{1-\alpha}$</td></tr>
<tr><td rowspan="2">Decision:
Reject H_0 if the following condition satisfies</td><td colspan="3">Reject H_0 (Accept H_A) at the significant level α if :</td></tr>
<tr><td>$Z > Z_{1-\alpha/2}$
Or $Z < -Z_{1-\alpha/2}$</td><td>$Z > Z_{1-\alpha}$
(one – Sided Test)</td><td>$Z < -Z_{1-\alpha}$
(one – Sided Test)</td></tr>
</table>

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

$\bar{p}$ = proportion of females in the sample

$$\hat{p} = \frac{X}{n} = \frac{24}{45} = 0.5333$$

$$p_o = \frac{70}{100} = 0.7$$

$\alpha = 0.10$

Hypotheses:

$H_o: p = 0.7 \quad (p_o = 0.7)$

$H_A: p \neq 0.7$

Level of significance:

$\alpha = 0.10$ $\quad 1-\alpha/2 = 1-0.10/2 = 0.95$

Test Statistic (T.S.):

$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

$$= \frac{0.5333 - 0.70}{\sqrt{\frac{(0.7)(0.3)}{45}}} = -2.44$$

Rejection Region of H_o (R.R.):

Critical values: $\quad Z_{1-\alpha/2} = Z_{0.95} = 1.645$

We reject H_0 If:

$Z(test) > Z_{0.95}$ or $Z(test) < -Z_{0.95}$

$-2.44 > 1.645$ or $-2.44 < -1.645$

Since one of the conditions is valid then,

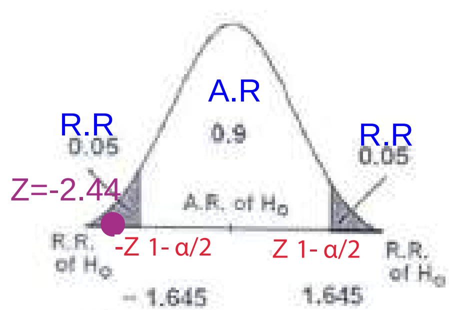
Decision : Reject H_0

From curve

Since Z(test)= -2.44 in R.R

Decision :

Reject H_0



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

Decision:

Since Z= −2.44 ∈Rejection Region of H_o (R.R), we reject

H_o:p=0.7 and accept H_A:$p \neq$0.7 at α=0.1. Therefore, we do not agree with the claim stating that 70% of the patients in this population are females.

Example:

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

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

p_0=0.25

α=0.025

Hypotheses:

H_o: $p \geq 0.25$ (p_o=0.25)

H_A: $p < 0.25$ (research hypothesis)

Level of significance:

α=0.025

Test Statistic (T.S.):

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.3 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{200}}} = 1.633$$

Rejection Region of H_o (R.R.):

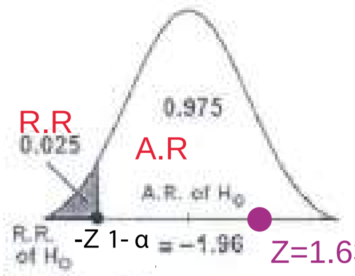
Critical value: $Z_{1-\alpha} = Z_{0.975} = 1.96$

Critical Region:

We reject H_o if:

$Z < -Z_{1-\alpha}$

$1.633 < -1.96$ Accept H_0 (condition not satisfy)



Another solution:
From curve :

Since Z(test)=1.633 in A.R

Decision:

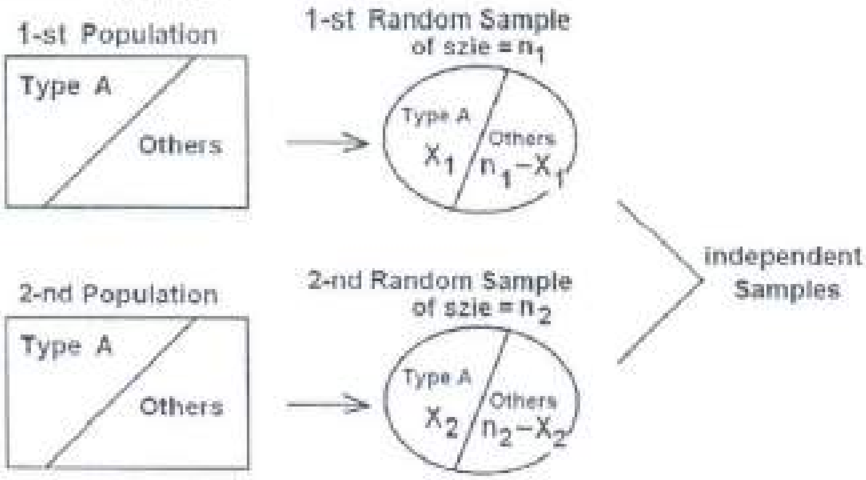
Accept H_0

Decision:

Since $Z=1.633 \in$ Acceptance Region of H_o (A.R.), we accept (do not reject) H_o: $p \geq 0.25$ and we reject H_A: $p < 0.25$ at $\alpha=0.025$. Therefore, we do not agree with claim stating that the proportion of adults in this city who hesitate to take dental appointment is less than 0.25.

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 2-nd sample
- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about $p_1 - p_2$.
- For large n_1 and n_2, we have

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

- $q = 1 - p$

Hypotheses:

We choose one of the following situations:

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

or equivalently,

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

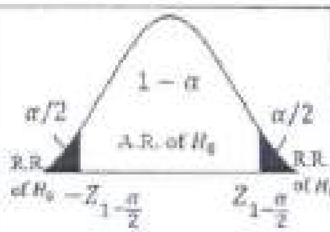
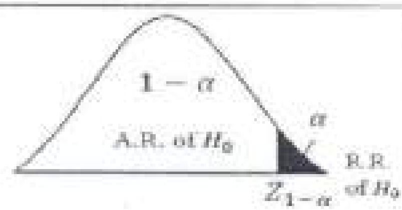
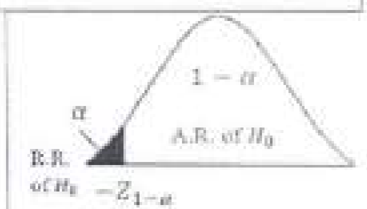
Note, under the assumption of the equality of the two population proportions (H_o: $p_1 = p_2 = p$), the pooled estimate of the common proportion p is:

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

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 Procedure:

Hypothesis	$H_0: P_1 - P_2 = 0$ $H_A: P_1 - P_2 \neq 0$	$H_0: P_1 - P_2 \leq 0$ $H_A: P_1 - P_2 > 0$	$H_0: P_1 - P_2 \geq 0$ $H_A: P_1 - P_2 < 0$
Test Statistic (T.S.)	$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\,\bar{q}}{n_1} + \frac{\bar{p}\,\bar{q}}{n_2}}}$, Pooled proportion: $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ where $\bar{q} = 1 - \bar{p}$		
Rejection Region(R.R) & Acceptance Region(A.R)			
Reliability Coefficient	$-Z_{1-\alpha/2}$ or $Z_{1-\alpha/2}$	$Z_{1-\alpha}$	$-Z_{1-\alpha}$
Decision : Reject H_0 if the following condition satisfies	Reject H_0 (Accept H_A) at the significant level α if :		
	$Z > Z_{1-\alpha/2}$ Or $Z < -Z_{1-\alpha/2}$	$Z > Z_{1-\alpha}$ (one – Sided Test)	$Z < -Z_{1-\alpha}$ (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 $\alpha = 0.05$ and assume that the two population proportions are equal.

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

Males	Females
$n_1 = 150$	$n_2 = 200$
$X_1 = 21$	$X_2 = 48$
$\hat{p}_1 = \frac{X_1}{n_1} = \frac{21}{150} = 0.14$	$\hat{p}_2 = \frac{X_2}{n_2} = \frac{48}{200} = 0.24$

The pooled estimate of the common proportion p is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{21+48}{150+200} = 0.197$$

Hypotheses:

$H_o: p_1 = p_2$
$H_A: p_1 \neq p_2$

or

$H_o: p_1 - p_2 = 0$
$H_A: p_1 - p_2 \neq 0$

Level of significance: $\alpha = 0.05$ $\quad 1-\alpha/2 = 1 - 0.05/2 = 0.975$

Test Statistic (T.S.):

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.14-0.24)}{\sqrt{\frac{0.197\times 0.803}{150} + \frac{0.197\times 0.803}{200}}} = -2.328$$

Rejection Region (R.R.) of H_o:

Critical values:

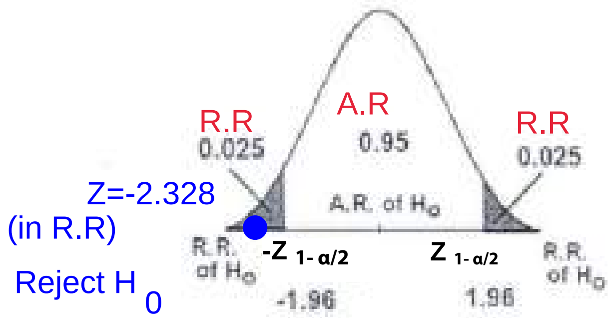
$$Z_{1-\alpha/2} = Z_{0.975} = 1.96$$

Critical region:

Reject H_o if: $Z_{test} < -1.96$ or $Z_{test} > 1.96$

$-2.328 < -1.96$ ✓

Decision: Reject H_0 (Since one of the conditions satisfied)



Another solution:
From curve

Since Z(test) in R.R

Decision:
Reject H 0

Decision:

Since Z= −2.328 ∈R.R., we reject H_0: $p_1 = p_2$ and accept H_A: $p_1 \neq p_2$ at α=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

TABLE OF CONTENTS

TABLE OF CONTENTS