Advance Statistics Project

by Anisha Sharma

PGPDSBA.O.Mar23.A

Great Learning

CONTENT:

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?

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

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

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

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

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?

2.2 What is the probability of a radiation leak?

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.

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)**

3.1 What 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.?

3.3 What 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.?

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?

4.2 What is the probability that a 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?

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?

5.2 Is the mean hardness of the polished and unpolished stones the 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 hypothesis must be formed accordingly.

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 hypothesis separately for the two types of alloys.?
2. Before the hypothesis may be tested, state the required assumptions. Are the assumptions fulfilled? Comment separately on both alloy types.?
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?
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?
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?
6. Consider the interaction effect of dentist and method and comment on the interaction plot, separately for the two types of alloys?

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?

Problem1:

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

Solution : P(Injured) = Total injured/Total

= 145/235

= 0.617

Probability of injured people can be checked by dividing Total injured by Total population. So the probability of randomly chosen player getting injured is 0.617.

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

Solution : P(Forward or Winger) = P(Forward) + P(Winger)

= (94/235) + (29/235)

= 0.523

The Probability that a player is a forward or a winger is 0.523 which is calculated by adding the probability getting total forward and winger players dividing by Total players.

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

Solution: $P(\text{Striker} \cap \text{Foot injured}) = 45/235$

$= 0.191$

The Probability that a randomly chosen player plays in striker position has a foot injury is 0.191.

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

Solution: $P(\text{Striker}|\text{Injured}) = (\text{Striker} \cap \text{Foot injured})/P(\text{injured})$

$= (45/235)+(145/235)$

$= 0.310$

The Probabilty that a randomly chosen injured is a striker is 0.310 by adding the striker injured and total injured.

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

Solution: P(Either forward or attacking midfielder)=Total forward and attacking midfielder injured/Total injured player =80/235

=0.551

Total Forward injured players and injured attacking midfielder is 80(56+25).

Total injured player is 145.

So, The Probabilty of randomly chosen injured player is either forward or an attcking midfielder is 0.551.

Problem2:

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

Solution: According to the studies carried out by the organization,

- The probability of a radiation leak in case of a Fire is 20%, we can formulate this as, $P(R|F) = 0.2$.
- The probability of a radiation leak in case of a mechanical 50%, we can formulate this as, $P(R|M) = 0.5$.
- The probability of a radiation leak in case of a human error is 10%, we can formulate this as, $P(R|H)= 0.1$.
- The probability of a radiation leak occurring simultaneously with a fire is 0.1%, we can formulate this as,$P(R\cap F) = 0.001$.
- The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%, we can formulate this as, $P(R\cap M) = 0.0015$.
- The probability of a radiation leak occurring simultaneously with a human error is 0.12%, we can formulate this as, $P(R\cap H) = 0.0012$. So, the

probabilities of a fire

P(F): $P(R \cap F)/ P(R \mid F) = 0.001/0.2 = 0.005$

probabilities of a mechanical failure

P(M): $P(R \cap M)/ P(R \mid M) = 0.0015/0.5 = 0.003$

2.2 What is the probability of a radiation leak?

Solution: The types of accidents possible at the plant are, fire hazards, mechanical failure, or human error, so using Addition Rule we can calculate

probability of a radiation leak(P(R))

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

$= 0.001 + 0.0015 + 0.0012 = 0.0037$

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: By using Bayes's Theorem:

- The probability of a radiation leak in the reactor due to a Fire

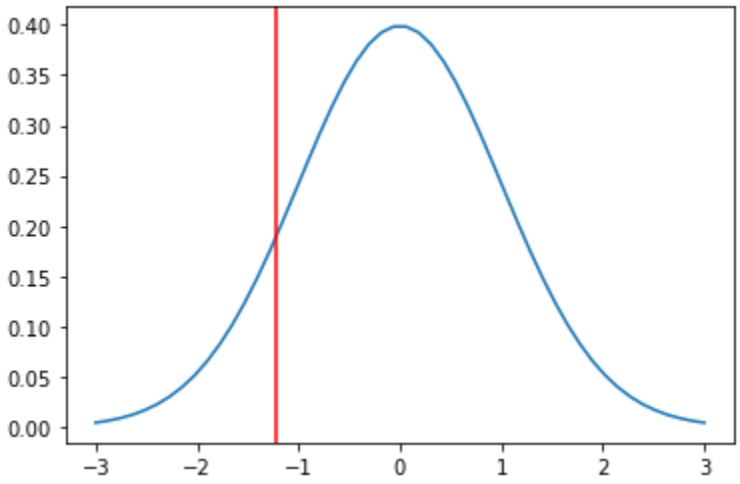
  $P(F \mid R): P(R \cap F)/P(R) = 0.001/0.0037 = 0.2702$

- The probability of a radiation leak in the reactor due to a Mechanical Failure. P(M | R): P(R∩M)/P(R) =0.0015/0.0037 = 0.4054
- The probability of a radiation leak in the reactor due to Human Error P(H | R): P(R∩H)/P(R) =0.0012/0.0037 = 0.3243

Problem3:

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

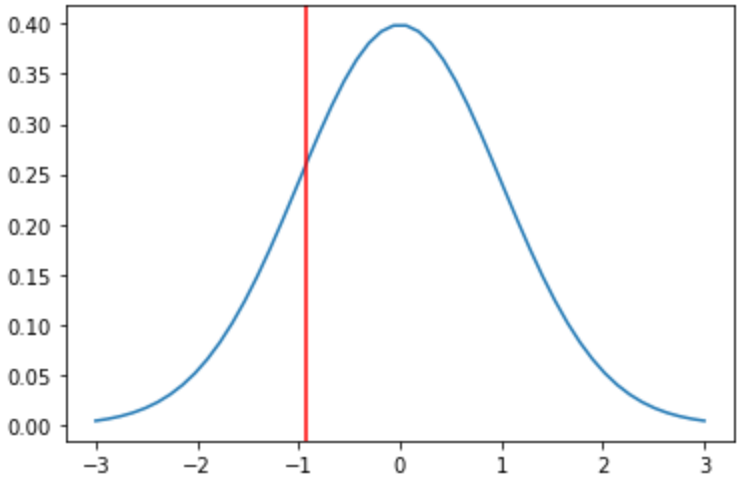
Solution: The probability of the gunny bags have a strength less than 3.17kg kg sq.cm is 0.1112



Conclusion: 11.12% proportion of the 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.?

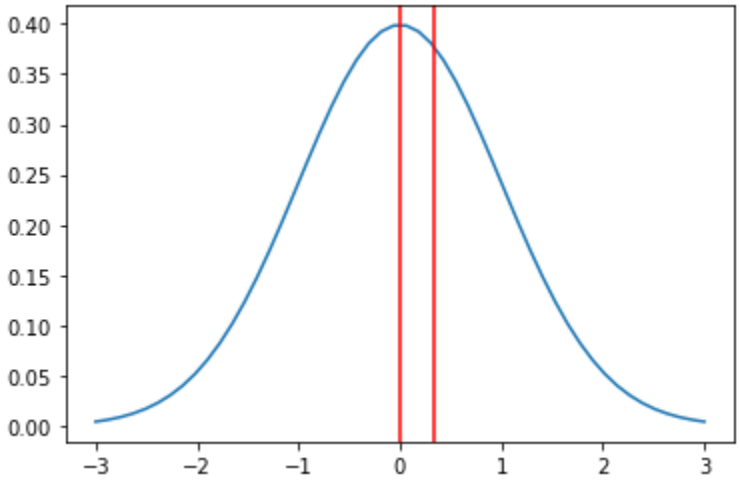
Solution: The probability of the gunny bags have a breaking strength at least 3.6 kg sq.cm is 0.8246.



Conclusion: 82.46% 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.?

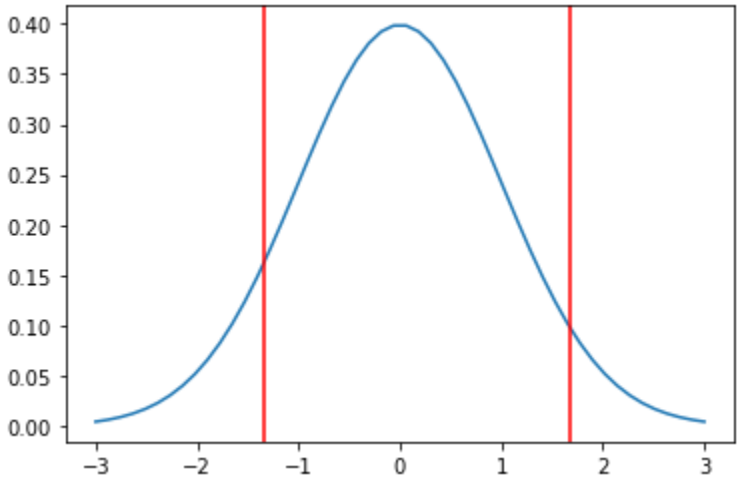
Solution: The proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm is 0.1305.



conclusion: 13.05% of the gunny bags have 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.?

Solution: The proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm is 0.1390.



Conclusion: ~14% of gunny bags have breaking strength not between 3 and 7.5 kg per sq.cm.

Problem4:

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

Solution: For finding the probability we have to use z score which tells how far the data point is from mean and also it's a measure of how many standard deviation below or above the population mean a raw score is.

z = (x-mean)/std

=(85-77)/8.5

= 0.94111

And will use cumulative frequency distribution to find the probability of a random variable having values less than or equal to.

- stats.norm.cdf(z)

The probability that a randomly chosen student gets a grade below 85 on this exam is 0.8266.

Conclusion: 82.66% of the people gets grade below 85 on this exam.

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

Solution: Calculating z score between two random variables:

z1 = (65-77)/8.5

z2 = (87-77)/8.5

Calculate by cdf:

- stats.norm.cdf(z2)-stats.norm.cfd(z1)

The probability that a randomly selected student scores between 65 and 87 is 0.8012

Conclusion: 80.12% is the probability of a 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?

Solution: For that we will ppf(). Ppf() is a function that calculates the normal distribution value for which a given probability is the required value.

norm.ppf() takes a percentage and returns a standard deviation multiplier for what value that percentage occurs at.

- stats.norm.ppf(0.75,loc=77,scale=8.5)

which will give the output value 82.73

Conclusion: 82% marks are required so that 75% of the students clear the exam.

Problem5:

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?

Solution: To check the surface of the stone having Brinell's hardness index of at least 150. We use one sample t test.

Step 1: Define Null and Alternate hypothesis

H0 = Mean brinell's hardness index of unpolished stone is greater than or equal to 150. H0: μ=>150.

Ha = Mean brinell's hardness index of unpolished stone is less than 150. Ha: $\mu<150$.

Step 2: Decide significance level, apha=0.05.

Step 3: Identify the test statistics.

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

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.

One sample t test , t statistic: -4.164629601426758, p value: 8.342573994839285e-05

Step 5: Decide to reject or accept the null hypotheses.

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= 4.1712869974196425e-05

Conclusion: We have analyse that there is sufficient evidence for Zingaro company to believe that unpolished stones are not suitable for printing as brinell's hardness index is less than 150.

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

Solution: By comparing the mean hardness of the polished and unpolished stones, Here we need to test the hardness of unpolished stone with polished stone. As per Zingaros believe unpolished stones may not be suitable for printng.

We observed:

Mean hardness of polished stone is 134.1

Mean hardness of unpolished stone is 147.7

Hence, means are not same as mean of polished and unpolished stones.

Problem6:

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%)

Solution: To check if the null hypothesis need to be rejected or accepted we will use paired two sample t test.

Step 1: Define Null and Alternate hypotheses.

We have define if candidate is able to do more than 5 pushups as compared to when he/she joined the program.

$\mu 1$ = count of pushups after joining the program.

$\mu 2$ = count of pushups before joining the program.

H0 = Difference after joining the program pushup counts has been increased to more than 5. H0: $\mu 1-\mu 2<5$.

Ha = Difference after joining the program pushup counts has not been increased to more than 5. Ha: $\mu 1-\mu 2>5$.

Step 2: Decide significance level.

alpha=0.05 and standard deviation is not known.

Step 3: Identify the test statistics.

- We have two samples and we do not know the population standard deviation.
- Sample sizes for both samples are same.
- The sample is n=100. So you 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. This test assumes that the populations have identical variances.

Step 5: Decide to reject or accept the null hypotheses.
Paired two-sample t-test p-value= 1.1460209626255983e-35
We have enough evidence to reject the null hypothesis in favour of alternative hypothesis
We conclude that the difference after joining the program pushup counts has been increased.

Conclusion: We have enough evidence to reject the null hypothesis in favor of alternative hypothesis, so we can conclude that the claim of Aquarius health club training program is unsuccessful, that the candidates of Aquarius health club fail to do more than 5 push-ups.

Problem7 :

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 hypothesis separately for the two types of alloys.?

Solution: For that we will use the hypothesis for the one way Anova.
H0: The mean response is same for both the alloys

Ha: The mean response is not same for both the alloys.

formula='Response ~ C(Dentist)'

model=ols(formula,dset).fit()

aov_table=anova_lm(model)

aov_table

P_value is greater than alpha which is 0.05, thus we fail to reject the null hypotheses.

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

Solution: Assumptions:-

- Response variable of all population are continuous and normally distributed.
- There should be no significance outliers.
- Number of observation is same for both.
- The sample data of population is independent.
- Dependent variable should be measured at the continuous level.
- Dependent variable should be approximately normally distributed.
- Independent variable should be independent and categorical.
- The variation within each group being compared is similar for every group.

Insights:

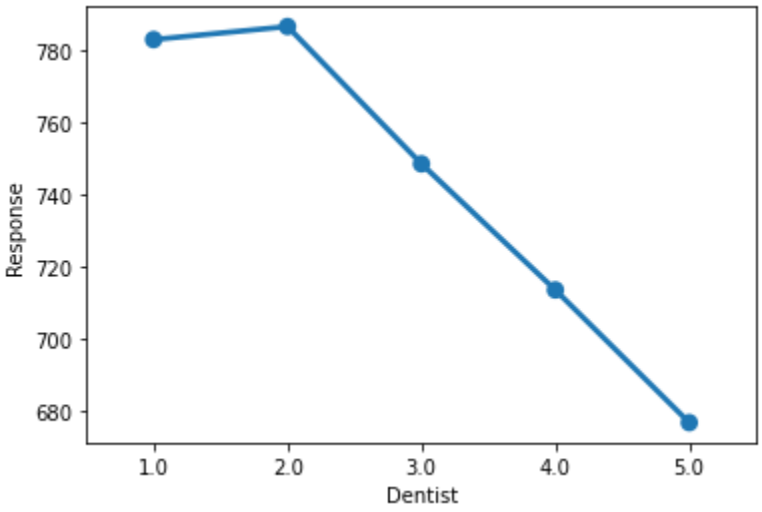
- Response have some outliers. So we have treated it.
- Observation of each group is same.

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?

Solution: For that we will use the hypothesis for the one way Anova.
H0: The mean response is same for both the alloys

Ha: The mean response is not same for both the alloys.

P_value is greater than alpha which is 0.05, thus we fail to reject the null hypotheses.

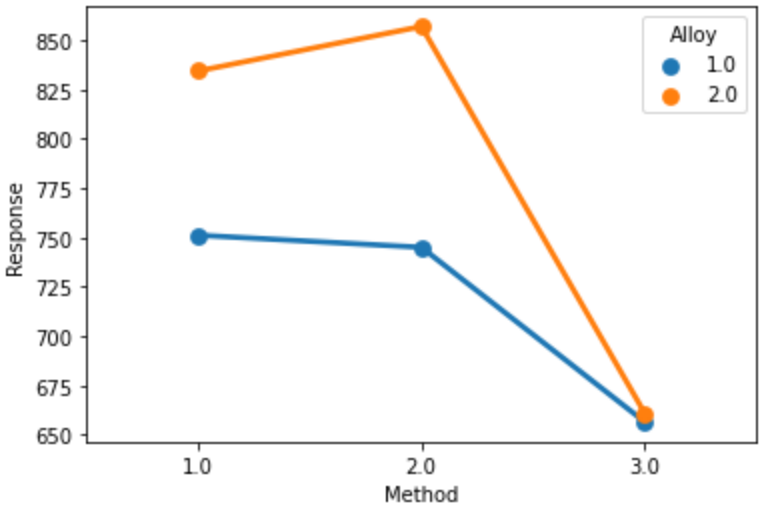


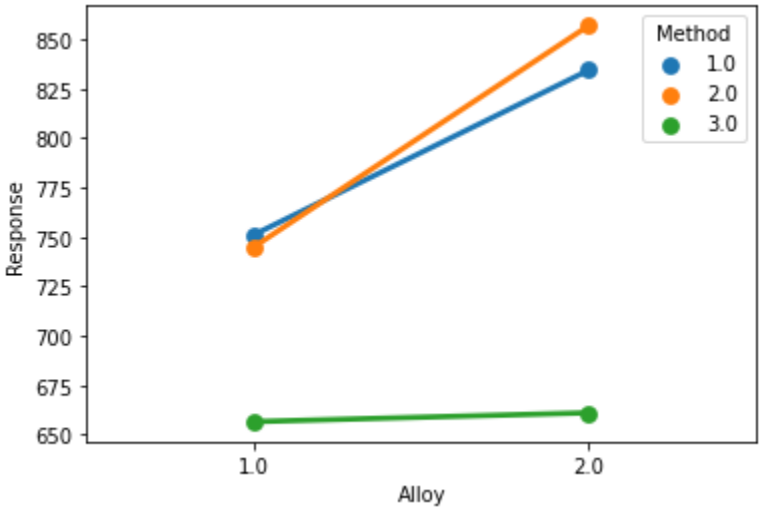
Conclusion: We have enough evidence that not all responses are the same for both alloys , at 5% significance level.

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?

Solution: For that we will use the hypothesis for the one way Anova.
H0: The mean response is same for both the alloys

Ha: The mean response is not same for both the alloys.

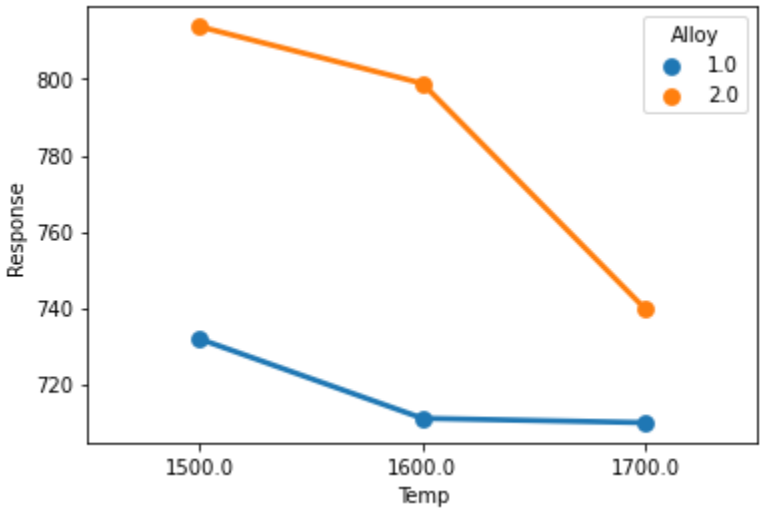




Since p-value is significantly less than alpha = 0.05, there is sufficient evidence against the null hypothesis that all methods have equal impact on hardness.so we can reject null hypothesis in that scenario.

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?

Solution:



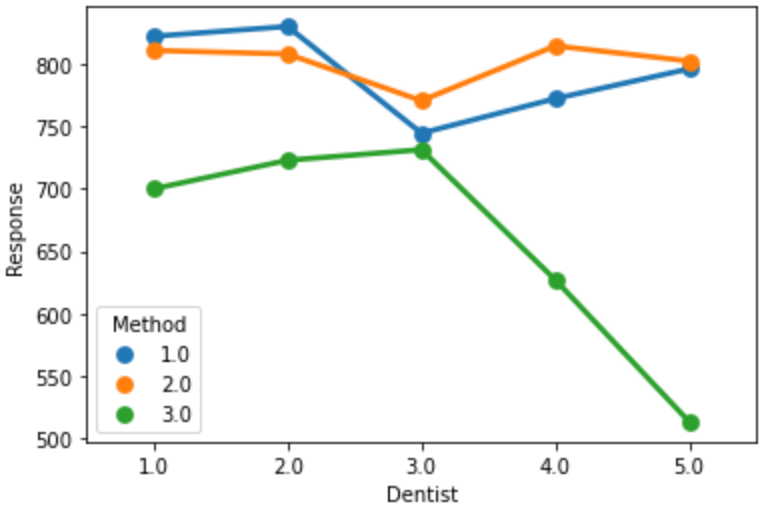
Since p-value> alpha = 0.05, failed to reject the null. So, there is no difference among the temperature levels on the hardness of dental implant, separately for the two types of alloys.

Conclusion: There is huge difference in the temp for both alloys. Alloy 2 has higher response than the alloy 2. and after 1600, there is sudden fall whereas alloy 1 remains the same approximate. We have no evidence to reject the null hypothesis.

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

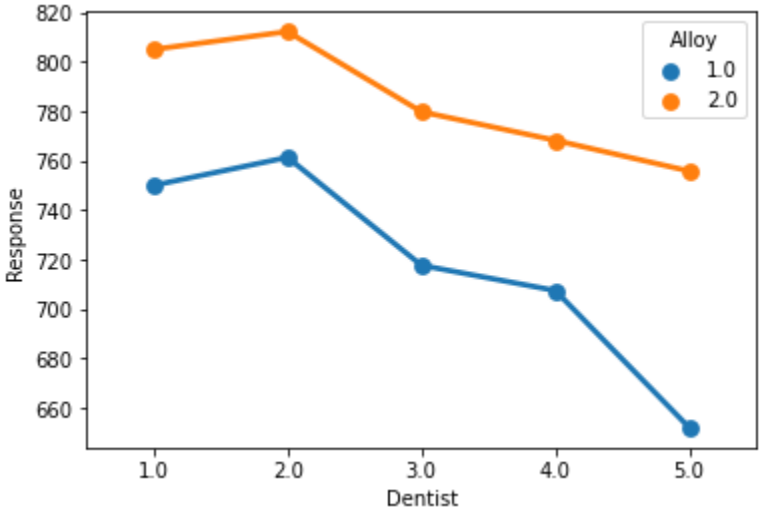
Solution: We have done testing using Anova.

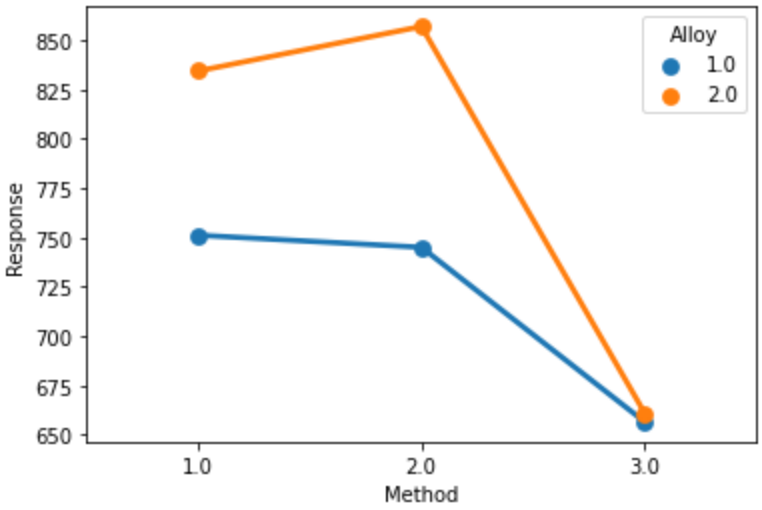
Here we can see for the method 1 and 2, responses of dentist is approximately equal. And there is sudden fall for method 3 for consecutively 3,4,5.



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?

Solution:





Conclusion: Effect of dentist for alloy 2 is higher than the alloy 1. but both of them gradually decrease after 2.

Effect of method, there is a sudden fall in both the alloys after 2. but alloy 2 has higher response for method than the alloy 1.