2022

ADVANCED STATISTICS GRADED PROJECT REPORT

DSBA

Girish Chadha

16/10/2022

Contents

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?

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?

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?

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

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?

EXECUTIVE SUMMARY

Problem 1

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

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

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

Solution : Using Marginal Probabilities

P(Injured) = 145/235 = 0.617

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

Solution : Using Marginal Probabilities

P(Forward U Winger) = 94/235 + 29/235 = 0.523

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

Solution : Using Marginal Probabilities

P(Striker ∩ Foot Injury) = 45/235 = 0.191

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

Solution : Using Marginal Probabilities

P(Striker / Injured) = 45/145 = 0.310

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

Solution : Using Marginal Probabilities

P(Forward/Injured) + P(Attacking midfielder/Injured) = 56/145 + 24/145 = 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:

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

<u>Solution</u> : <u>GIVEN</u>: *P(R/F)=0.2 , P(R/M)=0.5, P(R/H)=0.1, P(R intersection F)=0.001, P(R intersection M)= 0.0015, P(R intersection H)=0.0012.* <u>Using Conditional probability-</u> >

The probabilities of a fire, a mechanical failure, and a human error respectively are 0.005,0.003 & 0.012.

$$P(F) = \frac{P(R \cap F)}{P(R|F)} = 0.001/0.2$$
$$P(F) = 0.005$$
$$P(M) = \frac{P(R \cap M)}{P(R|M)} = 0.0015/0.5$$
$$P(M) = 0.003$$
$$P(H) = \frac{P(R \cap H)}{P(R|H)} = 0.0012/0.1$$
$$P(H) = 0.012$$

2.2 What is the probability of a radiation leak?

<u>Solution</u> : P(R) = P(R/F)P(F) + P(R/M)P(M) + P(R/H)P(H) = (0.2)0.001/0.2 + (0.5)0.0015/0.5 + (0.1)0.0012/0.1

<u>Using total probability theorem,</u>

$$P(R) = P(R \cap F) + P(R \cap M) + P(R \cap H)$$

$$P(R) = 0.001 + 0.0015 + 0.0012$$
$$P(R) = 0.0037$$

Probability of a radiation leak 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.

Solution : Using Bayes' theorem, probability that it has been caused by a Fire, a Mechanical Failure , a Human Error is 0.270,0.405 & 0.324 respectively.

$$P(F|R) = \frac{P(R \cap F)}{P(R)} = 0.001/0.0037$$
$$P(F|R) = 0.27027027$$
$$P(M|R) = \frac{P(R \cap M)}{P(R)} = 0.0015/0.0037$$
$$P(M|R) = 0.40540541$$
$$P(H|R) = \frac{P(R \cap H)}{P(R)} = 0.0012/0.0037$$
$$P(H|R) = 0.32432432$$

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;

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

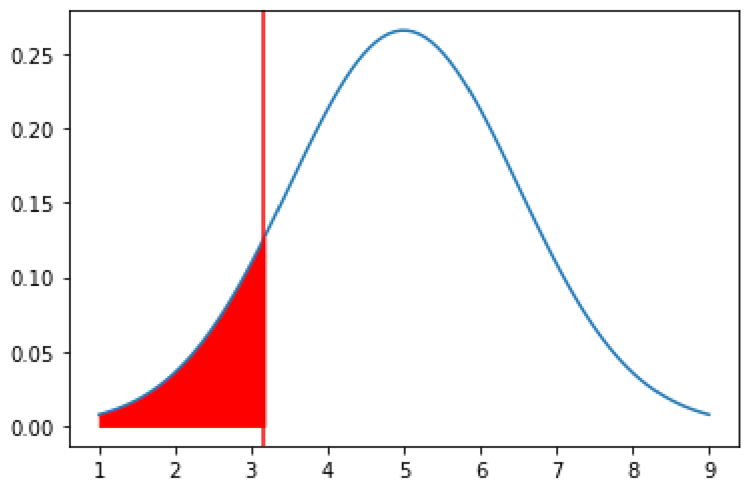
Using Standard Normal Distribution -

$\mu = 5$

$\sigma = 1.5$

Answer: The required probability is $p(X < 3.17)$.

$$
\begin{aligned}
p(X < 3.17) &= p\left(\frac{X-\mu}{\sigma} < \frac{3.17-5}{1.5}\right) \\
&= p\left(z < -1.22\right) \\
&= 0.5 - p(-1.22 \leq z < 0) \\
&= 0.5 - 0.3888(\text{Obtained from standard normal distribution table}) \\
&= 0.1112
\end{aligned}
$$



Therefore, 0.1112 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.?

Answer: The required probability is $p(X \geq 3.6)$.

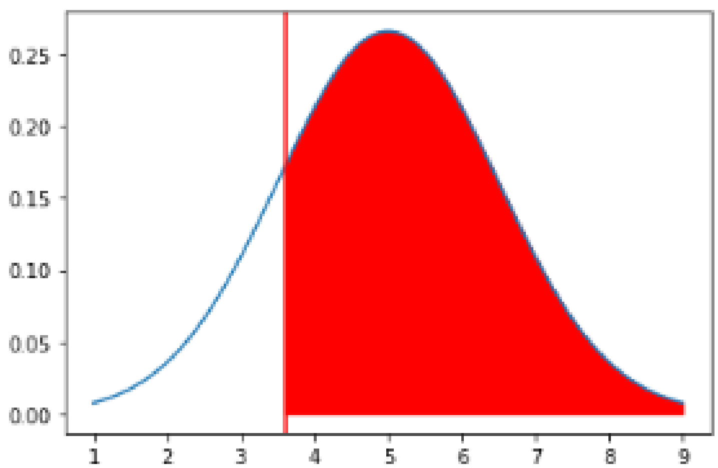
$$p(X > 3.6) = p\left(\frac{X-\mu}{\sigma} > \frac{3.6-5}{1.5}\right)$$

$$= p\,(z > -0.93)$$

$$= 0.5 + p(-0.93 \leq z < 0)$$

$$= 0.5 + 0.3238(\text{Obtained from standard normal distribution table})$$

$$= 0.8238$$



Therefore, 0.8238 proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm.

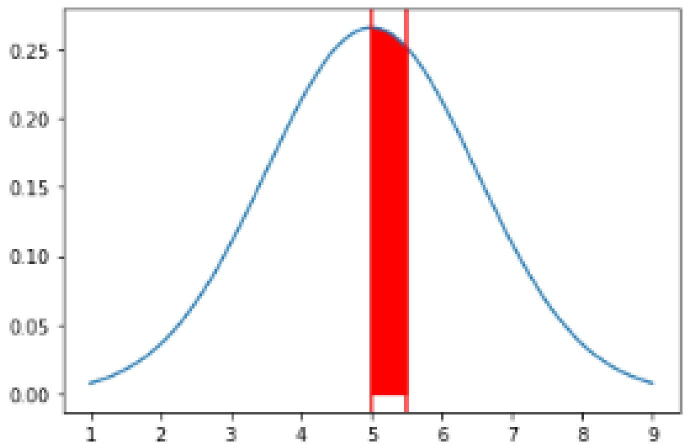
3.3 What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?

Answer: The required probability is $p(5 < X < 5.5)$.

$$p(5 < X < 5.5) = p\left(\frac{5-5}{1.5} < \frac{X-\mu}{\sigma} < \frac{5.5-5}{1.5}\right)$$

$$= p\,(0 < z < 0.33)$$

$$= 0.1293(\text{Obtained from standard normal distribution table})$$



Therefore, 0.1293~0.13 proportion of the 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.?

Answer: We first find the probability $p(3 < X < 7.5)$

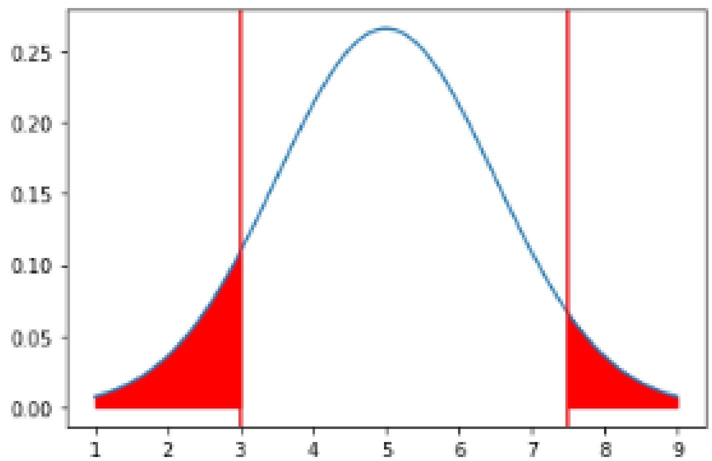
$$p(3 < X < 7.5) = p\left(\frac{3-5}{1.5} < \frac{X-\mu}{\sigma} < \frac{7.5-5}{1.5}\right)$$

$$= p\left(-1.33 < z < 1.67\right)$$

$$= p(-1.33 < z < 0) + p(0 < z < 1.67)$$

$$= 0.4082 + 0.4525$$

$$= 0.8607$$



Therefore, proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm is $= 1 - 0.8607 = 0.1393$.

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?

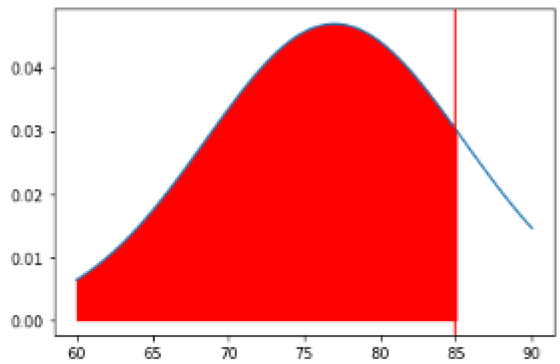
Using Standard Normal Distribution -

Mean = 77

Standard Deviation = 8.5

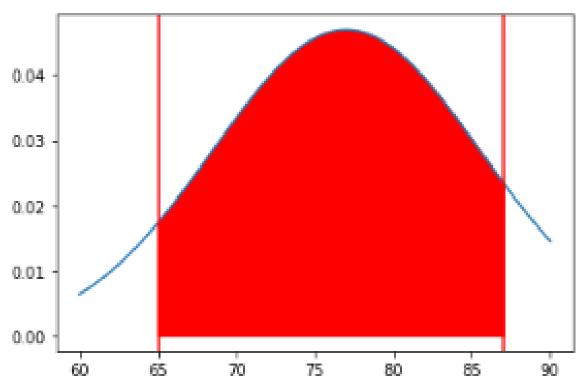
Using Python-

Probability that a randomly chosen student gets a grade below 85 on this exam is 0.826.



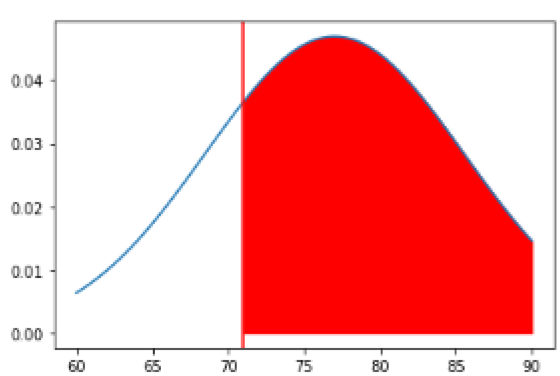
4.2 What is the probability that a randomly selected student scores between 65 and 87?

Probability that a randomly selected student scores between 65 and 87 is 0.801.



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

The passing cut-off so that 75% of the students clear the exam is 71.26~ 71.



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);

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?

<u>Solution 5.1</u>-

Reading the data

	Unpolished	TreatedandPolished
0	164.481713	133.209393
1	154.307045	138.482771
2	129.861048	159.665201
3	159.096184	145.663528
4	135.256748	136.789227

Describing the data

	Unpolished
count	75.000000
mean	134.110527
std	33.041804
min	48.406838
25%	115.329753
50%	135.597121
75%	158.215098
max	200.161313

Step 1: Define null and alternative hypotheses

In testing that unpolished stones may not be suitable for printing.Let μ′ be mean hardness index of unpolished stones

Null hypothesis states that mean hardness index of unpolished stones, μ′ is greater than or equals to 150.

Alternative hypothesis states that mean hardness index of unpolished stones, μ′ is less than 150.

$H0$: μ′ ≥ 150

HA: μ′ < 150

Step 2: Decide the significance level

Here we select α = 0.05.

The sample size for this problem is 75

We do not know the population standard deviation and n = 75. So we use the t distribution and tSTAT test statistic.

In Python, One sample T Test is implemented in ttest_1samp() function in the scipy package. However, it does a Two tailed test by default, and reports a signed T statistic. That means, the reported P-value will always be computed for a Two-tailed test. To calculate the correct P value, you need to divide the output P-value by 2.

Apply the following logic if you are performing a one tailed test:

For greater than test: Reject H0 if p/2 < alpha (0.05).

For lesser than test: Reject H0 if p/2 < alpha (0.05).

Using Python-

t Statistic: [-4.1646296]

P Value: [4.171287e-05]

Level of significance: 0.05

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

Our one-sample t-test p-value= [4.171287e-05]

Hence Zingaro is justified in thinking so Zingaro has reason to believe now that the unpolished stones may not be suitable for printing as mean hardness index of unpolished stones, μ′ is less than 150. So at 95% confidence level, there is sufficient evidence to prove that mean hardness index of unpolished stones is less than 150.

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

Solution - Describing the data

	Unpolished	TreatedandPolished
count	75.000000	75.000000
mean	134.110527	147.788117
std	33.041804	15.587355
min	48.406838	107.524167
25%	115.329753	138.268300
50%	135.597121	145.721322
75%	158.215098	157.373318
max	200.161313	192.272856

Step 1: Define null and alternative hypotheses

In testing whether the mean hardness index of unpolished stones is same as mean hardness index of unpolished stones, the null hypothesis states that the mean hardness index of unpolished stones is same as mean hardness index of unpolished stones, μA equals

μB. The alternative hypothesis states that the mean hardness index of unpolished stones is different from mean hardness index of unpolished stones, μA is not equal to μB.

$H0$: μA - μB = 0 i.e μA = μB

HA: μA - μB ≠ 0 i.e μA ≠ μB

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 use the t distribution and the $tSTAT$ test statistic for two sample unpaired test.

Step 4: Calculate the p - value and test statistic ** We use the scipy.stats.ttest_ind to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.**

** This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values.

Using Python-

```
Ttest_indResult(statistic=array([-3.24223205]), pvalue=array([0.00158838]))
```

two-sample t-test p-value= 0.00158838

We have enough evidence to reject the null hypothesis in favour of alternative hypothesis

We conclude that mean hardness index of unpolished stones is different from mean hardness index of unpolished stones.

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.

Solution - Reading the data

	Sr no.	Before	After
0	1	39	44
1	2	25	25
2	3	39	39
3	4	6	13
4	5	40	44

Describing the data

	Sr no.	Before	After
count	100.000000	100.000000	100.000000
mean	50.500000	26.940000	32.490000
std	29.011492	8.806357	8.779562
min	1.000000	3.000000	10.000000
25%	25.750000	21.750000	26.000000
50%	50.500000	28.000000	34.000000
75%	75.250000	32.250000	39.000000
max	100.000000	47.000000	51.000000

Step 1: Define null and alternative hypotheses

In testing if the candidate is able to do more than 5 push-ups after enroling in the program

The null hypothesis states that the the mean difference is 5

The alternative hypothesis is that the mean difference is less than 5

$H0$: μa - μb ≥ 5

HA: μa - μb < 5

Step 2: Decide the significance level

Here we select α = 0.05 as given in the question.

Step 3: Identify the test statistic

Sample sizes for both samples are same.

We have two paired samples and we do not know the population standard deviation.

So you use the t distribution and the $tSTAT$ test statistic for two sample paired test.

Step 4: Calculate the p - value and test statistic

We use the scipy.stats.ttest_rel to calculate the T-test on TWO RELATED samples of scores. This is a two-sided test for the null hypothesis that 2 related or repeated samples have identical average (expected) values. Here we give the two sample observations as input. This function returns t statistic and two-tailed p value.

Using Python -

```
tstat  -36.730
p-value for one-tail: 4.36643115031875e-60
```

As pvalue is less than 0.05. Thus, We have enough evidence to reject the null hypothesis in favour of alternative hypothesis. Hence we can conclude that the program is unsuccesful as the mean difference is less than 5.

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 favour one method above another and may work better in his/her favourite 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.?

For alloy of Type 1 :

H0 : Mean Implant hardness is same among all dentists for type 1 alloy

H1: Mean Implant hardness is different among dentists for type 1 alloy

Using one way ANOVA(F test) in python we obtain the following result:

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	106683.688889	26670.922222	1.977112	0.116567
Residual	40.0	539593.555556	13489.838889	NaN	NaN

For alloy of Type 2 :

H0 : Mean Implant hardness is same among all dentists for type 2 alloy

H1: Mean Implant hardness is different among dentists for type 2 alloy

Using one way ANOVA(F test) in python we obtain the following result:

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	5.679791e+04	14199.477778	0.524835	0.718031
Residual	40.0	1.082205e+06	27055.122222	NaN	NaN

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

Assumption of randomness is fulfilled by both alloy types A & B.

ANOVA Assumptions

• Randomness and Independence

• Select random samples from the c groups (or randomly assign the levels)

• Normality

• The sample values for each group are from a normal population

• Homogeneity of Variance

• All populations sampled from have the same variance

• Can be tested with Levene's Test

<u>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?</u>

FOR ALLOY TYPE 1 :

The p-value is 0.116 so we would fail to reject the null hypothesis. As p value is more than 0.05, this means we don't have sufficient evidence to say that there is a statistically significant difference between the group means thus implant hardness doesn't depend on dentists for alloy type 1.

FOR ALLOY TYPE 2 :

The p-value is 0.71 so we would fail to reject the null hypothesis. As p value is more than 0.05, this means we don't have sufficient evidence to say that there is a statistically significant difference between the group means thus implant hardness doesn't depend on dentists for alloy type 2.

Yes it is possible to identify which pairs of dentists differ if null hypothesis is rejected for that we have to do more analysis.

<u>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?</u>

For alloy of Type 1 :

H0 : Mean Implant hardness is same among all methods for type 1 alloy

H1: Mean Implant hardness is different among methods for type 1 alloy

Using one way ANOVA(F test) in python we obtain the following result:

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	148472.177778	74236.088889	6.263327	0.004163
Residual	42.0	497805.066667	11852.501587	NaN	NaN

For alloy of Type 2 :

H0 : Mean Implant hardness is same among all methods for type 2 alloy

H1: Mean Implant hardness is different among methods for type 2 alloy

Using one way ANOVA(F test) in python we obtain the following result:

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	499640.4	249820.200000	16.4108	0.000005
Residual	42.0	639362.4	15222.914286	NaN	NaN

Conclusion -

As P value is less than 0.05 for both types of alloys 1 & 2 thus we reject the null hypothesis as there is enough evidence to say that there is mean difference among the methods on hardness of dental implant. Yes it is possible to identify which pairs of methods differ for that we have to do more analysis.

<u>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?</u>

For alloy of Type 1 :

H0 : Mean Implant hardness is same among all Temperature for type 1 alloy

H1: Mean Implant hardness is different among Temperatures for type 1 alloy

Using one way ANOVA(F test) in python we obtain the following result:

	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2.0	10154.444444	5077.222222	0.335224	0.717074
Residual	42.0	636122.800000	15145.780952	NaN	NaN

For alloy of Type 2 :

H0 : Mean Implant hardness is same among all Temperatures for type 2 alloy

H1: Mean Implant hardness is different among Temperatures for type 2 alloy

Using one way ANOVA(F test) in python we obtain the following result:

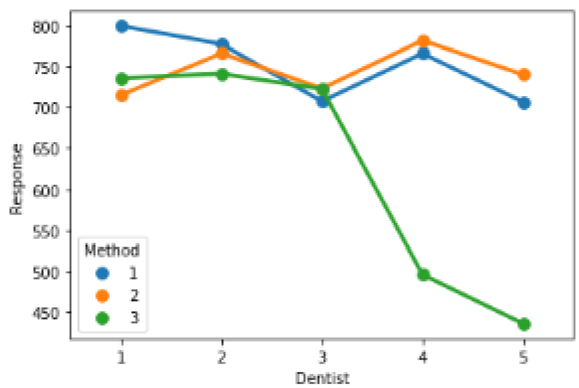
	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2.0	9.374893e+04	46874.466667	1.883492	0.164678
Residual	42.0	1.045254e+06	24886.996825	NaN	NaN

Conclusion -

As P value is more than 0.05 for both types of alloys 1 & 2 thus we accept the null hypothesis as there is not enough evidence to say that there is mean difference among the temperatures on hardness of dental implant. Yes if null hypothesis was rejected it is possible to identify which pairs of temperature differ for that we have to do more analysis.

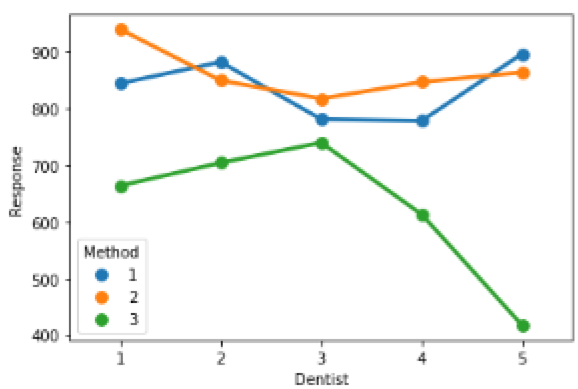
<u>6. Consider the interaction effect of dentist and method and comment on the interaction plot, separately for the two types of alloys?</u>

<u>FOR ALLOY TYPE 1</u>



We can see that there is some sort of interaction between the two treatments for alloy type 1. So, we will introduce a new term while performing the Two Way ANOVA.

FOR ALLOY TYPE 2



We can see that there is some sort of interaction between the two treatments for alloy type 2. So, we will introduce a new term while performing the Two Way ANOVA.

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?

FOR ALLOY TYPE 1

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	106683.688889	26670.922222	2.591255	0.051875
C(Method)	2.0	148472.177778	74236.088889	7.212522	0.002211
Residual	38.0	391121.377778	10292.667836	NaN	NaN

The p-value in the one of the treatments(Dentist) is greater than α (0.05) thus we accept the null hypothesis but in case of Method we reject the null hypothesis.

FOR ALLOY TYPE 2

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	56797.911111	14199.477778	0.926215	0.458933
C(Method)	2.0	499640.400000	249820.200000	16.295479	0.000008
Residual	38.0	582564.488889	15330.644444	NaN	NaN

The p-value in the one of the treatments(Dentist) is greater than α (0.05) thus we accept the null hypothesis but in case of Method we reject the null hypothesis.

We will introduce a new term while performing the Two Way ANOVA.

FOR ALLOY TYPE 1

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	106683.688889	26670.922222	3.899638	0.011484
C(Method)	2.0	148472.177778	74236.088889	10.854287	0.000284
C(Dentist):C(Method)	8.0	185941.377778	23242.672222	3.398383	0.006793
Residual	30.0	205180.000000	6839.333333	NaN	NaN

Due to the inclusion of the interaction effect term, we can see a slight change in the p-value of the first two treatments 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 for alloy type 1 in this case.

FOR ALLOY TYPE 2

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	56797.911111	14199.477778	1.106152	0.371833
C(Method)	2.0	499640.400000	249820.200000	19.461218	0.000004
C(Dentist):C(Method)	8.0	197459.822222	24682.477778	1.922787	0.093234
Residual	30.0	385104.666667	12836.822222	NaN	NaN

Due to the inclusion of the interaction effect term, we can see a slight change in the p-value of the first two treatments 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 accepted for alloy type 2 in this case.

THANK YOU