Advance Statistics

PROJECT REPORT

Contents

PROJECT REPORT- Advance statistics

Problem 1

A physiotherapist with a male football team is interested in studying the relationship between foot injuries and the positions at which the players play from the data collected

Total	Winger	Attacking Midfielder	Forward	Striker	
145	20	24	56	45	Players Injured
90	9	11	38	32	Players Not Injured
235	**29**	**35**	**94**	**77**	**Total**

1.1 What is the probability that a randomly chosen player would suffer an injury?

-The total number of players are 235.

-The total number of players Injured are 145

-The probability that a randomly chosen player would suffer an injury is 145/235= 0.62 or 62%

1.2 What is the probability that a player is a forward or a winger?

-The total number of players are 235.

-The total number of players playing as a forward are 94

-The total number of players playing as a winger are 29

-The probability that a player is a forward or a winger 0.52 or 52.34%

1.3 What is the probability that a randomly chosen player plays in a striker position and has a foot injury?

-The total number of players which are strikers are 77

-The total number players which are striker and have injury are 45

So, the probability for player which are strikers and have foot injury are 45/77 which is 0.5844 or 58.44%

The probability that a randomly chosen player plays in a striker position and has a foot injury is 0.19

1.4 What is the probability that a randomly chosen injured player is a striker?

-The Total number of injured players are 145

-Total number of players which are injured and striker are 45

So, the probability of player chosen which are injured and is a striker is 45/145 which is 0.31 or 31%

1.5 What is the probability that a randomly chosen injured player is either a forward or an attacking midfielder?

-The total number of injured players are 145

-Total number of players which are injured and forward and midfielder are 56 and 24 respectively.

So, the probability of players chosen injured and are forward and midfielder are (56+24)/145 which is 0.5517 or 55.17%

Problem 2

An independent research organization is trying to estimate the probability that an accident at a nuclear power plant will result in radiation leakage. The types of accidents possible at the plant are, fire hazards, mechanical failure, or human error. The research organization also knows that two or more types of accidents cannot occur simultaneously.

According to the studies carried out by the organization, the probability of a radiation leak in case of a fire is 20%, the probability of a radiation leak in case of a mechanical 50%, and the probability of a radiation leak in case of a human error is 10%. The studies also showed the following;

- **The probability of a radiation leak occurring simultaneously with a fire is 0.1%.**
- **The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%.**
- **The probability of a radiation leak occurring simultaneously with a human error is 0.12%.**

On the basis of the information available, answer the questions below:

GIVEN-

- ★ The probability of a radiation leak in case of a fire is 20%,
- ★ The probability of a radiation leak in case of a mechanical 50%,

- ★ The probability of a radiation leak in case of a human error is 10%
- ★ The probability of a radiation leak occurring simultaneously with a fire is 0.1%
- ★ The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%
- ★ The probability of a radiation leak occurring simultaneously with a human error is 0.12

2.1 What are the probabilities of a fire, a mechanical failure, and a human error respectively?

Probability of Fire

= Probability of a radiation leak in case of a fire/probability of a radiation leak in case of a fire

= 0.10/20

=0.005

Probability of a mechanical failure

=Probability of a radiation leak in case of a Mechanical failure / probability of a radiation leak in case of a Mechanical failure

= 0.15/50

=0.003

Probability of Human Error

probability of a radiation leak in case of a Human Error / probability of a radiation leak in case of a Human Error

= 0.15/50

=0.012

2.2 What is the probability of a radiation leak?

Probability of a radiation leak = The probability of a radiation leak occurring simultaneously with a fire + probability of a radiation leak occurring simultaneously with a mechanical failure + The probability of a radiation leak occurring simultaneously with a human error

Probability of a radiation leak = 0.1+0.15+0.12 = 0.37

Probability of Radiation Leak is 0.37

2.3 Suppose there has been a radiation leak in the reactor for which the definite cause is not known. What is the probability that it has been caused by:

- A Fire.
- A Mechanical Failure.
- A Human Error.

A-Probability that it has been caused by Fire = probability of a radiation leak in case of a fire/Probability of a radiation leak

Probability that it has been caused by Fire = 0.1/0.37

Probability that it has been caused by Fire is 0.27

B- Probability that it has been caused by Mechanical failure = probability of a radiation leak in case of a mechanical failure/Probability of a radiation leak

Probability that it has been caused by Mechanical failure = 0.15/0.37

Probability that it has been caused by Mechanical failure is 0.41

Problem 3:

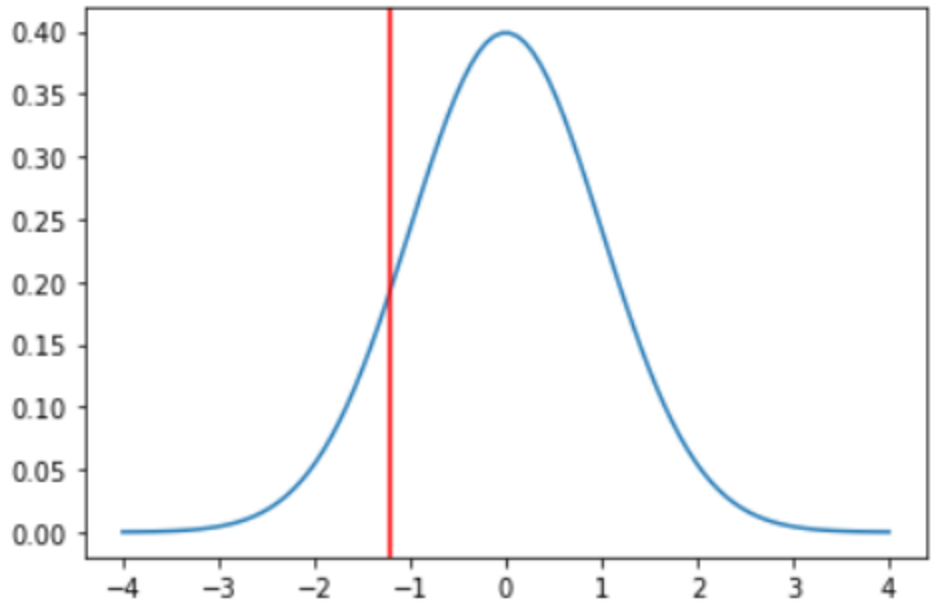
The breaking strength of gunny bags used for packaging cement is normally distributed with a mean of 5 kg per sq. centimeter and a standard deviation of 1.5 kg per sq. centimeter. The quality team of the cement company wants to know the following about the packaging material to better understand wastage or pilferage within the supply chain; Answer the questions below based on the given information; (Provide an appropriate visual representation of your answers, without which marks will be deducted)

Given

- Gunny bags used for packaging cement is normally distributed
- Mean is 5 kg per sq. centimetre
- Standard deviation is 1.5 kg per sq. centimetre

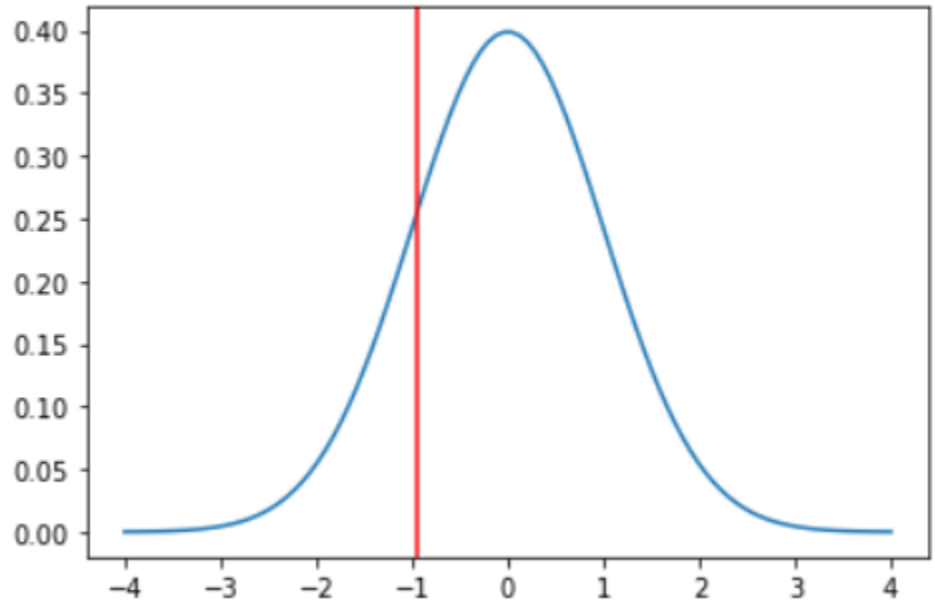
3.1 What proportion of the gunny bags have a breaking strength less than 3.17 kg per sq cm?

11.12% proportion of the gunny bags have a breaking strength less than 3.17 kg per sq cm



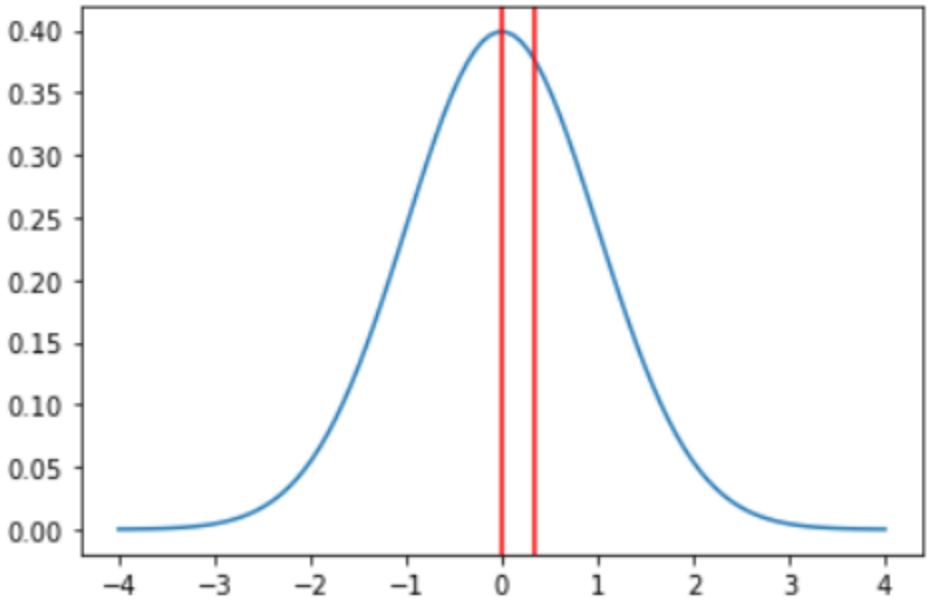
3.2 What proportion of the gunny bags have a breaking strength at least 3.6 kg per sq cm.?

82.45% of the gunny bags have a breaking strength at least 3.6 kg per sq cm.



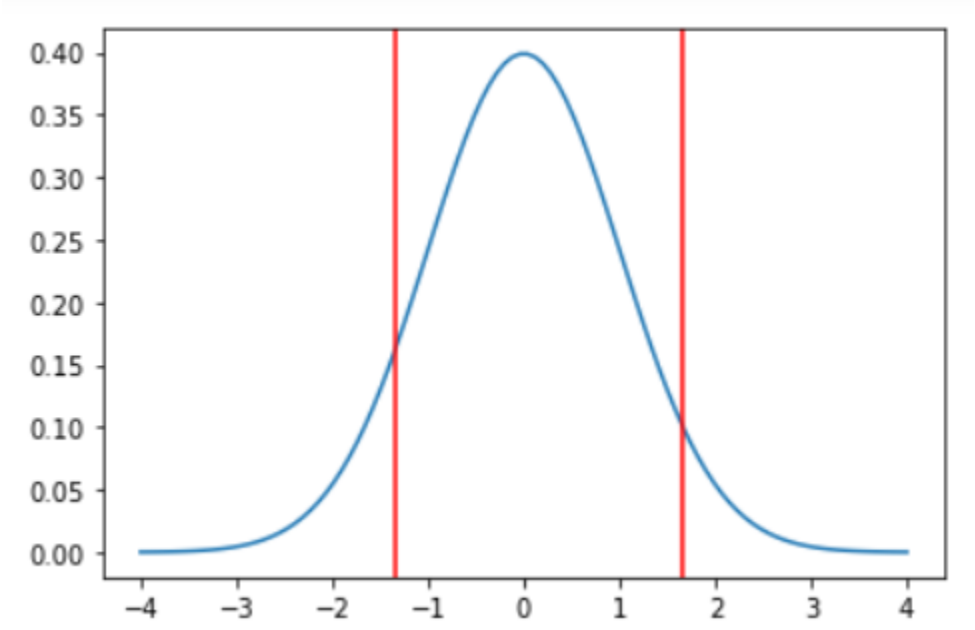
3.3 What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?

13.05% of gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.



3.4 What proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.?

13.90% of proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.



Problem 4:

Grades of the final examination in a training course are found to be normally distributed, with a mean of 77 and a standard deviation of 8.5. Based on the given information answer the questions below.

GIVEN

- Grades of the final examination in a training course are found to be normally distributed
- Mean-77
- Standard deviation-8.5

4.1 What is the probability that a randomly chosen student gets a grade below 85 on this exam?

82.66% of student gets a grade below 85 on this exam.

4.2 What is the probability that a randomly selected student scores between 65 and 87?

80.12% is the probability that a randomly selected student score between 65 and 87.

4.3 What should be the passing cut-off so that 75% of the students clear the exam?

82.73 marks is the passing cut-off so that 75% of the students clear the exam

Problem 5:

Zingaro stone printing is a company that specializes in printing images or patterns on polished or unpolished stones. However, for the optimum level of printing of the image the stone surface has to have a Brinell's hardness index of at least 150. Recently, Zingaro has received a batch of polished and unpolished stones from its clients. Use the data provided to answer the following (assuming a 5% significance level);

Load the required packages, set the working directory and load the data file.

Dataset has two variables Unpolished and Treated and polished

Pre-analysis

- The Data Set consist of 75 Rows (Samples) and 2 Columns (Features).
- Mean for Unpolished is 134.11 and for Treated and Polished is 147.79
- Standard Deviation for Unpolished is 33.04 and for Treated and Polished is 15.59
- There are no Missing Values
- Both the data are Normally distributed across the samples

These are two and one outliers in **Treated and Polished and Unpolished respectively**

5.1 Earlier experience of Zingaro with this particular client is favorable as the stone surface was found to be of adequate hardness. However, Zingaro has reason to believe now that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

Step 1: Define null and alternative hypotheses

Null Hypotheses states that Mean Brinell's hardness Index of unpolished stone surface, u is greater than or equal to 150.

Alternate Hypothesis states that Mean Brinell's hardness Index of unpolished stone surface, u is less than 150.

OR

Ho: u>=150

Ha: u<150

STEP 2-Decide the significance Level

Here we select alpha, α=0.05

Step 3: Identify the test statistic

We do not know the population standard deviation although the sample size is more than 30 still we use the t distribution and the $tSTAT$ test statistic. It is left tailed t test.

Step 4: Calculate the p - value and test statistic

SciPy. stats. ttest_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.

Values of t statistic= -4.1646296
Value of p = 4.171286995e-05

Step 5 Decide to reject or accept null hypothesis

At Level of significance: 0.05

We reject the null hypothesis since p value < Level of significance

So the statistical decision is we reject the null hypothesis at 5% level of significance

It means that there is sufficient evidence for Zingaro stone printing company to believe that unpolished stones are not suitable for printing, that is they have Brinell's hardness index of less than 150.

5.2 Is the mean hardness of the polished and unpolished stones the same?

- Mean for Unpolished is 134.11 and for Treated and Polished is 147.79. Therefore, the mean is not same.

Problem 6:

Aquarius health club, one of the largest and most popular cross-fit gyms in the country has been advertising a rigorous program for body conditioning. The program is considered successful if the candidate is able to do more than 5 push-ups, as compared to when he/she enrolled in the program. Using the sample data provided can you conclude whether the program is successful? (Consider the level of Significance as 5%)

Note that this is a problem of the paired-t-test. Since the claim is that the training will make a difference of more than 5, the null and alternative hypotheses must be formed accordingly.

Step 1: Define null and alternative hypotheses

Step 1: Define null and alternative hypotheses

In testing whether the program is successfull or not which is defined if the candidate is able to do more than 5 push ups as compared to when he or she joined the program.

Here u1 is the count of push ups after joining the program.

u2 is the count of push ups before joining the program.

the null hypothesis states that the difference after joining the program push up counts has increased to more than 5.

the alternate hypothesis states that after joining the program push up count has not increased to more than 5.

Ho : u1-u2 < 5

Ha : u1 - u2 >5

Step 2: Decide the significance level

Here we select α = 0.05 and the population standard deviation is not known.

Step 3: Identify the test statistic

- We have two samples and we do not know the population standard deviation.
- Sample sizes for both samples are same.
- We you use the t distribution and the t_{STAT} test statistic for two sample unpaired test.

Using the sample data provided can you conclude whether the program is successful?

- The T statistic is: 19.322619811082458
- The corresponding pvalue is: 1.1460209626255983e-35
- In this scenario, the p value is very less than the 0.05.
- Hence we reject the null hypothesis
- The program is not successful. .

Problem 7:

Dental implant data: The hardness of metal implant in dental cavities depends on multiple factors, such as the method of implant, the temperature at which the metal is treated, the alloy used as well as on the dentists who may favor one method above another and may work better in his/her favorite method. The response is the variable of interest.

1. Test whether there is any difference among the dentists on the implant hardness. State the null and alternative hypotheses. Note that both types of alloys cannot be considered together. You must state the null and alternative hypotheses separately for the two types of alloys.?

Hypothesis for the Anova

H_0: The mean response is the same for all three dentists.

H_a: For at least one pair of dentists the mean response will be different.

H_0: The mean response is same for both types of alloys.

H_a: The mean response is different for both types of alloys.

Testing of the null Hypothesis

After performing one way Anova on 'Dentist' with respect to 'response' we get p value as 0.11

Sine the p value is greater than alpha (0.05) we fail to reject null hypothesis.

Thus, the mean response for all the three types of dentist is same.

2. Before the hypotheses may be tested, state the required assumptions. Are the assumptions fulfilled? Comment separately on both alloy types.?

Required assumptions-

The samples drawn from different populations are independent and random.

There should be no significant outliers.

Dependent variable should be measured at the continues level.

Independent variables should each consist of tow or more categorical , independent groups.

Dependent variable should be approximately normally distributed for each combination of the group so two independent variables.

Number of observations in each group are same.

There is homogeneity of variance

IN OUR DATASET THE FOLLOWING ASSUMPTIONS ARE FULLFILLED.

- The samples drawn from different populations are independent and random.
- Outliers are removed
- Dependent variable is measured at continuous level.
- Independent variables consist of two or more categorical, independent groups.
- In three variables pvalue is more than alpha, so fail to reject H0, and variances are equal
- We used Shapiro test and Anderson Darling test to check weather are sample data is from normal distribution or not. Some of the variables are not normal.

3. Irrespective of your conclusion in 2, we will continue with the testing procedure. What do you conclude regarding whether implant hardness depends on dentists? Clearly state your conclusion. If the null hypothesis is rejected, is it possible to identify which pairs of dentists differ?

Hypothesis for the Anova

H_0: The mean response is the same for all three dentists.
H_a: For at least one pair of dentists the mean response will be different.

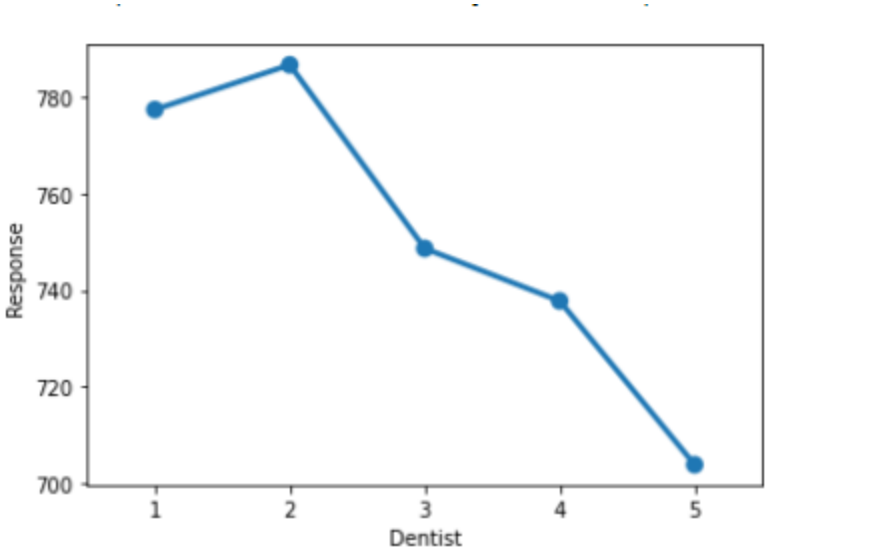
Testing of the null Hypothesis

After performing one way Anova on ‘Dentist’ with respect to ‘response’ we get p value as 0.11

Sine the p value is greater than alpha (0.05) we fail to reject null hypothesis.

Thus, the mean response for all the three types of dentist is same.

After drawing the poinplot we can clearly see that mean count for dentist 1 is way highest, but the sample data is not enough to conclude that. Also Anova does not help us identify which pairs of dentist differ.



4. Now test whether there is any difference among the methods on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which pairs of methods differ?

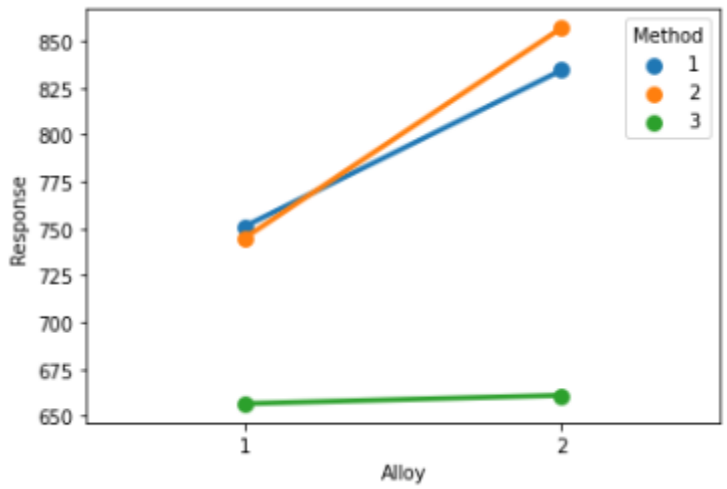
Hypothesis for the Anova

H_0: The mean response is the same for all three types of methods on the hardness of dental implant.
H_a: For at least one pair of dentists the mean response will be different.

Testing of the null Hypothesis

After performing Anova and looking at the interaction effect we get p value as smaller than alpha thus we reject null hypothesis.

After drawing the pointplot we can clearly see there is an intereaction effect, but the sample data is not enough to conclude that. Also Anova does not help us identify which pairs of dentist differ.



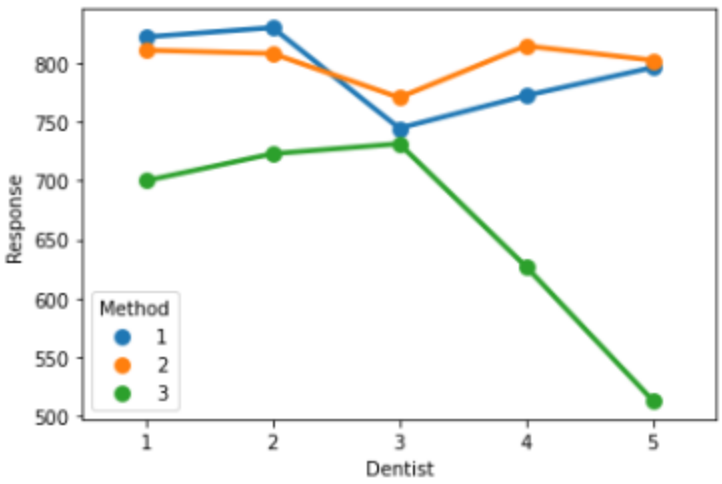
5. Now test whether there is any difference among the temperature levels on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which levels of temperatures differ?

Null hypothesis-

6. Consider the interaction effect of dentist and method and comment on the interaction plot, separately for the two types of alloys?

P value for interaction effect of dentist and method is 1.657388e-02

Since the p value for interaction effect is way less than alpha, we can conclude that there is no effect of interaction effect on our response variable.



Since the lines are not parallel to each other and clearly two lines are intersecting each other it means that there is significant interaction between the dentist and method used.

7. Now consider the effect of both factors, dentist, and method, separately on each alloy. What do you conclude? Is it possible to identify which dentists are different, which methods are different, and which interaction levels are different?

It is not possible to find out which methods are different, and which interaction levels are different using anova.