**ADVANCE STATISTICS PROJECT REPORT**

**CONTENTS:**

|  |  |  |
| --- | --- | --- |
| **1** | **Physiotherapist and football team Problem** | **5** |
|  |  |  |
| **1.1** | **What is the probability that a randomly chosen player would suffer an injury?** | **5** |
| **1.2** | **What is the probability that a player is a forward or a winger?** | **6** |
| **1.3** | **What is the probability that a randomly chosen player plays in a striker position and has a foot injury?** | **6** |
| **1.4** | **What is the probability that a randomly chosen injured player is a striker?** | **7** |
| **1.5** | **What is the probability that a randomly chosen injured player is either a forward or an attacking midfielder?** | **7** |
|  |  |  |
| **2** | **Nuclear power plant and Radiation leakage** | **8** |
|  |  |  |
| **2.1** | **What are the probabilities of a fire, a mechanical failure, and a human error respectively?** | **9** |
| **2.2** | **What is the probability of a radiation leak?** | **9** |
| **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.** | **10** |
|  |  |  |
| **3** | **Breaking strength of gunny bags problem** | **11** |
|  |  |  |
| **3.1** | **What proportion of the gunny bags have a breaking strength less than 3.17 kg per sq cm?** | **11** |
| **3.2** | **What proportion of the gunny bags have a breaking strength at least 3.6 kg per sq cm.?** | **12** |
| **3.3** | **What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?** | **13** |
| **3.4** | **What proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.?** | **14** |
|  |  |  |
| **4** | **Scores of student’s problem** | **15** |
|  |  |  |
| **4.1** | **What is the probability that a randomly chosen student gets a grade below 85 on this exam?** | **15** |
| **4.2** | **What is the probability that a randomly selected student scores between 65 and 87?** | **16** |
| **4.3** | **What should be the passing cut-off so that 75% of the students clear the exam?** | **17** |
|  |  |  |
| **5** | **Zingaro Stone Printing Company Probelm** |  |
|  |  | **17** |
| **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?** | **19** |
| **5.2** | **Is the mean hardness of the polished and unpolished stones the same?** | **19** |
|  |  |  |
| **6** | **Aquarius Health club problem** | **20** |
|  |  |  |
|  | **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** | **20** |
|  |  |  |
| **7** | **Dentist Implants Problem** | **22** |
|  |  |  |
| **7.1** | **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.?** | **24** |
| **7.2** | **Before the hypotheses may be tested, state the required assumptions. Are the assumptions fulfilled? Comment separately on both alloy types.?** | **26** |
| **7.3** | **Irrespective of your conclusion in 7.2, 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?** | **27** |
| **7.4** | **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?** | **28** |
| **7.5** | **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?** | **29** |
| **7.6** | **Consider the interaction effect of dentist and method and comment on the interaction plot, separately for the two types of alloys?** | **30** |
| **7.7** | **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?** | **31** |

**List of Figures:**

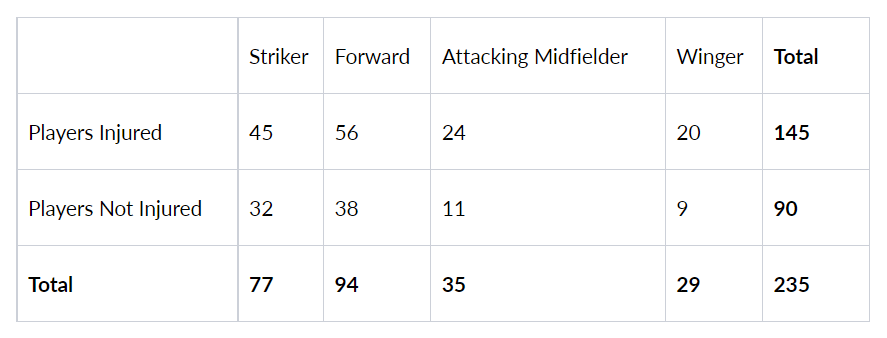
|  |  |  |
| --- | --- | --- |
| **1** | **Breaking strength less than 3.17 kg per sq cm** | **12** |
| **2** | **Breaking strength at least 3.6 kg per sq cm** | **13** |
| **3** | **breaking strength between 5 and 5.5 kg per sq cm.** | **14** |
| **4** | **breaking strength NOT between 3 and 7.5 kg per sq cm.** | **15** |
| **5** | **randomly chosen student gets a grade below 85 on this exam** | **16** |
| **6** | **a randomly selected student scores between 65 and 87** | **17** |
| **7** | **Interaction between Method, Response, Alloy** | **30** |
| **8** | **Interaction between Dentist, Response, Alloy** | **31** |
| **9** | **Interaction between Dentist, Method, Alloy A** | **32** |
| **10** | **Interaction between Dentist, Method, Alloy B** | **34** |

**List of Tables:**

|  |  |  |
| --- | --- | --- |
| **1** | **First five rows of Zingaro stone printing data** | **18** |
| **2** | **Basic information on Zingaro stone printing data** | **18** |
| **3** | **First five rows of Aquarius health data** | **20** |
| **4** | **Basic information on Aquarius health data** | **21** |
| **5** | **Head of the dental implant data** | **22** |
| **6** | **Describe information of the dental implant data** | **23** |
| **7** | **Basic information on the dental implant data** | **23** |
| **8** | **ANOVA table for Response and Dentist** | **24** |
| **9** | **ANOVA table for Dentist with Alloy A** | **25** |
| **10** | **ANOVA table for Dentist with Alloy B** | **25** |
| **11** | **ANOVA table for Method with Alloy A** | **28** |
| **12** | **ANOVA table for Method with Alloy B** | **29** |
| **13** | **ANOVA table for Temperature with Alloy A** | **29** |
| **14** | **ANOVA table for Temperature with Alloy B** | **30** |
| **15** | **ANOVA table for Method and Dentist with Alloy A** | **31** |
| **16** | **ANOVA table for Method and Dentist of Alloy A with interaction effect** | **32** |
| **17** | **ANOVA table for Method and Dentist with Alloy B** | **33** |
| **18** | **AN1OVA table for Method and Dentist of Alloy B with interaction effect** | **34** |

**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.**



* 1. **What is the probability that a randomly chosen player would suffer an injury?**

The probability of Players being Injured is to be found.

To calculate Players Injured we should calculate the total number of players who are injured to the total number of players.

**P (Players Injured) = Total Players Injured / Total Players**

**From the data,**

**Total Players Injured = 145**

**Total players = 235**

**P (Players Injured) = 145 / 235 => 0.6170212765957447**

**Rounding the value gives 0.62**

**P (Players Injured) = 0.62**

**The probability of a randomly chosen player would suffer an injury is 0.62**

* 1. **What is the probability that a player is a forward or a winger?**

The Probability of a player being a forward or a winger is to be calculated.

This can be done by:

Forward and Winger are independent variables.

**P (Forward U Winger) = P(Forward) + P(Winger) – P (Forward n Winger)**

**P (Forward n** Winger**) = 0, as Forward and Winger are mutually exclusive events.**

**P (Forward or Winger) = P(Forward) + P(Winger)**

P(forward) = Total forward players/Total number of players => 94 / 235

P(winger) = Total winger players/Total number of players => 29 / 235

**P (Forward U Winger) = 94 / 235 + 29 / 235 => 0.5234042553191489**

**Rounding the value gives 0.52**

**P (Forward U Winger) = 0.52**

**The probability of a player being a forward or a winger is 0.52**

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

The probability of a player being a striker and having an injury is to be calculated.

This can be done by dividing the number of injured striker players by the total number of players.

**P (Injured Player, striker) = Total number of injured striker players / Total number of players**

This is a Joint Probability

The total number of injured striker players from the table is 45

Total number of players = 235

**P (Injured Player, striker) = 45/235 => 0.19**

**The probability of a player being a Striker and having an Injury is 0.19**

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

Probability of an injured person in which person being a striker

The randomly chosen player is an injured person and the player is from striker.

P (injured player | striker) = P (Injured player, Striker) / P(Striker)

**P (Injured Player, striker) = Total number of injured striker players / Total number of players**

This is a Conditional Probability.

P (Injured Player, striker) = 45 / 235

P (Injured Player, striker) = 77 / 235

**P (injured player | striker) = 45 / 77 => 0.58**

**The probability of a player being an Injured player and from a Striker is 0.58**

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

The probability that a randomly chosen injured player is either a forward or an attacking midfielder is to be calculated.

This can be done by:

Forward and attacking midfielders two are independent variables.

**P (forward | injured) + P (attacking midfielder | injured**)

P (injured forward or injured attacking midfielder) = P (injured, forward) + P (injured, attacking midfielder)

P (forward, injured) = Total injured forward players/Total number of injured players= P (forward, injured) = 56/145

P (attacking midfielder, injured) = Total injured attacking midfielder players/Total number of injured players

P (attacking midfielder, injured) = 24/145P (injured forward or injured attacking midfielder) = 56/145 + 24/145

**The Probability to be injured player from a forward or attacking midfielder is 0.55**

## **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 radiation due to fire, mechanical failure, and human error.

**P (radiation, fire)** = Probability of radiation when fire = 20% => **0.2**

**P (radiation, mechanical)** = Probability of radiation when mechanical failures = 50% => **0.5**

**P (radiation, human)** = Probability of radiation when human errors = 10% => **0.1**

**P (fire, radiation)** = probability of radiation leak occurring with fire = 0.1% => **0.001**

**P (mechanical, radiation)** = probability of radiation leak occurring with mechanical failure = 0.15% **=> 0.0015**

**P (human, radiation)** = probability of radiation leak occurring with human error = 0.12% => **0.0012**

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

Given that the probabilities are in the percentages.

Let’s calculate the probabilities of fire, a mechanical failure, and a human error as P(x) and x be fire or mechanical or human error.

P(fire) = P (fire, radiation) / P (radiation, fire) = 0.001 / 0.2

**P(fire) = 0.005**

P(mechanical) = P (mechanical, radiation) / P (radiation, mechanical) = 0.0015 / 0.5

**P(mechanical) = 0.003**

P(human) = P (human, radiation) / P (radiation, human) = 0.0012 / 0.1

**P(human) = 0.012**

### 2.2 What is the probability of a radiation leak?

Here we are about to find the probability of radiation leak P(radiation).

To calculate no. of times the radiation leak is done by multiplying P(no\_of\_Accidents) and P (radiation, no\_of\_Accidents).

P (radiation, no\_of\_Accidents) = 0, as that no 2 or more accidents never occur together, so that probability will be 0 because you can't have all 3 of them acting at the same time.

This will 0 obviously as this is multiplied by P (r, no\_of\_Accidents) which is 0.

Therefore, we can do this by adding all the probabilities that we have for the radiation leak P(radiation).

This can be done by summing all the probabilities of causing radiation, i.e., radiation with fire, radiation with mechanical failures, and radiation with human errors.

P (radiation leak) = P (radiation, fire) + P (radiation, mechanical) + P (radiation, human).

P (radiation leak) = 0.001 + 0.0015 + 0.0012 => 0.0037

The probability of radiation leak P(radiation) is 0.0037

### 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.**

The probability of radiation caused by fire, mechanical failures, human errors is to found.

This is conditional probability.

The radiation is caused given that by fire

**P (fire | radiation) = P (fire, radiation)/P(radiation**)

**P (fire | radiation) = 0.001 / 0.0037 => 0.2702702702702703**

**The probability of a radiation leak in the reactor for which the cause is fire is 0.27**

The radiation is caused given that by mechanical failure

**P (mechanical | radiation) = P (mechanical, radiation)/P(radiation)**

**P (mechanical | radiation) = 0.0015 / 0.0037 => 0.4054054054054054**

**The probability of a radiation leak in the reactor for which the cause is mechanical failure is 0.41**

The radiation is caused given that by human error

**P (human | radiation) = P (human, radiation)/P(radiation)**

**P (human | radiation) = 0.0012 / 0.0037 => 0.3243243243243243**

**The probability of a radiation leak in the reactor for which the cause is human error is 0.32**

### 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 that, mean = 5

standard deviation = 1.5

To calculate this, we will calculate the proportion of gunny bags breaking strength by cumulative probability (CDF)

CDF is the probability function that X will take a value less than or equal to x

𝐹𝑋(𝑥)=𝑃(𝑋≤𝑥)

𝐹𝑋(𝑥) = function of X

X = real value variable

P = probability that X will have a value less than or equal to x

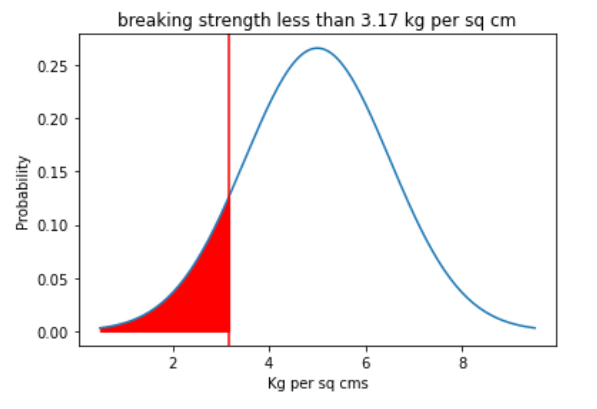
### 3.1 What proportion of the gunny bags has a breaking strength of less than 3.17 kg per sq cm?

The proportion of the gunny bags having a breaking strength of less than 3.17 kg per sq cm is to be calculated.

**P (x ≤ 3.17) = 0.1112**

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

**Fig 1: Breaking strength less than 3.17 kg per sq cm**

****

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

The proportion of the gunny bags having a breaking strength of at least 3.6 kg per sq cm is to be calculated.

That is greater than 3.6 kg per sq cm.

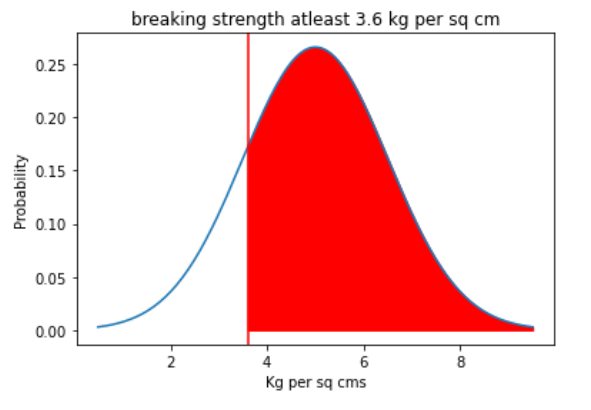
**P (x ≥ 3.17) = 1 - P (x ≤ 3.17)**

P (x ≥ 3.17) = 1 – 0.17532

**P (x ≥ 3.17) = 0.8246**

**The proportion of the gunny bags have a breaking strength at least 3.6 kg per sq cm is 82.47** **%**

**Fig 2: 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.?

The proportion of the gunny bags having a breaking strength between 5 and 5.5 kg per sq cm is to be calculated.

That is above 5 and below 5.5 kg per sq cm

**P (5 ≤ x ≤ 5.5) = P (x ≤ 5.5) – P (x ≤ 5)**

The proportion of the gunny bags have a breaking strength below 5.5 kg per sq cm is 63.06 %

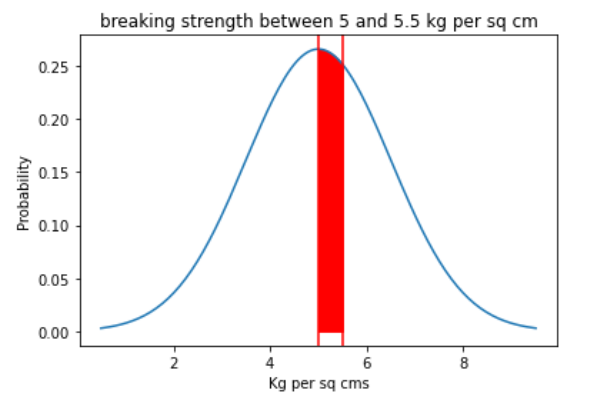
The proportion of the gunny bags have a breaking strength below 5 kg per sq cm is 50.0 %

P (5 ≤ x ≤ 5.5) = 0.63055 – 0.5

**P (5 ≤ x ≤ 5.5) = 0.1306**

**The proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm is 13.06 %**

**Fig 3: 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.?

From the above problem, we are asked to find the not between 3 and 7.5 kg per sq. cm that is we should find less than 3 and greater than 7.5 kg per sq. com

To do this we calculate the cumulative probability of 3 and 1-(CDF (7.5)) gives the probability of bags having a breaking strength NOT between 3 and 7.5 kg per sq cm.

That is below 3 and above 7.5 kg per sq cm

P (x ≤ 3) = 0.09121

**The proportion of the gunny bags that have a breaking strength of less than 3 kg per sq cm is 9.12 %**

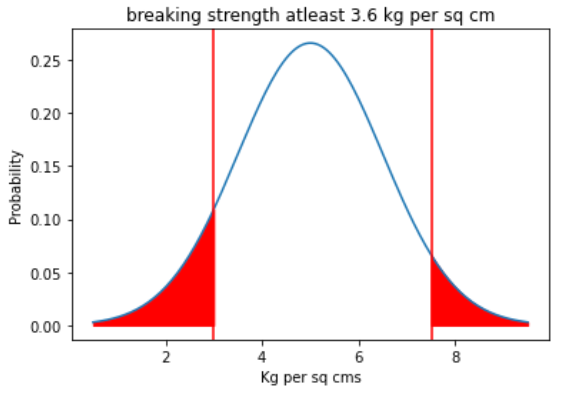
P (x ≥ 7.5) = 1 - P (x ≤ 7.5)

P (x ≥ 7.5) = 0.478

**The proportion of the gunny bags that have a breaking strength of more than 7.5 kg per sq cm is 4.78 %**

**The proportion of the gunny bags that have a breaking strength NOT between 3 and 7.5 kg per sq cm is 9.12 % and 4.78 %**

**Fig 4: 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.

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

The probability of a randomly chosen student getting grade below 85 can be calculated by:

**1. Cumulative Distribution Function (CDF):**

The Cumulative Distribution Function (CDF) of a real-valued random variable X, evaluated at x, is the probability function that X will take a value less than or equal to x. It is used to describe the probability distribution of random variables in a table.

**P (x ≥ 85) = 0.83**

**2. By Z score:**

𝑧=(𝑥−𝜇) / 𝜎

The z-score value is from 0. The mean is in the middle of the normal distribution, so the z score for 85 will be from 77 to 85. We also need the area under 77 assuming 0. As the grades are normally distributed, the other half would be 0.5.

So, we should add 0.5 + z-score.

Mean = 77

Standard Deviation = 8.5

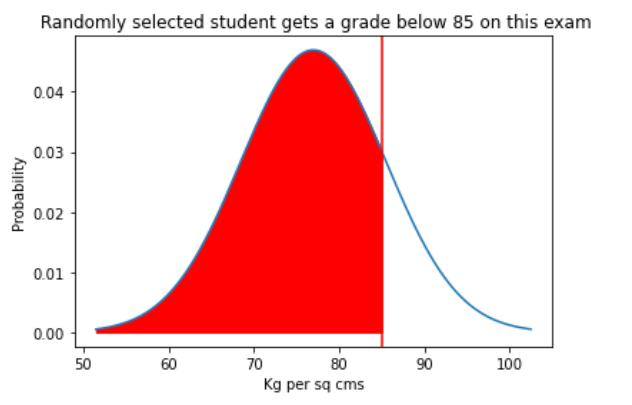
z = (85-77) / 8.5

The z is 0.94

The z-score value of 0.94 is 0.3264

**The probability of a randomly chosen student getting a grade below 85 on this exam is 0.83**

**Fig 5: Randomly chosen student getting a grade below 85 on this exam**

****

### 4.2 What is the probability that a randomly selected student scores between 65 and 87?[¶](http://localhost:8888/notebooks/great%20learning/Advance_statistics/Advance%20Statistics%20project/sahithi_kocharlakota.ipynb#4.2-What-is-the-probability-that-a-randomly-selected-student-scores-between-65-and-87?)

The proportion of a randomly selected student score between 65 and 87 is to be calculated.

That is above 65 and below 87 scores.

**P (65 ≤ x ≤ 87) = P (x ≤ 87) – P (x ≤ 65)**

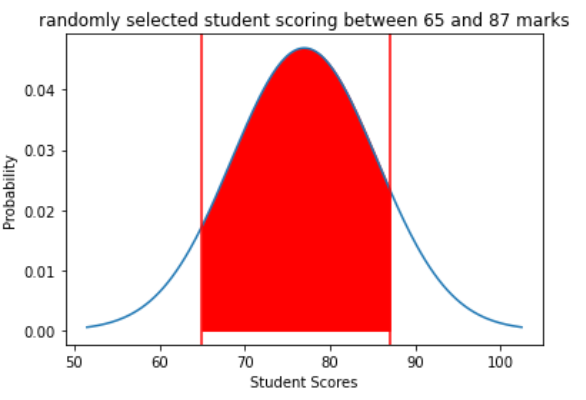
The probability of a randomly selected student scoring below 65 is 0.079

The probability of a randomly selected student scoring below 87 is 0.88

**P (65 ≤ x ≤ 87) = 0.079 – 0.88 => 0.801**

**The probability of a randomly selected student scores between 65 and 87 is 0.801**

**Fig 6: Randomly selected student scores between 65 and 87**

****

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

Here we have to find the cut-off for 75% students to clear the exam

This can be done by percent point function (PPF) (or inverse cumulative distribution function)

**percent point function (PPF):** returns the value x of the variable that has a given cumulative distribution probability (CDF)

Here CDF is 75% => 0.75

Mean = 77

Standard Deviation = 8.5

Cut-off = 82.77

**The cut-off marks so that 75% of the students clear the exam is 83**

## 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);[¶](http://localhost:8888/notebooks/great%20learning/Advance_statistics/Advance%20Statistics%20project/sahithi_kocharlakota.ipynb#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);)

To check the data, load the packages and data file.

The data has 75 rows and 2 columns. For this we use shape.

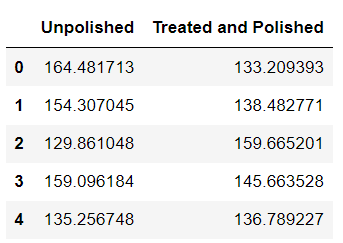
To get a clear view of the data, we use a head () function.

For a quick look at the data information, we use the info () function. From this, we can get basic information about rows, columns, and null values.

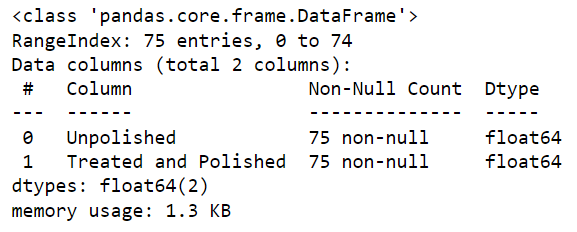
From the info (), we find there are no null values.

Let’s check again with isnull(). sum() to get more clarity on the data.

**Table 1: First five rows of the data**



**Table 2: Basic information**



### 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?

According to Zingaro the unpolished stones may not be suitable for printing.

Checking the describe () output for Unpolished,

We can see that the mean of the Brinell's hardness for Unpolished stones is around 134 which is less than 150.

There are 75 Unpolished stones in which there are only 24 stones with hardness above 150.

This might be the reason why Zingaro has to believe that the unpolished stones may not be suitable for printing.

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

To check the means hardness of polished and unpolished stones to be same or different, we use t test.

The statistic t value can be found by

t = xdiff / (sdiff/√n)

xdiff: sample mean of the differences s: sample standard deviation of the differences n: sample size (i.e., number of pairs)

Degree of freedom (df)= n-1 =>100-1

df = 99

Level of significance is 5%

i.e., alpha(α) = 0.05

**Null hypothesis (H0): There is no difference in the means of the Unpolished and Polished stones.**

**Alternate hypothesis (Ha): At least one of the means is different for Unpolished and Polished stones.**

The t value is -3.5589113215869057

The p-value is =0.0006545976110249849

**The p-value is less than alpha (α==0.05). So, we reject the NULL HYPOTHESIS**

**Therefore, we conclude that means of the Unpolished and Polished stones are different.**

### 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.

To check the data, we should upload the data which is in csv.

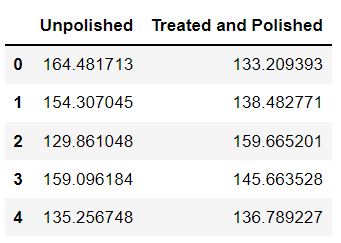
After uploading data, we should first check what is the content of the data. To check this, we have head () function.

We should know the shape of the data before proceeding to the further process so we use shape attribute to get number of rows and columns in the data set.

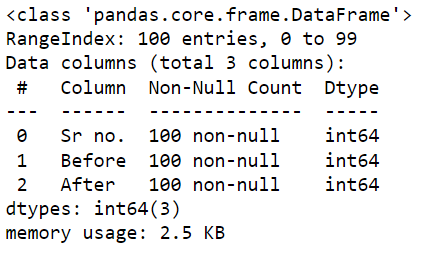
Next step is to check the description of the data which given mean, Standard deviation, 25%, Median (50%),75%, minimum, maximum values.

To get a clear approach about the data regarding the datatypes, columns, null values we use info () function.

**Table 3: First five rows of Aquarius health club data**



**Table 4: Basic information about the Aquarius health club data**



We use paired t test for this business problem to be solved.

Paired t-test is the type of t-test that we apply when we want to explore whether the two means of two related samples are significantly different.

For paired t-test always uses the following null hypothesis:

𝜇μ1= 𝜇μ2

The alternative hypothesis can be either two-tailed, left-tailed, or right-tailed:

H1 (two-tailed): μ1 ≠ μ2 (Both population means are not equal)

H1 (left-tailed): μ1 < μ2 (population 1 mean is less than population 2 mean)

H1 (right-tailed): μ1> μ2 (population 1 mean is greater than population 2 mean)

**In this case it is two tailed.**

**Null hypothesis (H0): There is no difference in the number of pushups before and after the cross-fit gyms program.**

**Alternate hypothesis (Ha): There is difference in the number of pushups before and after the cross-fit gyms program.**

**T statistics = 5.550/ (2.872/100\*\*0.5) => 19.138**

**The p value is less than 0.0001 so we reject the** null hypothesis.

Since the p-value is less than 0.05, we reject the null hypothesis. We have sufficient evidence to say that the mean count of pushups before and after enrolling in cross-fit gym are different.

From this, we can conclude that the program was 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.

To check the data, we should upload the data which is in excel.

After uploading data, we should first check what is the content of the data. To check this, we have head () function.

### Table 5: Head of the dental implant data

### 

We should know the shape of the data before proceeding to the further process so we use shape attribute to get number of rows and columns in the data set.

From this we can see that the data set has \*\*106 rows and 14 columns\*\*.

The next step is to check the description of the data which given mean, Standard deviation, 25%, Median (50%),75%, minimum, maximum values. But we can see that Dentist, Method, Alloy is having the above description which are actually not numeric but categorical.

### Table 6: Describe information of the dental implant data

### 

To get a clear approach about the data regarding the datatypes, columns, null values we use info () function.

### Table 7: The basic information of the dental implant data

### 

Have observing the information from the info () function we see that Dentist, Method, Alloy is in numerical data type which are in encoded form. So must be in Categorical data type.

To change the data types of these columns we use pd.Categorical() function.

From the info () output we have also seen few null values in some columns. These are to be treated.

If the data is very large number of null values, it is better to drop those columns.

We confirmed that there are null values in the data set. So, let's find the total number of null values in each column by isnull(). sum() functions.

From this, we see that Unnamed: 5 and 6 have 107 null values that is those columns are filled with null values. Also Unnamed:12 and 13 have more than 100 null values.

But from the business problem that we have, we can drop the all-Unnamed columns as there is not much useful information.

Now we should deal with the null values in the Dentist, Method, Alloy, Temp, and, Response which have 17 null values.

As Dentist, Method, and Alloy are categorical values, we can use mode () to fill in the null values, and Temp, and Response are numerical values so mean () we can be used to fill in null values.

### 7.1 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.?

Points from the business problem to remember:

The implant hardness depends on the method of implant, the temperature at which the metal is treated, the alloy used, and dentists.

The Response is the variable of interest.

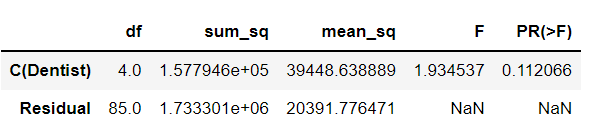
The two types of alloys cannot be considered together.

There are five different dentists.

**Null hypothesis (H0): There is no difference in the means of implant hardness among 5 different Dentists**

**Alternate hypothesis (Ha): At least one of the means is different for among different Dentists.**

**Table 8: ANOVA table for Response and Dentist**

****

**The corresponding p-value for Dentist with respect to response is more than alpha (𝛼 = 0.05).**

**Thus, we fail to reject NULL HYPOTHESIS (𝐻0)**

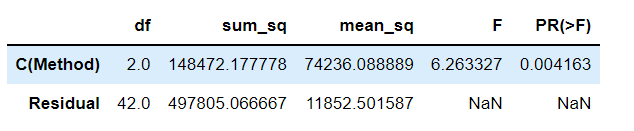
Hence, we could say that there is no difference in the hardness of the implants made among different dentists.

There is no impact by different Dentists on the implant hardness.

**Null hypothesis for alloy A (H0\_A) = There is no difference in the means of implant hardness for alloy A**

**Alternate hypothesis for alloy A (Ha\_A) = At least one of the means is different for alloy A**

**Table 9: ANOVA table for Dentist with Alloy A**

****

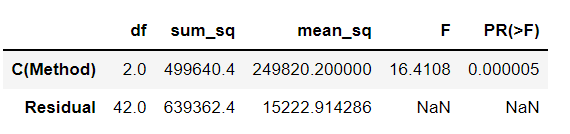
**The corresponding p-value for Dentist for Alloy A with respect to response is less than alpha (𝛼 = 0.05). Thus, we reject NULL HYPOTHESIS (𝐻0)**

That is, we could say that there is a difference in the hardness of the implants made among different dentists when Alloy A is used.

**Null hypothesis for alloy B (H0\_B) = There is no difference in the means of implant hardness for alloy B**

**The alternate hypothesis for alloy B (H0\_B) = At least one of the means is different for alloy B**

**Table 10: ANOVA table for Dentist with Alloy** **B**

****

**The corresponding p-value for Dentist for Alloy\_B with respect to response is less than alpha (𝛼 = 0.05). Thus, we reject NULL HYPOTHESIS (𝐻0)**

That is, we could say that there is difference in the hardness of the implants made among different dentists when Alloy B is used.

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

The Assumptions are:

1. Scale of measurement: The data collected is continuous. Here the Response is continuous variable.
2. Simple random sample: Randomly selected portion
3. The data must be normally distributed.
4. Homogeneity of variance

The **Shapiro-Wilk** **test** tests the null hypothesis that the data was drawn from a normal distribution.

alpha(𝛼α) = 0.05

if p-value is less than 0.05 we say null hypothesis is passed.

**The Null Hypothesis (H0): The Sample data is normally distributed data**

**The Alternate Hypothesis (Ha): The Sample data is not from normal distribution data**

The Anderson Darling test is used when Shapiro test is failed.

if p-value is greater than 0.05 we say null hypothesis is passed**.**

**The p-value of most of the samples are less than the alpha value (𝛼 = 0.05).**

**We reject Null hypothesis**. The data is not normally distributed.

**The Shapiro test for two types Alloys**:

For the ALLOY A: The p-value (1.1945) is high than the alpha (α = 0.05).

Therefore, for **ALLOY A we fail to reject Null Hypothesis.**

For the ALLOY B: The p-value (0.00040) is less than the alpha (α = 0.05).

Therefore, for **ALLOY B we reject Null Hypothesis**.

**From this, ALLOY A is not normally distributed and ALLOY B is normally distributed.**

One of the assumptions is failed when compared all the variables.

Next assumption is Homogeneity in variances

The**Levene test**tests the null hypothesis that all input samples are from populations with equal variances.

alpha (α) = 0.05

**The Null Hypothesis (H0): The Sample data has equal variances**

**The Alternate Hypothesis (Ha): The Sample data has no equality in variances.**

The p-value of two samples is less than the alpha value (α = 0.05).

**We reject Null hypothesis**. The data is don’t have equal variances.

**The Levene test for Two types of Alloys:**

**The p-value for alloys is (0.23669) which means that the alloys have equality in the variances.**

Another assumption is failed when compared all the variables.

**The main assumptions of the ANOVA test failed. The sample is not normally distributed and has no equality in the variance.**

### 7.3 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?

p-value (0.112) > alpha (α = 0.05)

No, the null hypothesis is not rejected.

That is, we could say that there is no difference in the hardness of the implants made among different dentists.

The means are the same for all the different dentists.

There is no effect on the implant hardness by the dentists.

**When the considering the two types of Alloys separately We can conclude that we reject Null Hypothesis as the p-value is less than alpha (α = 0.05)**

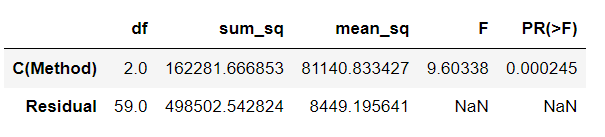
That is there is some impact in the hardness of the implant among the dentist and the alloys they use.

### 7.4 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?

Null hypothesis for alloy A (𝐻0\_A) = There is no difference among the methods in the means of implant hardness for alloy A

Alternate hypothesis for alloy A (𝐻𝑎\_A) = At least one of the means among the methods is different for alloy A

**Table 11: ANOVA table for method with Alloy A**



**The corresponding p-values for both Method for Alloy A with respect to response are less than alpha (**𝛼α**= 0.05). Thus, we reject NULL HYPOTHESIS (**𝐻0**)**

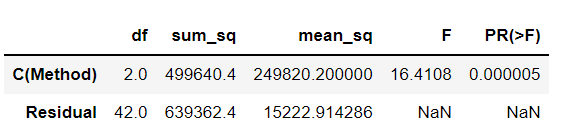
That is there is difference in the implant hardness when different methods and different alloys are used.

When Alloy A is used there is difference in the hardness of the implant.

Null hypothesis for alloy B (𝐻0\_B) = There is no difference among the methods in the means of implant hardness for alloy B

Alternate hypothesis for alloy B (𝐻𝑎\_B) = At least one of the means among the methods is different for alloy B

**Table 12: ANOVA table for Methods with Alloy B**



**The corresponding p-values for both Method for Alloy B with respect to response are less than alpha (**𝛼α**= 0.05). Thus, we reject NULL HYPOTHESIS (**𝐻0**)**

That is there is difference in the implant hardness when different alloys are used.

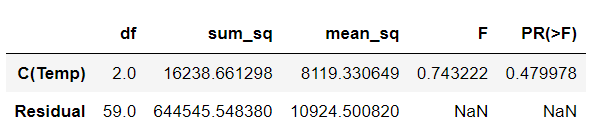
When Alloy B is used there is difference in the hardness of the implant.

### 7.5 Now test whether there is any difference among the temperature levels on the hardness of dental implants, 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 for alloy A (𝐻0\_A) = There is no difference among the temperatures levels in the means of implant hardness for alloy A

Alternate hypothesis for alloy A (𝐻𝑎\_A) = At least one of the means among the temperature levels is different for alloy A

**Table 13: ANOVA of Temperature with Alloy A**



**The corresponding p-values for Temperature for Alloy A with respect to response is greater than alpha (**α**= 0.05). Thus, we fail to reject NULL HYPOTHESIS (**𝐻0𝐻0**)**

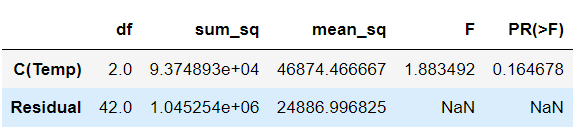
That is there is no difference in the temperature for the implant hardness when different alloys are used.

When Alloy A is used there is no impact on the implant hardness due to temperature.

Null hypothesis for alloy B (𝐻0\_B) = There is no difference among the temperature levels in the means of implant hardness for alloy B

Alternate hypothesis for alloy B (𝐻𝑎\_B) = At least one of the means among the temperature levels is different for alloy B

**Table 14: ANOVA of Temperature with Alloy B**



**The corresponding p-values for Temperature for Alloy B with respect to response is greater than alpha (**α**= 0.05). Thus, we fail to reject NULL HYPOTHESIS (**𝐻0𝐻0**)**

That is there is no difference in the temperature for the implant hardness when different alloys are used.

When Alloy B is used there is no impact on the implant hardness due to the temperature

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

### Fig 7: Interaction between Method, Response, Alloy

### 

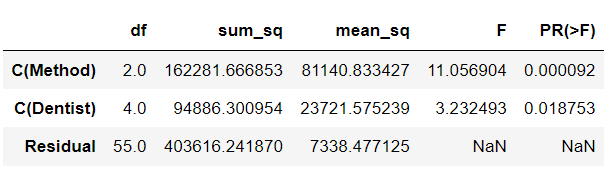
### Fig8: Interaction between Dentist, Response, Alloy

### 7.7 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?

**Null hypothesis for alloy A (𝐻0 A) = There is no difference among the dentist and methods in the means of implant hardness for alloy A**

**Alternate hypothesis for alloy A (𝐻𝑎 A) = At least one of the means among the dentist and methods is different for alloy A**

**Table 15: ANOVA table for Method and Dentist with Alloy A**

****

**Considering the effect of Dentist and method for two alloys separately, we can conclude that the p-value is less than the alpha(**𝛼α**= 0.05).Thus we reject NULL HYPOTHESIS (**𝐻0𝐻0**)**

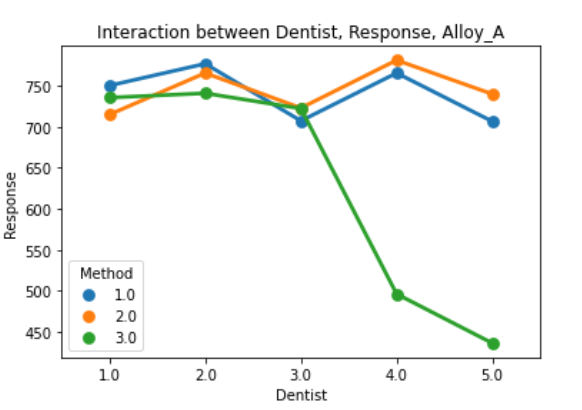
There is some impact in the hardness of the implant due to difference among the dentists and methods when Alloy A used.

**The p-value in the one of the treatments is greater than**𝛼α**(0.05)**

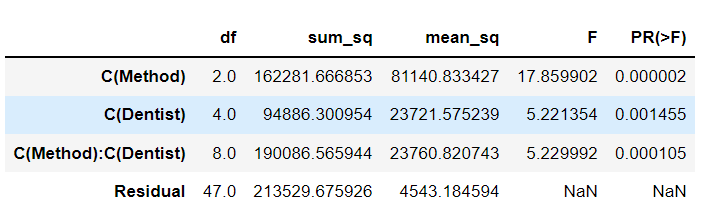
The p-value of Dentist is greater than alpha and Method is less than alpha.

**Let us check whether there is any interaction effect.**

**Fig 9: Interaction Effect between Method, Dentist, Alloy**

****

**Table 16: ANOVA table for Method and Dentist of Alloy A with interaction effect**

****

**Considering the effect of Dentist and method for two alloys separately, we can conclude that the p-value is less than the alpha (α = 0.05). Thus we reject NULL HYPOTHESIS (𝐻0)**

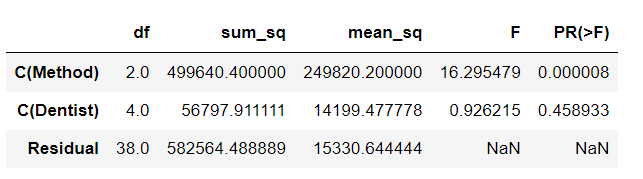
**There is some impact on the hardness of the implant due to differences among the dentists and methods when Alloy A used.**

There is more interaction among dentists, and methods with alloys A.

Null hypothesis for alloy B (𝐻0\_B) = There is no difference between the dentist and methods in the means of implant hardness for alloy B

Alternate hypothesis for alloy B (𝐻𝑎\_B) = At least one of the means among the dentist and methods is different for alloy B

**Table 17: ANOVA table for for Method and Dentist of Alloy B**

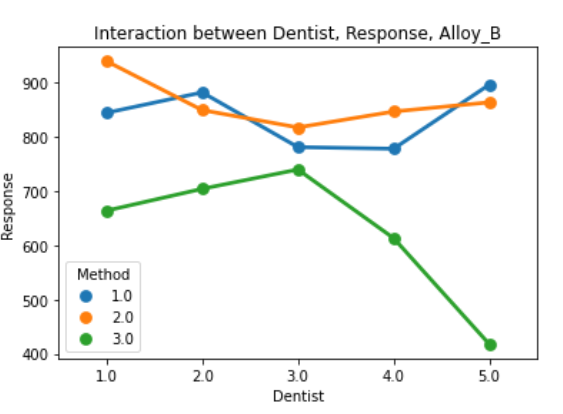


The p-value in the one of the treatments is greater than α(0.05)

The p-value of Dentist is greater than alpha and Method is less than alpha.

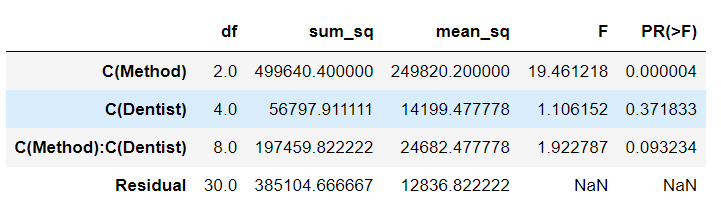
**Let us check whether there is any interaction effect between the treatments.**

**Fig 10: Interaction Effect between Method, Dentist, Alloy**



**The interaction between the methods 1,2 and dentist with alloy B can be seen**.

**Table 18: ANOVA table for for Method and Dentist of Alloy B with interaction effect**



The P-value of the Dentist is greater than alpha and the P-value of the Method is less than alpha.

By checking the P-value of the interaction effect between the Dentist and Method is more than alpha.

**So, we fail to reject Null Hypothesis in this case that id for Alloy B**.

Therefore, we can conclude that - Dentist and method variables

**For Alloy A**: We reject Null Hypothesis – There is some impact on the implant hardness due to the factors – Dentist, Method, Alloy A.

**For Alloy B**: We fail to reject Null Hypothesis – There is no impact on the implant hardness due to the factors – Dentist, Method, Alloy B.

**Due to the inclusion of the interaction effect term, we can see a slight change in the p-value of the Dentist and Method as compared to the Two-Way ANOVA without the interaction effect terms.**

**We see that the p-value of the interaction effect term of 'Dentist' and 'Method' suggests that the Null Hypothesis is rejected in the case with Alloy\_B and Null Hypothesis is accepted in case of Alloy\_B**.