Applied Statistics

Agenda - Common Statistical Tests - Week 3

1. Some important tests
   a. Test for one mean
   b. Test for equality of means
   c. Test for equality of means - equal std dev
   d. Test for equality of means - unequal std dev
   e. Paired test for equality of means [*additional content*]
   f. Test for one proportion
   g. Test for two proportions
   h. Test for one variance [*additional content*]
   i. Test for equality of variances [*additional content*]
   j. Test of independence
   k. ANOVA test

Some important Statistical Tests

Hypothesis Testing Frameworks

Choice of test depends on test statistic and data availability

Means

Compare the sample mean to the population mean when std dev is known
1-sample z-test

Compare the sample mean to the population mean when std dev is unknown
1-sample t-test

Compare the sample means from 2 independent populations when std devs are known
2-sample ind. z-test

Compare the sample means from 2 independent populations when std devs are unknown
2-sample ind. t-test

Compare the sample means from 2 related populations when std devs are unknown
Paired t-test

Compare the sample means from 2 or more independent populations
ANOVA Test

Proportions

Compare the sample proportion to the population proportion
1-sample z-test

Compare the sample proportions from two populations
2-sample z-test

Variances

Compare the sample variance to the population variance
Chi-Square test

Compare the sample variances from two populations
F-test

Frequencies

Check whether the categorical variables from a population are independent
Chi-Square Test of Independence

Test for one mean

Example

A certain food aggregator ZYX is facing stiff competition from its main rival SWG during Corona period. To retain business, ZYX is advertising that, within a radius of 5 km from the restaurant where the order is placed, it can still deliver in 40 minutes or less on the average (and changed condition has not made any impact on them).

The delivery times in minutes of 25 randomly selected deliveries are given in a CSV file.

Assuming the delivery distribution is approximately normal, is there enough evidence that ZYX's claim is false?

This is clearly a one-tailed hypothesis problem, concerning population mean μ, the average delivery time.

Test for One Mean - Unknown Std Dev

Significance of the test	Assumptions	Test Statistic Distribution
Test for population mean $H_0 : \mu = \mu_0$	• Continuous data • Normally distributed population and sample size < 30 • Unknown population standard deviation • Random sampling from the population	t distribution (The test is also known as **One-sample t-test**)

Test for equality of means (Known std dev)

Example

To compare customer satisfaction levels of two competing media channels, 150 customers of Channel 1 and 300 customers of Channel 2 were randomly selected and were asked to rate their channels on a scale of 1-5, with 1 being least satisfied and 5 most satisfied.
(The survey results are summarized in a CSV file)

Test at 0.05 level of significance whether the data provide sufficient evidence to conclude that channel 1 has a higher mean satisfaction rating than channel 2.

This is a two-sample problem where the channel 1 and channel 2 populations are independent. Further, this is a one-tailed hypothesis problem, concerning population means μ_1 and μ_2, the mean customer satisfaction for channel 1 and channel 2 respectively.

Test for Equality of Means - Known Std Devs

Significance of the test	Assumptions	Test Statistic Distribution
Test for equality of two population means $H_0 : \mu_1 = \mu_2$	• Continuous data • Normally distributed population or sample size > 30 • Independent populations • Known population standard deviations σ_1 and σ_2 • Random sampling from the population	Standard Normal distribution (The test is also known as **Two independent sample z-test**)

Test for equality of means
(Equal and unknown std dev)

Example

In the lockdown period, because of working from home and increased screen time, many opted for listening to FM Radio for entertainment rather than watching Cable TV. An advertisement agency randomly collected daily usage time data (in minutes) from both type of users and stored it in a CSV file.

Assuming daily Radio and TV usage time are normally distributed, do we have enough evidence to conclude that there is any difference between daily TV and Radio usage time at 0.05 significance level?

This is a two-sample problem where FM Radio and Cable TV users are assumed independent. Further, this is a two-tailed hypothesis problem, concerning population means μ_1 and μ_2, the daily mean usage time of Radio and TV respectively.

Test for Equality of Means : Equal Std Devs

Significance of the test	Assumptions	Test Statistic Distribution
Test for equality of two population means $H_0 : \mu_1 = \mu_2$	• Continuous data • Normally distributed populations • Independent populations • Equal population standard deviations • Random sampling from the population	t distribution (The test is also known as **Two independent sample t-test**)

Test for equality of means (Unequal and unknown std dev)

Example

SAT verbal scores of two groups of students are given in a CSV file. The first group, **College**, contains scores of students whose parents have at least a bachelor's degree and the second group, **High School**, contains scores of students whose parents do not have any college degree.

The Education Department is interested to know whether the sample data support the theory that students show a higher population mean verbal score on SAT if their parents attain a higher level of education.

Assuming SAT verbal scores for two populations are normally distributed, do we have enough statistical evidence for this at 5% significance level?

This is a two-sample problem as the College and High School populations are different. Further, this is a one-tailed hypothesis problem, concerning population means μ_1 and μ_2, the mean verbal score on SAT for College and High School groups.

Test for Equality of Means : Unequal Std Devs

Significance of the test	Assumptions	Test Statistic Distribution
Test for equality of two population means $H_0 : \mu_1 = \mu_2$	• Continuous data • Normally distributed populations • Independent populations • Unequal population standard deviations • Random sampling from the population	t distribution (The test is also known as **Two independent sample t-test**)

Test for One Proportion

Example

A researcher claims that Democratic party will win in the next United States Presidential election.

To test her belief the researcher randomly surveyed 90 people and 24 out of them said that they voted for Democratic party.

Is there enough evidence at $\alpha = 0.05$ to support this claim?

This is clearly a one-tailed test, concerning population proportion p, the proportion of people voted from Democratic party.

Test for One Proportion

Significance of the test	Assumptions	Test Statistic Distribution
Test for population proportion $H_0 : p = p_0$	• Binomially distributed population • Random sampling from the population • When both mean (np) and n(1-p) are greater than or equal to 10, the binomial distribution can be approximated by a normal distribution	Standard Normal distribution (The test is also known as **One proportion z-test**)

Test for Two Proportions

Example

A car manufacturer aims to improve its products' quality by reducing the defects. So, the manufacturer randomly checks the efficiency of two assembly lines in the shop floor. In line 1, there are 20 defects out of 200 samples and In line 2, there are 25 defects out of 400 samples.

At 5% level of significance, do we have enough statistical evidence to conclude that the two assembly procedures are different?

This is clearly a one-tailed test, concerning two population proportion p_1 and p_2, the proportion of defects in assembly line 1 and assembly line 2 respectively.

Test for Two Proportions

Significance of the test	Assumptions	Test Statistic Distribution
Test for equality of two population proportions $H_0 : p_1 = p_2$	• Binomially distributed populations • Independent populations • Random sampling from the populations • When both mean (np) and n(1-p) are greater than or equal to 10, the binomial distribution can be approximated by a normal distribution	Standard Normal distribution (The test is also known as **Two proportions z-test**)

Test of Independence

Chi Square Test for Independence

2x2 contingency table that describes two variables (smoking and gender) at two levels each and stores the number of observations at each cell

	Male	Female	Total
Smoker	120	100	220
Non-smoker	60	140	200
Total	180	240	420

We are interested to know whether the **two variables are independent**

H_0: Smoking and gender are independent.

H_a: Smoking and gender are not independent.

Example

The following table summarizes beverage preference across different age-groups.

	Beverage Preference		
Age	Tea/Coffee	Soft Drink	Others
21 - 34	25	90	20
35 - 55	40	35	25
> 55	24	15	30

Does beverage preference depend on age?

This is a problem of Chi-Square test of independence, concerning the two independent categorical variables, Age and Beverage Preference.

Chi-Square Test for Independence

Significance of the test	Assumptions	Test Statistic Distribution
In a contingency table **H_0 : The row and column variables are independent**	• Categorical variables • Expected value of the number of sample observations in each level of the variable is at least 5 • Random sampling from the population	Chi Square distribution (The test is also known as **Chi-square test of independence**)

Analysis of Variance (ANOVA)

ANOVA Test : Introduction

Analysis **o**f **Va**riance (ANOVA) is used to determine whether the means of more than two independent populations are significantly different.

H_0: $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$
H_a: **at least one of these means if not the same**

Why do we call it ANOVA? - The mathematical tools to calculate the p-value rely heavily on using the variances of the populations.

ANOVA is used in various problems such as comparing the yields of the crop from several varieties of seeds, comparing the gasoline mileage of various types of automobiles, etc.

ANOVA Test : Some important terms

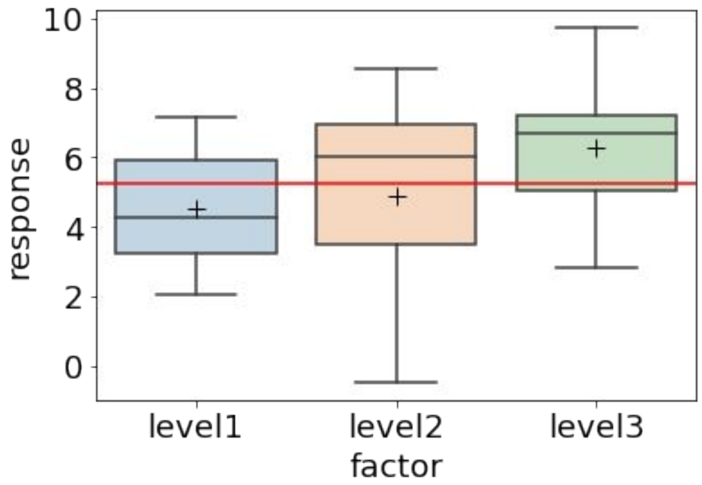
Response: Dependent variable which is continuous and assumed to follow a normal distribution

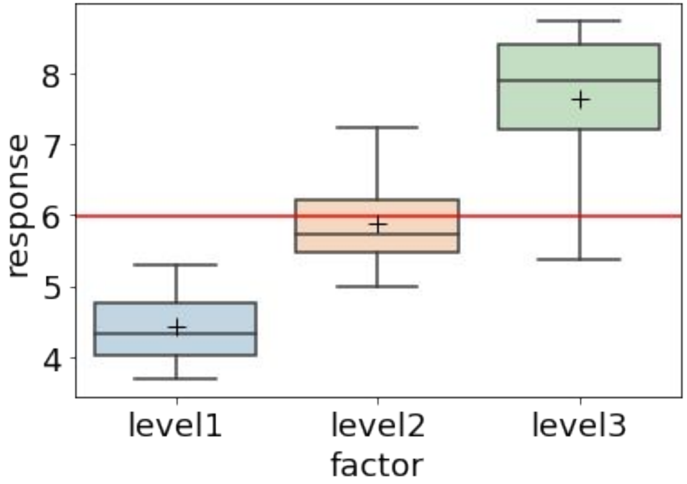
Consider, an example where interest lies in comparing the weekly volume of sales by different teams of sales executives.

Factor: Independent explanatory variable with several levels

ANOVA Test : One-way ANOVA

One-way ANOVA is used when the response variable depends on a single factor.





Between group variation is lower

Between group variation is higher

ANOVA Test : How it works

F-Statistic is the ratio of the between group variations to within group variations.

$$F - statistic = \frac{Between\ group\ variations}{Within\ group\ variations}$$

A large value of F-Statistic indicates that there is more variation between groups than within groups.

Thus, it will provide evidence against the null hypothesis.

Example

Traffic management inspector in a certain city wants to understand whether carbon emissions from different cars are different. The inspector has reasons to believe that Fuel type may be one important factor responsible for differences in carbon emission.

For this purpose, the inspector has taken random samples from all registered cars on the road in that city and would like to test if the amount of carbon emission release depends on fuel type at 5% significance level.

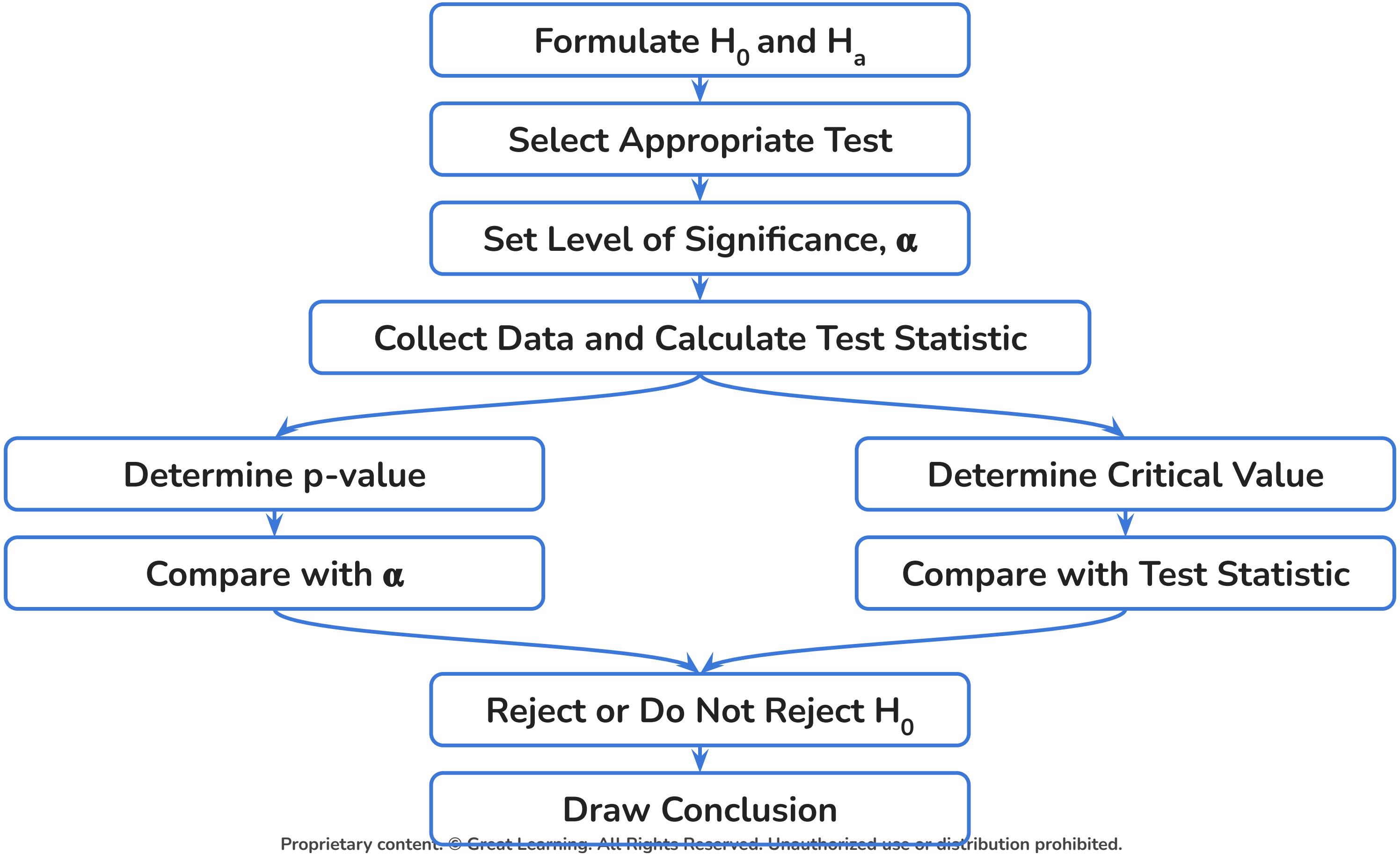
Here, we will compare the means of emission for the three different fuel types.

ANOVA Test : One-way ANOVA

Significance of the test	Assumptions	Test Statistic Distribution
Test for means for more than two populations **H_0 : All population means are equal**	• The populations are normally distributed • Samples are independent simple random samples • Population variances are equal	F distribution (The test is also known as **One-way ANOVA F-test**)

Now let’s summarize

Hypothesis Testing Steps



A flowchart to help you choose