

STATISTICAL METHODS FOR DECISION MAKING

An Internship Project Report

Submitted in partial fulfillment of the requirement for the award of the
Degree of Bachelor of Science in Mathematics

Submitted by

M.HEMALATHA

Register no: 18BM7450

Under the guidance of

Ms. S.Loganayaki M.Sc., M. Phil.,

Assistant Professor

Department of Mathematics (Self Finance)



SRI G.V.G VISALAKSHI COLLEGE FOR WOMEN (AUTONOMOUS)

(Affiliated to Bharathiyar University, Coimbatore)

ACCREDITED AT 'A+' GRADE WITH CGPA 3.27 BY NAAC

AN ISO 9001:2008 CERTIFIED INSTITUTION

UDUMALPET-642 128

March-2021

CERTIFICATE:

This is to certify that the Internship project work entitled “**STATISTICAL METHODS FOR DECISION MAKING**” is a bonafied record work done by **M.HEMALATHA (18BM7450)** submitted in partial fulfillment of the requirements for the award of the Degree of Bachelor of Science in Mathematics at Sri G.V.G Visalakshi college for women (Autonomous), Udumalpet during the academic year 2020-2021.

Signature of the Guide

Signature of the HOD

E-CERTIFICATE



INTRODUCTION:

The Internship training program was organized in **GREAT LEARNING** application launched by Mohan lakhamraju in 2013. **Great Learning** is one of India's leading ed-tech companies for professional and higher education. This is the no. 1 ranked online classroom course on Artificial Intelligence, Machine Learning, Data Science Engineering and Deep learning for college students professionally. The trainer Dr.Abhinanda Sarkar taught us about “**STATISTICAL METHODS FOR DECISION MAKING.**” In this he tells about sampling, normal distribution, hypothesis testing, chi-square test and Anova.

DIGITAL TOOL:



Great Learning
- Free Courses
Online &
Certificate
Great Learning

Uninstall

Update

Dr. Abhinanda Sarkar Professor of Statistics gave us training under the topic “STATISTICAL METHODS FOR DECISION MAKING.”

1. Sampling
2. Normal distribution
3. Hypothesis testing
4. Type 1 and Type 2 errors
5. Types of hypothesis tests
6. Confidence intervals
7. Examples of Hypothesis testing
8. Chi-Square test
9. ANOVA

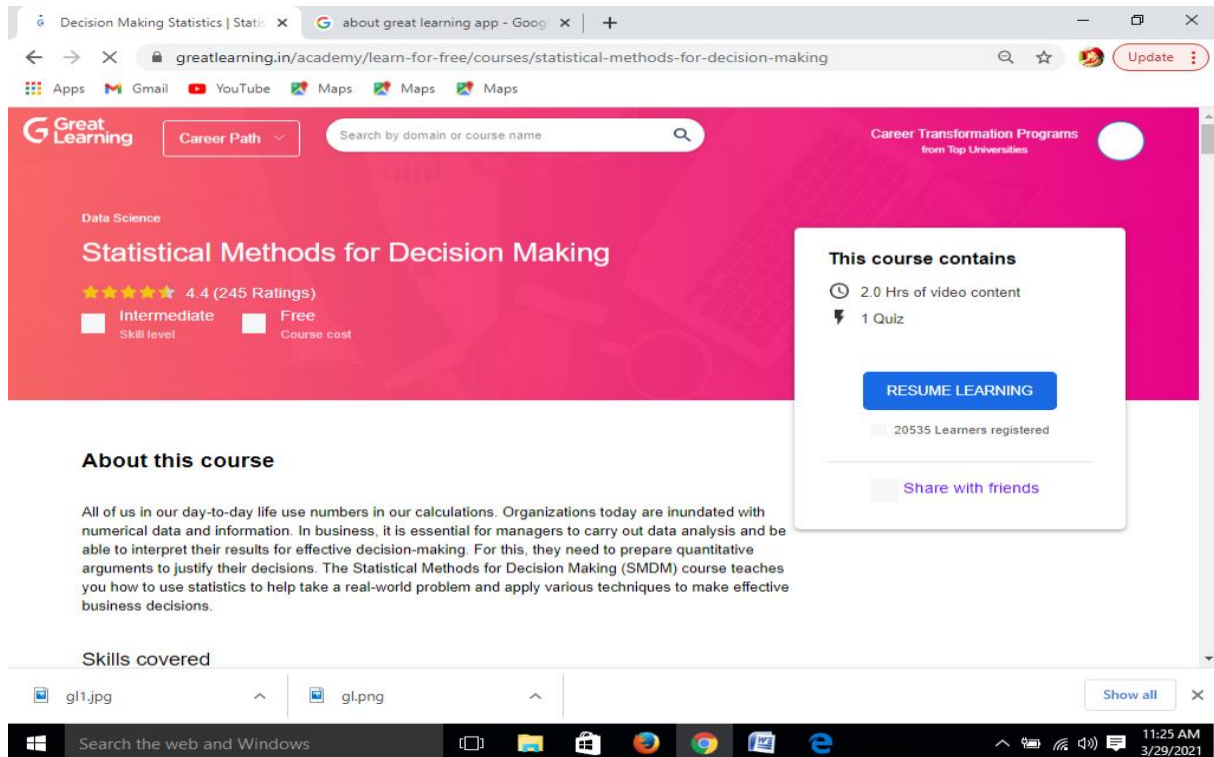
PG Programs Overview

10. Data Science and Business Analytic (DSBA)

11. Artificial Intelligence & Machine Learning

Uses of Statistical methods for decision making:

- ❖ **Statistics** can also aid the **decision making process** by enabling us to establish numerical benchmarks and monitor and evaluate the progress of our policy or program.
- ❖ **Statistics** can be used to inform **decision making** throughout the different stages of the **policy-making process**.



The screenshot displays a web browser window with the Great Learning website. The page features a pink header with the Great Learning logo, a search bar, and a dropdown menu for 'Career Path'. The main content area is titled 'Statistical Methods for Decision Making' under the 'Data Science' category. It includes a 4.4 rating from 245 reviews, an 'Intermediate' skill level, and a 'Free' course cost. A 'This course contains' box lists '2.0 Hrs of video content' and '1 Quiz', with a 'RESUME LEARNING' button and '20535 Learners registered'. Below this, there is a 'Share with friends' button. The 'About this course' section describes the importance of statistics in business decision-making. The 'Skills covered' section is partially visible at the bottom. The browser's taskbar shows the Windows logo, search bar, and various application icons, with the system tray displaying the time as 11:25 AM on 3/29/2021.

1. AGENDA

1. Sampling
2. Normal distribution
3. Hypothesis testing
4. Confidence interval
5. Chi-Square Test
6. ANOVA

The screenshot shows a web browser window with the URL olympus.greatlearning.in/courses/10913/pages/agenda?module_item_id=851734. The page features a green header with the Great Learning logo and a call to action: "Download Great Learning App Learn new skills anywhere anytime". Below the header, the "Agenda" section is displayed on a dark blue background with the Great Learning logo and tagline "Power Ahead". The agenda items are numbered 1 through 5, with the 6th item missing:

- 1. Sampling and Normal Distribution
- 2. Hypothesis Testing
- 3. Confidence Intervals
- 4. Chi-Square Test
- 5. ANOVA

On the right side, there is a blue sidebar with the text: "Become Job-Ready in Data Science PG Program in Data Science and Engineering" and a "KNOW MORE" button. The Windows taskbar at the bottom shows the time as 12:49 PM on 3/29/2021.

2. SAMPLING:

Need for sampling as opposed to using the entire population for analysis:

- Time factor
- Effort factor

It is usually not feasible to make a complete census of a population because of time and budget constraints. Therefore, a sample of the population is used to make inferences about the whole population. The goal of this type of sampling is to collect data that are representative of the entire population of interest.

Sampling

greatlearning
Power Ahead

Need for sampling as opposed to using the entire population for analysis:

- Time factor
- Effort factor

It is usually not feasible to make a complete census of a population because of time and budget constraints. Therefore, a sample of the population is used to make inferences about the whole population. The goal of this type of sampling is to collect data that are representative of the entire population of interest.

Become Job-Ready in Data Science
PG Program in Data Science and Engineering
[KNOW MORE](#)

3. NORMAL DISTRIBUTION-2

Central Limit Theorem

- Sampling Distribution of the mean of any independent random variable will be normal.
- This applies to both discrete and continuous distribution.
- The random variable should have a well defined mean and variance (standard deviation).
- Application even when the original variable is not normally distributed.

The screenshot displays a web browser window with the following elements:

- Browser Tabs:** "Normal Distribution-2: Statistical" and "Inbox (654) - subha482001@gm...".
- Address Bar:** "olympus.greatlearning.in/courses/10913/modules/items/851755".
- Navigation Bar:** Includes the Great Learning logo, a "Download Great Learning App" button with the text "Learn new skills anywhere anytime", and a user profile icon.
- Video Player:** The main content area shows a video titled "Normal Distribution-2" with the following text:

Central Limit Theorem

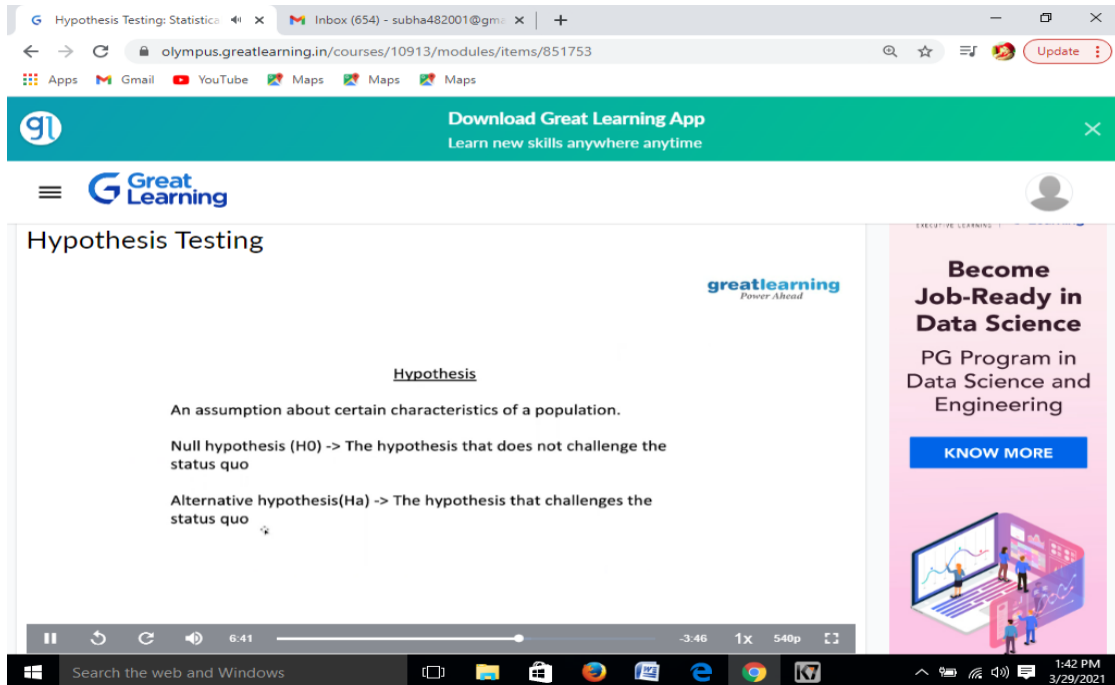
 - "Sampling Distribution of the mean of any independent random variable will be normal"
 - This applies to both discrete and continuous distributions.
 - The random variable should have a well defined mean and variance (standard deviation).
 - Applicable even when the original variable is not normally distributed.
- Right Sidebar:** A promotional banner for "Become Job-Ready in Data Science" featuring a "PG Program in Data Science and Engineering" and a "KNOW MORE" button.
- Taskbar:** Shows the Windows search bar, taskbar icons for various applications, and system tray information including the time "1:37 PM" and date "3/29/2021".

4. HYPOTHESIS TESTING:

An assumption about certain characteristics of a population.

Null hypothesis (H_0)-> The hypothesis that does not challenge the status quo

Alternative hypothesis (H_a)->The hypothesis that challenge the status quo



The screenshot shows a video player interface. The video content is a slide titled "Hypothesis Testing" with the following text:

Hypothesis

An assumption about certain characteristics of a population.

Null hypothesis (H_0) -> The hypothesis that does not challenge the status quo

Alternative hypothesis (H_a) -> The hypothesis that challenges the status quo

The slide also features the Great Learning logo and a sidebar advertisement for a "PG Program in Data Science and Engineering". The video player controls at the bottom show a progress bar at 6:41, a volume icon, and a play button.

5. TYPE 1 AND TYPE 2 ERRORS:

Type I Errors :

- Rejection of null hypothesis when it should not have been rejected

- Incorrectly rejecting the null hypothesis.

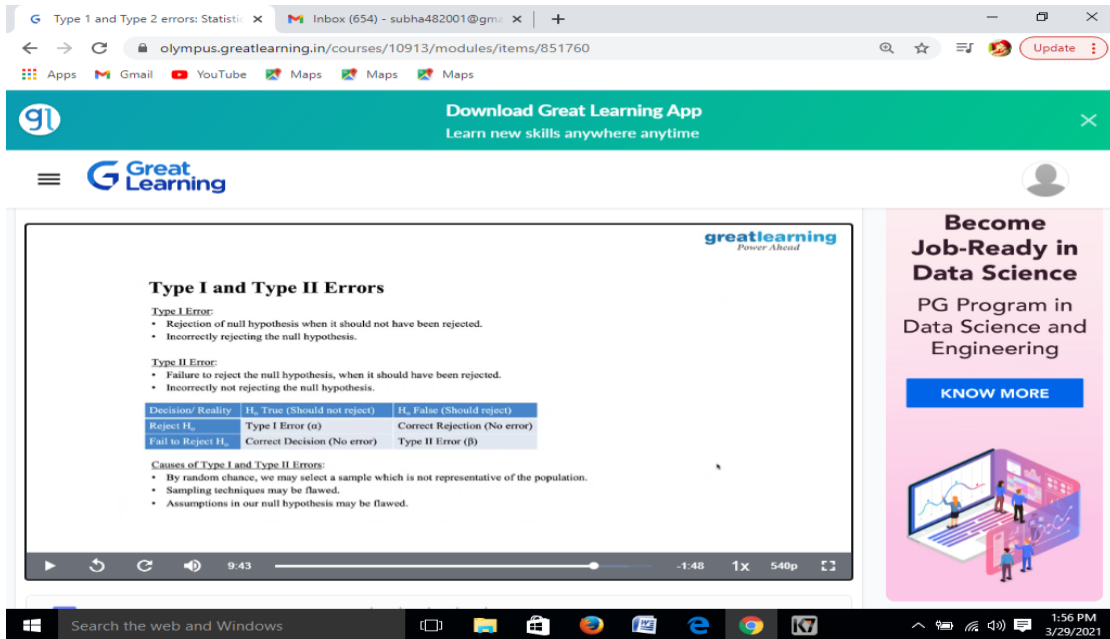
Type II Errors:

- Failure to reject the null hypothesis ,when it should have been rejected.
- Incorrectly not rejecting the null hypothesis.

Decision/Reality	H_0 True (Should not reject)	H_0 False (should reject)
Reject H_0	Type I Errors (α)	Correct Rejection (No error)
Fail to Reject H_0	Correct Decision (No error)	Type II Errors (β)

Causes of Type I Errors and Type II Errors :

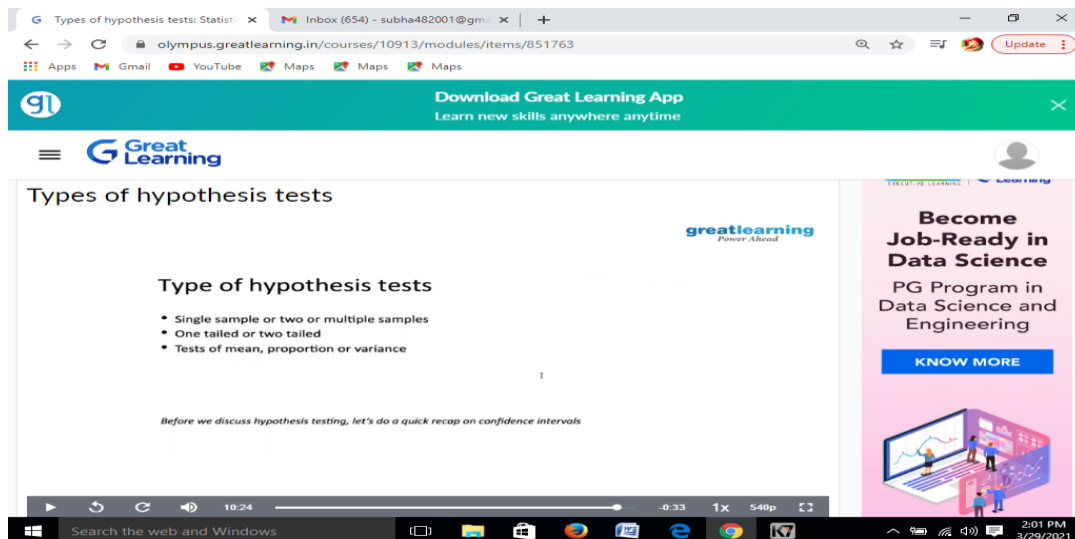
- By random chance ,we may select a sample which is not representative of the population.
- Sampling techniques may be flawed.
- Assumptions in our null hypothesis may be flawed.



Types of hypothesis tests

- Single sample or two or multiple samples
- One tailed or two tailed
- Tests of mean, proportion or variance

Before we discuss hypothesis testing ,let's do a quick recap on confidence intervals.



The screenshot shows a video player interface. The video content displays a slide with the following text:

Types of hypothesis tests

Type of hypothesis tests

- Single sample or two or multiple samples
- One tailed or two tailed
- Tests of mean, proportion or variance

Before we discuss hypothesis testing, let's do a quick recap on confidence intervals

On the right side of the video player, there is a promotional banner for a program:

Become Job-Ready in Data Science
PG Program in Data Science and Engineering
[KNOW MORE](#)

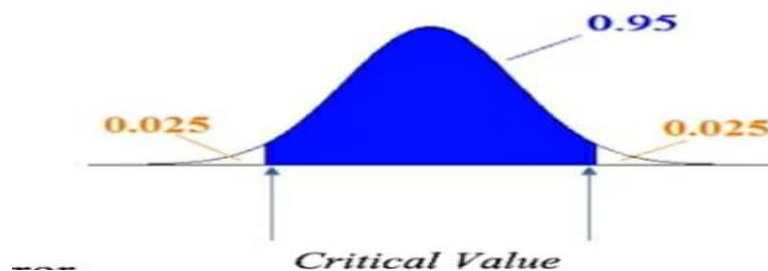
The video player interface includes a progress bar at the bottom showing 10:24, a volume icon, and a Windows taskbar at the very bottom with the search bar and system tray.

6.CONFIDENCE INTERVAL:

- 95% of all sample means (\bar{x}) are hypothesized to be in this region

=> This is called confidence interval.

- If sample mean is in the blue region ,we fail reject the null hypothesis

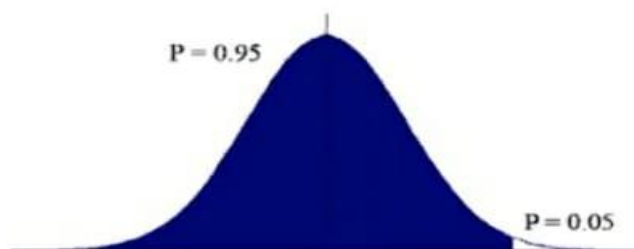


- If sample mean is in the white region ,we reject the null hypothesis .
- Here, $\alpha=0.05$

=> α is the null hypothesis is correct ,($\alpha*100$)% of the sample means should lie in the rejection region

In case of one-tailed situation:

- All of α is in one tail or the other ,depending on the alternative hypothesis
- H_a points to the tail ,where the critical value and the rejection region are(case when observed mean > hypothesis mean)



The screenshot shows a web browser window with the URL `olympus.greatlearning.in/courses/10913/pages/confidence-intervals?module_item_id=851746`. The page content includes a video player with a normal distribution curve. The text on the page explains confidence intervals and hypothesis testing. A sidebar on the right promotes a PG Program in Data Science and Engineering.

Example scenario to perform an hypothesis test

A study was done to see the effect of presence of dogs as pets on kids (ages 10 to 18). Two groups of teenagers, one group with teenagers who owned a dog for minimum 5 years and another group of kids who never owned a dog, were presented a questionnaire and score were computed. High score corresponds to higher cheerfulness and low score corresponds to lower cheerfulness.

Do dogs have a significant effect (either positive or negative) on the cheerfulness of kids?

Dog : 6.6,7.8,4.6,7.8,7.8,8.8,9.9,8.5,7.7,8.6,8.7,5.8,7.4

No - Dog : 9.8,8.3,7.1,7.2,8.1,8.9,6.7,7.5,7.8,7.6,7.3,6.4,6.8,7.6,4,7.9

What are the null and alternative hypothesis?

Is it a right tailed or a left tailed test ?

Is it a one sample or a two sample test ?

Is it a test of mean, proportion or variance?

Which statistical test do you think is appropriate?

Let's perform a two sample t-test to check if there is a significant difference in mean of the two sample

Avg-score of kids with dogs , $m_1, s_1=7.73, 1.24$

Avg-score of kids without dogs , $m_2, s_2=7.58, 1.23$

$|A-B|=0.10$

Is the difference significant at 5% significance level?

$H_0: m_1 = m_2$ (pets have no effect on the cheerfulness of kids)

$H_a: m_1 \neq m_2$ (pets have an effect on the cheerfulness of kids)

Alpha=0.05

t -critical = ± 2.11 (for a dot of 16, and a confidence of 95% in case of a two tailed test)

$$t = \frac{m_1 - m_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

t - statistic = 0.35

It is well inside the critical level .we could not prove that pets either increase or decrease the cheerfulness of kids We fail to reject the null hypothesis.

The screenshot shows a web browser window with the URL `olympus.greatlearning.in/courses/10913/modules/items/851748`. The page content includes:

Example scenario to perform an hypothesis test

A study was done to see the effect of presence of dogs as pets on kids (ages 10 to 18). Two groups of teenagers, one group with teenagers who owned a dog for minimum 5 years and another group of kids who never owned a dog, were presented a questionnaire and scores were computed. High score corresponds to higher cheerfulness and low score corresponds to lower cheerfulness.

Do dogs have a significant effect (either positive or negative) on the cheerfulness of kids?

Dog: 6.6, 7.8, 4.6, 7.8, 7, 8, 9, 9, 8.8, 9.9, 8.5, 7.7, 8.6, 8, 7, 5.8, 7.4
 No_dog: 9.8, 8.3, 7.1, 7.2, 8.1, 8.9, 6, 7, 7.5, 7.8, 7.6, 7.3, 6.4, 6.8, 7, 6.4, 7.9

What are the null and alternative hypothesis?
 Is it a right tailed or a left tailed test?
 Is it a one sample or a two sample test?
 Is it a test of mean, proportion or variance?
 Which statistical test do you think is appropriate?

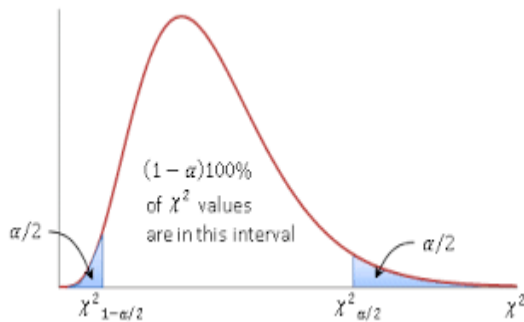
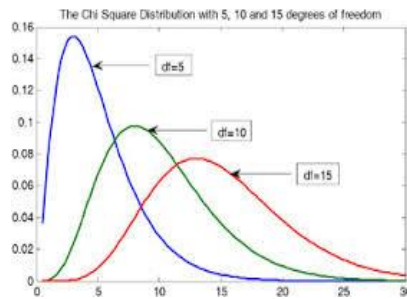
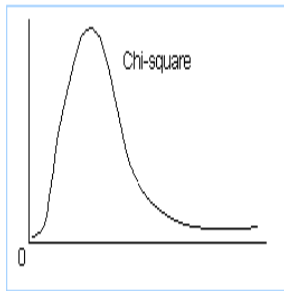
The sidebar on the right features the text: **Become Job-Ready in Data Science**, **PG Program in Data Science and Engineering**, and a **KNOW MORE** button.

- The test we performed is called a 2-sample t-test or independent samples t-test or student's t-test.
- We perform this test to see if there is a statistically significant difference between means of two independent groups.
- The null hypothesis will be that there is no difference in means.
- The alternative hypothesis will be that there is a significant difference in means.
- To perform this test, we will need one independent qualitative variable with two levels and one dependent qualitative variable.

7. CHI-SQUARE TEST:

When we take many samples of the same size from a normal population and find the sample means, they follow a normal distribution.

When we take many samples of the same size from a normal population and find the sample variances, they DO NOT follow a normal distribution; instead they follow a chi-square (χ^2) distribution, which is dependent on the degrees of freedom.



- Area under the curve is always 1
- Cumulative Probability runs from right to left; 1 is towards the left end, while 0 is towards the right.

Pop: $m_1, v_1 = 100, 16.11$

Lori's: $m_2, v_2 = 104.06, 40.99$

$N = 15$

Dof = $15 - 1 = 14$

$H_0: v_2 = v_1$ (Variance in income of Lori's is same as the population)

$H_a: v_2 > v_1$ (Variance in income of Lori's is higher than the population)

$\alpha = 0.05$

Chi_critical = 23.68 (any value beyond 23.68 falls in the rejection region)

$$\text{Chi_statistic} \Rightarrow x^2 = \text{dof} \left(\frac{v_2}{v_1} \right)$$

$$x^2 = 35.62$$

$$x^2 > 23.68$$

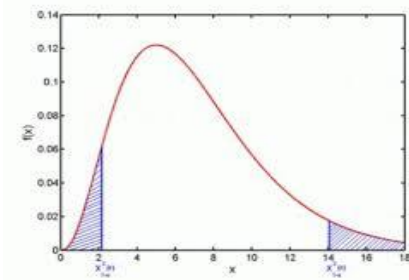
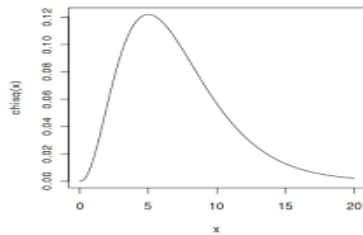
It is well beyond the critical value. The variation in income of blue collar workers at Lori's is significantly higher than the population variance.

Chi-square (χ^2) test compares the population variance, with the hypothesized variance.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad \text{where } n = \text{sample size}$$

s^2 = sample variance and σ^2 = population variance

At $\alpha = 0.05$ and $n = 5$ ($df = 4$)



p-value: How much of the area is above the test – statistic?

If it is less than the specific α , we reject the null hypothesis.

Chi-Square test: Statistical M | x | Inbox (655) - subha482001@gm... | x | chi-square test graphs - Google | x | +

olympus.greatlearning.in/courses/10913/pages/chi-square-test?module_item_id=851739

Download Great Learning App
Learn new skills anywhere anytime

Great Learning

Chi-Square test

Chi square test of variance

When we take many samples of the same size from a normal population and find the sample means, they follow a normal distribution.

When we take many samples of the same size from a normal population and find the sample variances, they DO NOT follow a normal distribution; instead they follow a **chi-square (χ^2) distribution**, which is dependent on the degrees of freedom.

- Area under the curve is always 1.
- Cumulative Probability runs from right to left; 1 is towards the left end, while 0 is towards the right.

download.png

Show all

4:55 PM 3/29/2021

8.ANOVA:

Hypothesis of One-Way ANOVA

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_k$$

All population means are equal

H_a : Not all of the population means are equal

For at least one pair ,the population means are unequal.

One Way ANOVA:

- The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant difference between the means of two or more independent (unrelated) groups

- For one -way ANOVA ,the ratio of the between -group variability to the within-group variability follows an F-distribution when the null hypothesis is true .when you perform a one -way ANOVA for a single study ,you obtain a single F-value.

Example

Three groups of samples of factory emissions of different plants of the same company were collected. The score is computed based on the composition of the emissions. We want to find out if there is any inconsistency or difference across the three groups.

A = 57,56,58,58,56,59,56,55,53,54,53,42,44,34,54,54,34,64,84,24

B = 49,47,49,47,49,46,45,46,41,42,42,42,14,14,34

C = 49,48,46,45,46,45,55,61,45,55,54,44,74,54,84,39

$$\text{Dof}(\text{between}) = k-1 = 3-1 = 2$$

$$\text{Dof}(\text{within}) = N - k = 59 - 3 = 56$$

$$\text{Dof}(\text{total}) = 56 + 2 = 58$$

For the above degrees of freedom,

$$F_{\text{critical}} = 3.161$$

$$\text{Mean}(A) = 52.45$$

$$\text{Mean}(B) = 41.36$$

$$\text{Mean}(C) = 51.45$$

$$\text{Overall Mean} = 2864/59 = 48.54$$

$$SS_{\text{total}} = \sum(x_i - \text{overall_mean})^2 = 8548.64$$

$$SS_{\text{within}} = \sum(a_i - \text{mean}(A))^2 + \sum(b_i - \text{mean}(B))^2 + \sum(c_i - \text{mean}(C))^2 = 7096.32$$

$$SS_{\text{between}} = SS_{\text{total}} - SS_{\text{within}} = 1452.32$$

$$MS_{\text{between}} = \frac{SS(\text{between})}{dof(\text{between})} = 726.16$$

$$MS_{\text{within}} = \frac{SS(\text{within})}{dof(\text{within})} = 126.72$$

$$F_{\text{statistic}} = \frac{MS(\text{between})}{MS(\text{within})} = 5.73$$

$$F_{\text{statistic}} > F_{\text{critical}}$$

Since our f-statistic is beyond the critical value, we reject the null.

The screenshot shows a web browser window with the URL `olympus.greatlearning.in/courses/10913/pages/annova?module_item_id=851736`. The page content includes the Great Learning logo and a navigation menu. The main content area is titled "Annova" and contains the following text and formulas:

greatlearning
Power Ahead

Dof(between) = $k - 1 = 3 - 1 = 2$
Dof(within) = $N - k = 59 - 3 = 56$
Dof(total) = $56 + 2 = 58$
For the above degrees of freedom,
 $F_{critical} = 3.161$

Mean(A) = 52.45
Mean(B) = 41.36
Mean(C) = 51.45
Overall Mean = $2864/59 = 48.54$

$$SS_{total} = \sum(x_i - overall_mean)^2 = 8548.64$$
$$SS_{within} = \sum(a_i - mean(A))^2 + \sum(b_i - mean(B))^2 + \sum(c_i - mean(C))^2 = 7096.32$$
$$SS_{between} = SS_{total} - SS_{within} = 1452.32$$

On the right side, there is a promotional banner for the "PG Program in Data Science and Engineering" with a "KNOW MORE" button and an illustration of people working at computers.

PG PROGRAMS OVERVIEW:

Kumar Muthuraman, Faculty director, Centre for Analytics and Transformative Technologies, PGP-AIML introduced that post graduate programs in

1. Data Science and Business Analytics (DSBA) in Great learning,

Browser tabs: Data Science and Business A, Inbox (655) - subha482001@gm..., chi-square test graphs - Google

Address bar: olympus.greatlearning.in/courses/10913/pages/data-science-and-business-analytics-dsba?module_ite...

Navigation: Apps, Gmail, YouTube, Maps, Maps, Maps

Download Great Learning App
Learn new skills anywhere anytime

Great Learning

Data Science and Business Analytics (DSBA)

The University of Texas at Austin
PG Program in DATA SCIENCE AND BUSINESS ANALYTICS
Program Partner: [greatlearning](#)

URL: https://olympus.greatlearning.in/courses/10913/pages/data-science-and-business-analytics-dsba?module_item_id=1099214

Windows taskbar: Search the web and Windows, 5:46 PM 3/29/2021

2. Artificial Science and Machine Learning

The screenshot shows a web browser window with the following elements:

- Browser Tabs:** "Artificial Intelligence & Machi...", "Inbox (655) - subha482001@gm...", "chi-square test graphs - Google...".
- Address Bar:** "olympus.greatlearning.in/courses/10913/pages/artificial-intelligence-and-machine-learning?module_ite...".
- Navigation:** Back, Forward, Refresh, Home, and Update buttons.
- App Bar:** "Download Great Learning App" with the tagline "Learn new skills anywhere anytime".
- Header:** "Great Learning" logo and a user profile icon.
- Main Content:**
 - Section Header:** "Artificial Intelligence & Machine Learning".
 - Video Player:** A video titled "Post Graduate Program in Artificial Intelligence and Machine Learning" featuring a robot head with a digital display. The video player shows a progress bar at 0:07 and a duration of -1:44.
 - Logos:** "GREAT LAKES EXECUTIVE LEARNING" and "TEXAS The University of Texas at Austin".
- Right Sidebar:**
 - Text:** "Become Job-Ready in Data Science", "PG Program in Data Science and Engineering".
 - Button:** "KNOW MORE".
 - Image:** An illustration of people interacting with data visualizations.

The Windows taskbar at the bottom shows the search bar, taskbar icons for various applications, and the system tray with the date and time: "5:52 PM 3/29/2021".

QUIZ

Checkpoint questions were provided to check our understandability of the concepts and to increase the real life application of probability. The quiz answers were checked and the correct answers were provided for our reference.

Quiz: Statistical Methods for Dec: x | Inbox (655) - subha482001@gm: x | chi-square test graphs - Google: x | +

olympus.greatlearning.in/courses/10913/quizzes/68515?module_item_id=851959

Apps | Gmail | YouTube | Maps | Maps | Maps

Great Learning

PG Programs Overview

Quiz

Quiz

Claim your course certificate

1 2 3 4 5 6 7 8 9 10

Q No. 1 Points: 1

The sampling distribution of the sample mean is approximately normal if

all possible samples are selected.

the sample size is large.

the standard error of the sampling distribution is small.

None of the above

Previous SUBMIT QUIZ Next

Search the web and Windows

5:56 PM 3/29/2021

CONCLUSION:

The internship was a useful experience. It helped to gain new knowledge and skills. It provides a path to achieve several of our learning goals. This internship

programme was not one sided, but it was a way of sharing knowledge, ideas and opinions. The new insights and motivation to pursue one's career using mathematics is gained through this internship programme.