manipalglobal
Academy of Data Science

Statistical Techniques for Data Science

Testing of Hypothesis

Parametric Tests

Dr. Subhabaha Pal
Manipal Global Academy of Data Science

Objective

manipalglobal
Academy of Data Science

After attending this session, you will be able to –

- **Describe Large sample test and Small sample test**
- **Describe Student’s t-test**
- **Describe F-test**
- **Describe Z-test**
- **Describe ANOVA**

Different Parametric Tests

- If the sample size for the test is more than 30, then the test is termed as large sample test, otherwise, it is termed as small sample test
- The major assumptions in large sample tests are as follows –
  - The random sampling distribution is approximately normal
  - Population standard deviation/variance is known
  - Values given by the samples are sufficiently close to the population value and can be used in its place for calculating the standard error of the estimate
- Large Sample Test ($n > 30$) Example - Standard Normal Test (Z-Test)
- Small Sample Test ($n \leq 30$) Example – Students' t-test (Paired and Unpaired t-test), Analysis of Variance (ANOVA), Repeated Measures of ANOVA

Large Sample Test: Z-Test

manipalglobal
Academy of Data Science

- Z-test is the large sample test
- Z-test is called standard normal variate test
- The major assumption here is that the corresponding populations whose parameters are to be tested follow normal distribution
- This test can be used for testing –
    - Mean of a single population (μ)
    - Difference between means of two populations ($\mu_1 - \mu_2$)
    - Proportion of a single population (P)
    - Difference between proportions of two populations ($P_1 - P_2$)
- Z-test can be applied when the population standard deviation/variance is known

Z- Test : Testing the Mean of Population

manipalglobal
Academy of Data Science

- The steps involved in testing the mean of population in Z-test are as follows –

1. Step 1 – Stating the Null and Alternative Hypothesis

   $H_0: \mu = \mu_0$

   versus

   (i) $H_1: \mu < \mu_0$

   or

   (ii) $H_1: \mu > \mu_0$

   or

   (iii) $H_1: \mu \neq \mu_0$

2. Step 2 – The distribution of the population whose parameter being tested to be assumed as normal

3. Step 3 – Need to fix level of significance $\alpha = 0.05$, i.e., if the p-value comes less than 0.05, to reject the null hypothesis

Steps in Testing the Mean of Population

manipalglobal
Academy of Data Science

4. **Step 4 – The test statistic needs to be computed –**

$$Z = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} \approx N(0,1)$$

5. **Step 5 - Define the rejection criteria/critical region in case of different alternative hypotheses:**

(i) H_1: $\mu < \mu_0$

or

(ii) H_1: $\mu > \mu_0$

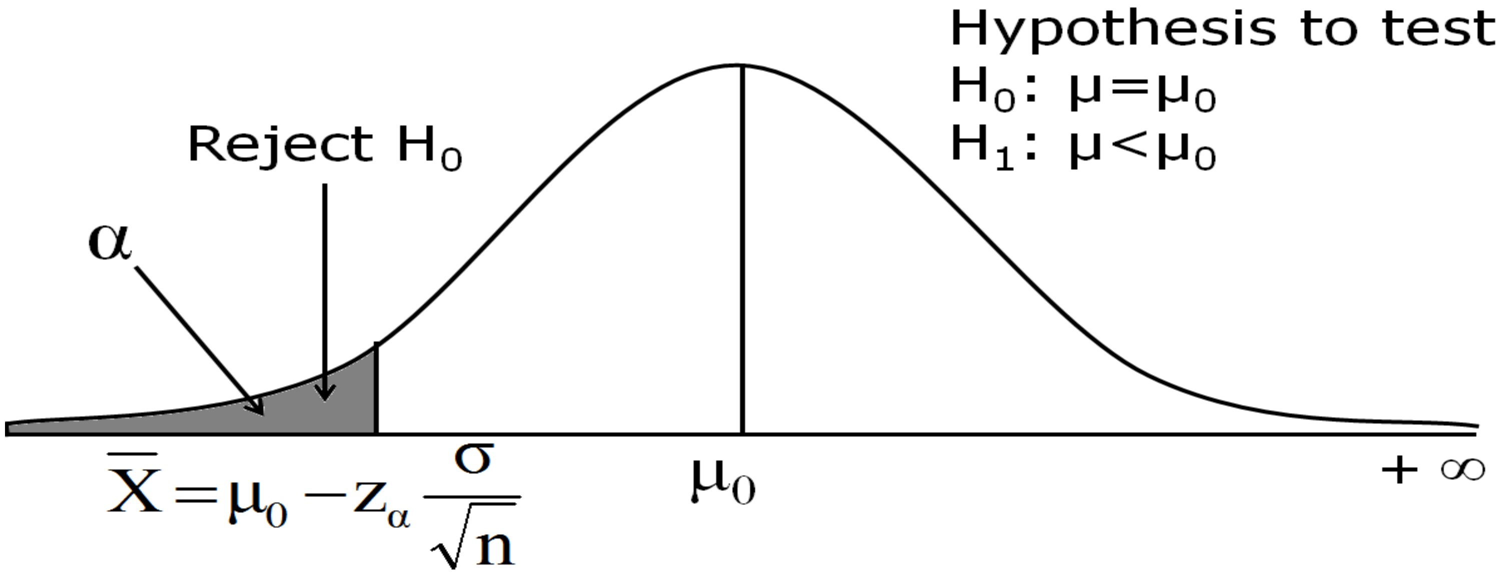
or

(iii) H_1: $\mu \neq \mu_0$

Steps in Testing the Mean of Population

5. Step 5 – When alternative hypothesis is H_1: $\mu < \mu_0$

Reject H_0 if the computed value of the test statistic Z is less than the critical value, i.e., $P(Z < -z_\alpha) < 0.05$, otherwise, accept the null hypothesis

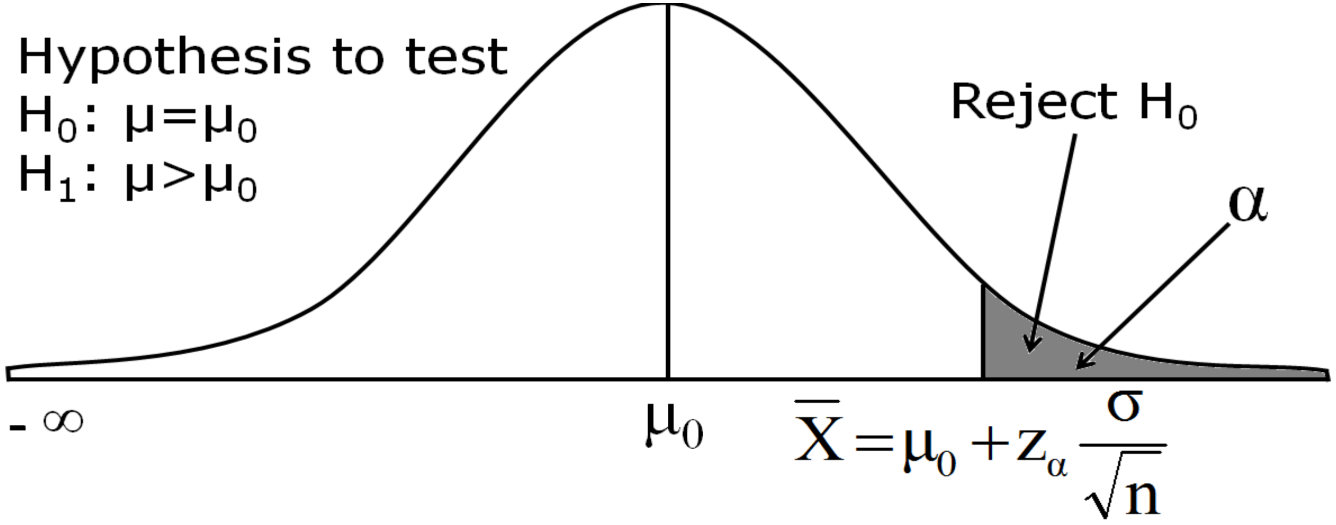


6. Step 6 – Conclusion of the test to be written based on result

Steps in Testing the Mean of Population

5. Step 5 – When alternative hypothesis is H_1: $\mu > \mu_0$

Reject H_0 if the computed value of the test statistic Z is greater than the critical value, i.e., $P(Z > z_\alpha) < 0.05$, otherwise, accept the null hypothesis



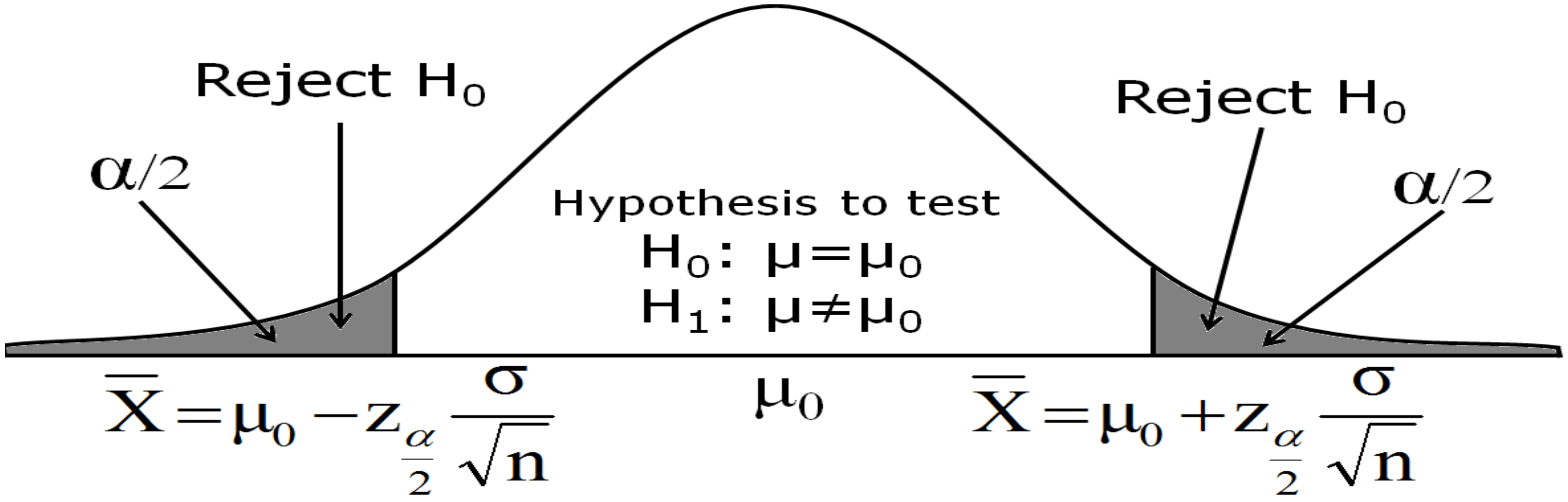
6. Step 6 – Conclusion of the test to be written based on result

Steps in Testing the Mean of Population

5. Step 5 – When alternative hypothesis is H_1: $\mu \neq \mu_0$

Reject H_0 if the computed value of the test statistic Z is less than or greater than the critical value, i.e.,

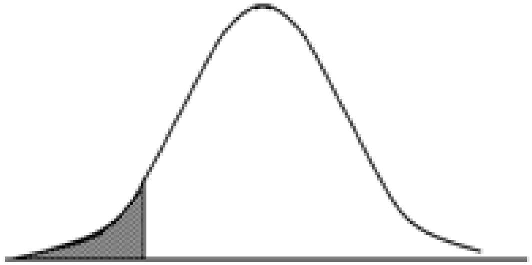
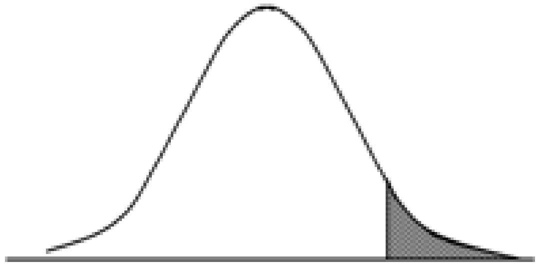
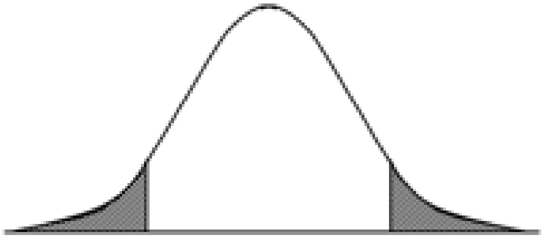
$P(Z < -z_{\alpha/2}) + P(Z > z_{\alpha/2}) < 0.05$, otherwise, accept the null hypothesis



6. Step 6 – Conclusion of the test to be written based on result

Testing of Mean of Population

Summary of One- and Two-Tail Tests…

One-Tail Test (left tail)	One-Tail Test (right tail)	Two-Tail Test
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$
		

Testing the Difference of Means of Two Populations

- **The Assumptions for this test are the following –**

1. **The samples are drawn from normally distributed population**
2. **The population variances should be known**
3. **The sample units from both populations should be drawn randomly**
4. **The two samples should be independent**
5. **The sample size should be more than 30 (i.e., n > 30)**

- **The null and alternative hypotheses are as follows –**

$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 < \mu_2$

Or

$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 > \mu_2$

Or

$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

Steps for Testing the Difference of Means of Two Populations

- **Step 1 – Null and alternative hypotheses to be stated**
- **Step 2 - Level of Significance to be determined, i.e., α = 0.05**
- **Step 3 – The population of the samples are assumed to follow normality**
- **Step 4 – The test statistic to be calculated as follows –**

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} \approx N(0,1)$$

Steps for Testing the Difference of Means of Two Populations

manipalglobal
Academy of Data Science

- **Step 5 – The rejection criteria/critical region needs to be defined –**

i. H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 < \mu_2$

Reject H_0 if computed value of Z is less than the critical value, i.e., $P(Z < -z_\alpha) < 0.05$, otherwise do not reject H_0

ii. H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 > \mu_2$

Reject H_0 if computed value of Z is greater than the critical value, i.e., $P(Z > z_\alpha) < 0.05$, otherwise do not reject H_0

iii. H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 \neq \mu_2$

Reject H_0 if computed value of Z is less than or greater than the critical value, i.e., $P(Z < - z_{\alpha/2}) + P(Z > z_{\alpha/2}) < 0.05$, otherwise do not reject H_0

Example on Z-test

A sample of 100 lung cancer patients on a new drug (Group I) are observed to have a mean survival of 27.5 months, in another sample of 100 lung cancer patients with old drug (Group II), the mean survival was observed to be 25.2 months. The Population variances are 6.25 months and 9.1 months respectively.

The investigators want to know on the basis of the data whether the new drug prolongs the survival?

Calculation of Z-statistic

- **Sample mean of Group I: 27.5**
- **Sample mean of Group II: 26.2**
- **Population variance of Group I: 6.25**
- **Population variance of Group II: 9.10**
- **Sample size $n_1 = 100$, $n_2 = 100$**

- $H_o: \mu_1 = \mu_2$
- $H_1: \mu_1 > \mu_2$
- Normal distribution

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} = \frac{27.5 - 26.2}{\sqrt{\frac{6.25}{100} + \frac{9.10}{100}}} = 3.316$$

- $Z_{0.05} = 1.645$
- 95% CI for difference = 0.516, 2.084
- P ??? (From Z table)
- Decision:

Testing Proportion of Single Population

- Samples are to be drawn from normal population
- Steps in Testing Process –

Step 1 – Null and Alternative hypotheses to be stated –

H_0: $P = P_0$ Vs. (i) H_1: $P < P_0$, (ii) H_1: $P > P_0$, (iii) H_1: $P \neq P_0$,

Step 2 – The level of significance α to be specified, i.e., α = 0.05

Step 3 – The underlying distribution is supposed to follow normality

Step 4 – The test statistic is as follows –

$$Z = \frac{p - P_0}{\left(\sqrt{\frac{pq}{n}}\right)} \approx N(0,1)$$

Step 5 – The testing rejection criteria is same as before as in means

Testing Difference between Proportions of Two Populations

manipalglobal
Academy of Data Science

- **The assumptions for this test are as follows –**
  - **The samples are to be drawn from normal population**
  - **Population variances are known**
- **Steps in testing difference between proportions of 2 populations**
  - **Step 1 – State the null and alternative hypotheses –**

H_0: $P_1 = P_2$ vs H_1: $P_1 < P_2$

Or

H_0: $P_1 = P_2$ vs H_1: $P_1 > P_2$

Or

H_0: $P_1 = P_2$ vs H_1: $P_1 \neq P_2$

  - **Step 2 – Test statistic needs to be calculated –**

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx N(0,1)$$

, where, $p = \frac{x_1 + x_2}{n_1 + n_2}, q = 1 - p$

  - **Step 3 – The testing criteria needs to be specified and test to be performed**

Small Sample Test

- Small sample test is generally undertaken under the following circumstances –
  - The sample size n is less than 30
  - Population is assumed to be distributed normally
  - Population variance is to be estimated from the sample data

- Assumptions for small sample test are the following –
  - Sample units are drawn from normally distributed population
  - Population variances are not known
  - The sample size n is less than 30
  - Samples should be allocated normally

Small Sample Test

- Generally the sample size should be less than 30 (n < 30) and even if the sample size is more than 30 (n > 30), but the population variances are unknown, the Student's t-test can be applied
- This is because due to Central Limit Theorem (CLT) for n > 30, the data converges to Standard Normal Distribution (SND)
- The following are the Small Sample tests –
  - Independent Sample t-test
  - Unpaired t-test
  - Paired t-test

Independent Sample t-Test

- Here we test the mean of a single population
- We assume that the samples are drawn from normal distribution
- We also assume that population variance is unknown

Steps involved in testing mean of a population

1. Stating null and alternative hypothesis

$H_0: \mu = \mu_0$

versus

(i) $H_1: \mu < \mu_0$

or

(ii) $H_1: \mu > \mu_0$

or

(iii) $H_1: \mu \neq \mu_0$

Steps involved in testing mean of a population

manipalglobal Academy of Data Science

2. The level of significance α = 0.05

3. Student's t-distribution

4. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{\left(\frac{S}{\sqrt{n}}\right)} \approx t_{(\alpha;n-1)}$$

5. Define the rejection criteria/ critical regional

(i) Reject H_0 if computed value of $|t|$ is greater than the critical value, ie., $P(|t| > t_{\alpha})$, otherwise do not reject H_0 for one-tailed test

(ii) Reject H_0 if computed value of $|t|$ is greater than the critical value, ie., $P(|t| > t_{\alpha/2})$, otherwise do not reject H_0 for two-tailed test

6. Conclusion

Steps involved in testing mean difference of two population

1. Stating null and alternative hypothesis

- $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 < \mu_2$

Or

- $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 > \mu_2$

Or

- $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

manipalglobal Academy of Data Science

Steps involved in testing mean difference of two population

2. The level of significance α = 0.05

3. Student's t-distribution

4. The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{(\alpha;\, n_1+n_2-2)}$$

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

5. Define the rejection criteria/ critical regional

(i) Reject H_0 if computed value of $|t|$ is greater than the critical value, ie., $P(|t| > t_{\alpha})$, otherwise do not reject H_0 for one-tailed test

(ii) Reject H_0 if computed value of $|t|$ is greater than the critical value, ie., $P(|t| > t_{\alpha/2})$, otherwise do not reject H_0 for two-tailed test

6. Conclusion

Examples on unpaired t-test

A sample of 15 lung cancer patients on a new drug (Group I) are observed to have a mean survival of 27.5 months, in another sample of 15 lung cancer patients with old drug (Group II), the mean survival was observed to be 25.2 months. The sample variances are 6.25 months and 9.1 months respectively.

The investigators want to know on the basis of the data whether the new drug prolongs the survival?

Calculation of t-test

- Sample mean of Group I: 27.5
- Sample mean of Group II: 26.2
- Sample variance of Group I: 6.25
- Sample variance of Group II: 9.10
- Sample size $n_1 = 15$, $n_2 = 15$

- $H_o: \mu_1 = \mu_2$
- $H_1: \mu_1 > \mu_2$
- Student's t-distribution

$$t = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{27.5 - 26.2}{\sqrt{\frac{6.25}{15} + \frac{9.10}{15}}} = 1.287$$

- t= 1.701, for 28 degrees of freedom
- P ??? (From t table)
- Decision:

Example 2

A sample of 30 lung cancer patients on a new drug (Group I) are observed to have a mean survival of 27.5 months, in another sample of 30 lung cancer patients with old drug (Group II), the mean survival was observed to be 25.2 months. The sample variances are 6.25 months and 9.1 months respectively.

The investigators want to know on the basis of the data whether the new drug prolongs the survival?

Calculation of t-test

- Sample mean of Group I: 27.5
- Sample mean of Group II: 26.2
- Sample variance of Group I: 6.25
- Sample variance of Group II: 9.10
- Sample size $n_1 = 30$, $n_2 = 30$

- $H_o: \mu_1 = \mu_2$
- $H_1: \mu_1 > \mu_2$
- Student's t-distribution

$$t = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}} = \frac{27.5 - 26.2}{\sqrt{\frac{6.25}{30} + \frac{9.10}{30}}} = 1.818$$

- t= 1.671 for 58 degrees of freedom
- P ??? (From t table)
- Decision:

Example 3

A sample of 50 lung cancer patients on a new drug (Group I) are observed to have a mean survival of 27.5 months, in another sample of 50 lung cancer patients with old drug (Group II), the mean survival was observed to be 25.2 months. The sample variances are 6.25 months and 9.1 months respectively.

The investigators want to know on the basis of the data whether the new drug prolongs the survival?

Calculation of t-test

- Sample mean of Group I: 27.5
- Sample mean of Group II: 26.2
- Population variance of Group I: 6.25
- Population variance of Group II: 9.10
- Sample size $n_1 = 50$, $n_2 = 50$

- H_o: $\mu_1 = \mu_2$
- H_1: $\mu_1 > \mu_2$
- Student's t-distribution

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}} = \frac{27.5 - 26.2}{\sqrt{\frac{6.25}{50} + \frac{9.10}{50}}} = 2.347$$

- t= 1.658 ???, for 98 degrees of freedom
- P ??? (From t table)
- Decision:

Paired sample t-test

Observations recorded on the same set of subjects at two different time intervals

Eg. Before and after treatment or Pre-test and Post-test

Assumptions on Paired t-test

- The difference between before and after observation should be normally distributed.
- The Samples should be selected randomly

Steps involved in testing mean of paired samples

1. Stating null and alternative hypothesis

$$H_0: \mu = \mu_d = 0$$

versus

(i) $H_1: \mu < \mu_d$

or

(ii) $H_1: \mu > \mu_d$

or

(iii) $H_1: \mu \neq \mu_d$

Steps involved in testing mean of a population

2. The level of significance α = 0.05

3. Student's t-distribution

4. The test statistic is $t = \dfrac{\text{Difference b'n before and after observations}}{\text{SE (Diff. b'n before and after observations)}}$

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \approx t_{(\alpha,\, n-1)}$$

5. Define the rejection criteria/ critical regional

(i) Reject H_0 if computed value of $|t|$ is greater than the critical value, ie., $P(|t| > t_\alpha)$, otherwise do not reject H_0 for one-tailed test

(ii) Reject H_0 if computed value of $|t|$ is greater than the critical value, i.e., $P(|t| > t_{\alpha/2})$, otherwise do not reject H_0 for two-tailed test

6. Conclusion

Analysis of Variance (ANOVA)

manipalglobal Academy of Data Science

- ►Samples should have been drawn from populations which are normally distributed
- ►Used to test equality of more than two population means

ANALYSIS OF VARIANCE

When there are more than two groups to be compared, it is not correct to compare the groups in pairs, as this type of comparison will not take the within variability into consideration

The Analysis procedure used in such comparisons is known as ANALYSIS OF VARIANCE

In this analysis the variability of observations between and within groups are taken into consideration

The total variability is split in these components and test is applied

The test used to compare these variability is F test

Assumptions on ANOVA

manipalglobal
Academy of Data Science

- ♠ Assume that the samples are drawn from normal distribution
- ♠ The population variances should be equal
- ♠ The samples should be selected randomly
- ♠ The groups should be independent

Assumptions on ANOVA

- **The data should be linearly additive**
- The sample size should be more than 30

One-way Analysis of variance (1-way ANOVA)

manipalglobal Academy of Data Science

Iron intake of four groups of patients in mg

Group 1	Group 2	Group 3	Group 4
11.5	19.5	18.5	30.0
12.5	18.5	16.5	26.5
18.5	16.0	24.5	27.0
21.0	22.0	30.0	34.0
28.0	30.0	28.5	20.0
26.0	24.5	14.0	22.5
14.0	19.0	19.0	28.0
22.0	24.0	17.0	32.0
20.0	19.5	18.0	27.0
22.0	15.0	29.0	25.5

One-way ANOVA table

Null hypothesis: No difference in the mean iron intake of four groups
To test this hypothesis One way Analysis of variance is used
The out put of the analysis will be as follows

Analysis of variance table

Source of Sum of squares	Degree of freedom	Sum of squares	Mean sum of squares	F
Between Groups	3	349.525	116.5083	4.544
Within Groups	36	922.950	25.6375	
Total	39	1272.475		

From the Variance ratio table, F value corresponding to the (3, 36) degrees of freedom is 4.38, for $P = 0.01$.
Conclusion: Iron intake between the Groups are significantly different

Post-hoc test for ANOVA

There are also other test for testing these differences between the groups

- **Bonferroni's test**
- **Scheffe's test,**
- **Tuckey's test**
- **Duncun's test**
- **Least significant difference test (LSD) etc.**

Post-hoc test for ANOVA

Analysis of Variance provides estimate of Standard error for testing which of the differences between the villages is significant. An estimate of the standard error of the differences between the group means is equal to $\sqrt{\dfrac{2S^2}{k}}$

Where S^2 is the 'Within groups mean sum of squares' and k is the number of observations in each of the group under comparison

S^2 = Within Groups Mean sum of squares
= 25.6375
k = Number of observations in each Group = 10

df of within villages = 36

t Statistics value corresponding to P = 0.05 for 36 df is 2.03

Least significant difference (LSD) $\left\{ t_{0.05}\sqrt{\frac{2S^2}{k}} \right\}$

$$\left\{2.03\sqrt{\frac{225.64}{10}}\right\} = 4.597$$

Group of patients			
1	2	3	4
Mean iron intake (mg)			
20	21	22	28

It can be seen that only the difference between Group 4 is different from other groups as only this difference is more than 4.597.

Two way Analysis

If there more than sub comparisons are to be made two-way analysis is applied

To test whether there is any significant difference in the man iron intake between Groups and also between trimesters of pregnancy

Ante-period (Trimester)	Group means for iron intake									
	1	2	3	4	5	6	7	8	9	10
I	11.5	19.5	18.5	12.5	18.5	16.5	26.5	18.5	16.0	24.5
II	27.0	28.0	22.0	21.0	15.0	19.5	20.0	26.0	30.0	28.5
III	28.0	3 0.0	26.0	30.0	24.5	28.5	26.0	30.0	27.0	25.5

Two-way Analysis of variance (2-way ANOVA)

Two way Analysis

If there more than sub comparisons are to be made two-way analysis is applied

To test whether there is any significant difference in the man iron intake between Groups and also between trimesters of pregnancy

Ante-period (Trimester)	Group means for iron intake									
	1	2	3	4	5	6	7	8	9	10
I	11.5	19.5	18.5	12.5	18.5	16.5	26.5	18.5	16.0	24.5
II	27.0	28.0	22.0	21.0	15.0	19.5	20.0	26.0	30.0	28.5
III	28.0	3 0.0	26.0	30.0	24.5	28.5	26.0	30.0	27.0	25.5

From the table of F values at P = 0.05, it is seen that F value for (9,18) df corresponding to 'Between Groups' is 2.46, while the calculated value is 0. 85 which is less than the Table value.

Table value of F at P = 0.05 for (2,18) d.f. corresponding to 'Between Trimesters' is 3.16, while the corresponding calculated value is 12.526 which is more than the table value.

It can be concluded that the differences in the mean iron intake of antenatal mothers between the Groups are not significantly different while the differences in the mean iron intake between the trimesters of pregnancy are significantly different.

Further analysis to look into the difference within the sub Group means can be done using least significant differences or other tests.

In many experiments, the outcome of a variable depends on the magnitude of the variable before subjecting the experimental units for experimentation.

As such, it may be necessary to analyse the outcome values in relation to initial values.

In some other cases, the outcome of a particular variable may be dependent on the outcome of another variable.

In such cases also it is desired to analyse the significance of the effect of this variable on the outcome of the experimental variable.

Analysis of co-variance is a technique that enables such analysis. This technique combines features of Analysis of variance and Regression analysis.

Example: Analysis of variance

A study on serum alkaline phosphatase activity levels in children with seizure disorders who were receiving anticonvulsant therapy under the care of a Paediatrician are as follows:

G1	G2	G3	G4
49.20	97.50	97.07	110.60
44.54	105.00	73.40	52.10
45.80	58.05	68.50	117.60
95.84	86.60	91.85	77.71
30.10	58.35	106.60	150.00
36.50	72.80	0.57	82.90
82.30	116.70	0.79	111.15
87.50	45.15	0.77	
105.00	70.35	0.81	
95.22	77.40		

G1	G2	G3	G4
49.20	97.50	97.07	110.60
44.54	105.00	73.40	52.10
45.80	58.05	68.50	117.60
95.84	86.60	91.85	77.71
30.10	58.35	106.60	150.00
36.50	72.80	0.57	82.90
82.30	116.70	0.79	111.15
87.50	45.15	0.77	
105.00	70.35	0.81	
95.22	77.40		

manipalglobal
Academy of Data Science

Copyright Manipal Global Education Services Pvt. Ltd. All Rights Reserved.

All product and company names used or referred to in this work are trademarks or registered trademarks of their respective holders.
Use of them in this work does not imply any affiliation with or endorsement by them.
This work contains a variety of copyrighted material. Some of this is the intellectual property of Manipal Global Education, some material is owned by others which is clearly indicated, and other material may be in the public domain. Except for material which is unambiguously and unarguably in the public domain, permission is not given for any commercial use or sale of this work or any portion or component hereof. No part of this work (except as legally allowed for private use and study) may be reproduced, adapted, or further disseminated without the express and written permission of Manipal Global Education or the legal holder of copyright, as the case may be.