IS GRADED PROJECT - CODED

DSBA

Krishnabhamini Sinha

Contents

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**

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

Solution:

Total number of players in the given table = 235.
Total number of players who are injured = 145.

The probability that a randomly chosen player would suffer an injury

= injured players / total number of players

= 145 / 235

= 0.61

The probability that a randomly chosen player would suffer an injury is 61%.

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

Solution:

Total number of players in the given table = 235.
Total number of forward players = 94.
Total number of winger players = 29.

The probability that a player is a forward or a winger

= (total forward players + total winger players) / total number of players

= 0.52

The probability that a player is a forward or a winger is 52%.

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

Solution:

Total number of players in the given table = 235.
Total number of striker players who are injured = 45.

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

= total number of injured striker players / total number of players

= 45 / 235

= 0.19

The probability that a randomly chosen player plays in a striker position and has a foot injury is 19%.

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

Solution:

Total number of injured players in the given table = 145.
Total number of striker players = 45

The probability that a randomly chosen injured player is a striker is

= total number of striker players / total number of injured players

= 0.31

The probability that a randomly chosen injured player is a striker is 31%.

Problem 2

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)**

Understanding the Problem Statement:

According to the given problem statement,

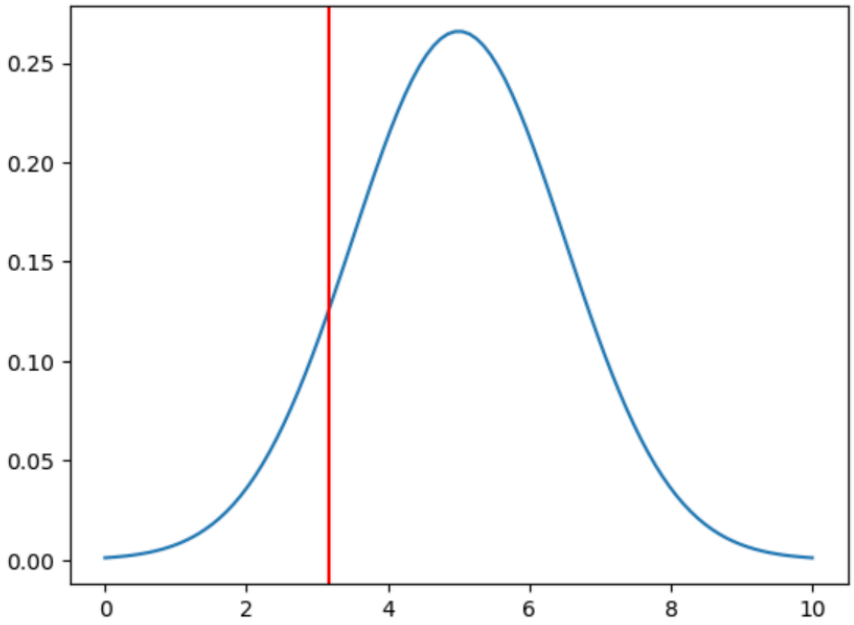
Given: mean (mu) = 5 and standard deviation (sigma) = 1.5
We will use the Z-test for this problem.

The formula for calculating z-statistic is = (x_bar - mu) / sigma

x_bar is the given value for calculating the z-statistic, whereas mean and sigma are mean and standard deviation of the given population that follows normal distribution.

Problem 2.1 - What proportion of the gunny bags have a breaking strength of less than 3.17 kg per sq cm?

Solution:



1. Z-statistic of the given value is = (x_bar - mu) / sigma

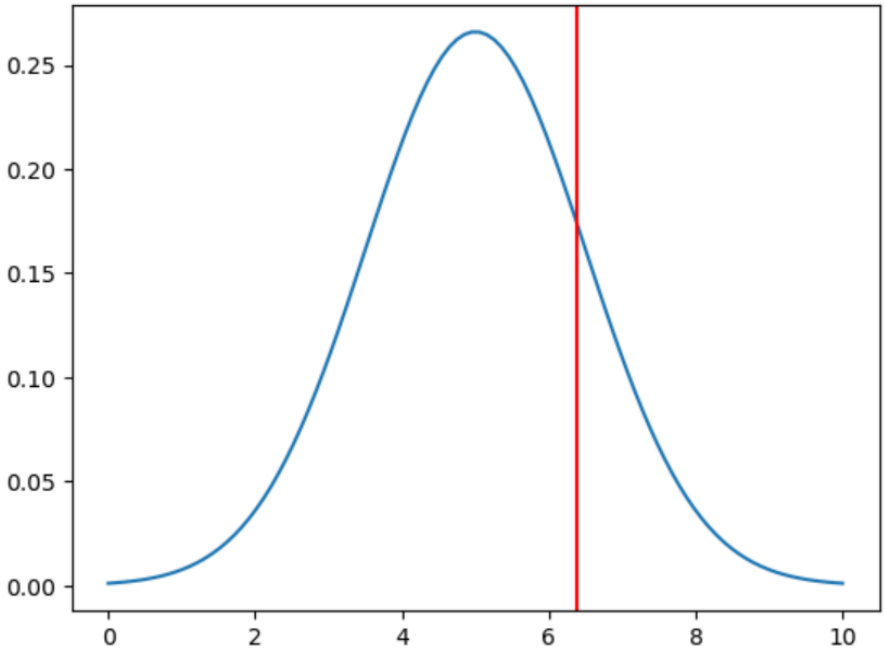
   = (3.17-5)/1.5

   = -1.22

2. The cumulative area corresponding to the z-statistic calculated above (-1.22) is 0.111 (calculated by stats.norm.pdf). So, the cumulative probability that gunny bags have a breaking strength of less than 3.17 kg per sq cm - as represented by the red vertical line in the figure above is **11.1%**.

Problem 2.2 - What proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm.?

Solution:



1. The z-statistic for the given value 3.6 is = (x_bar - mu) / sigma

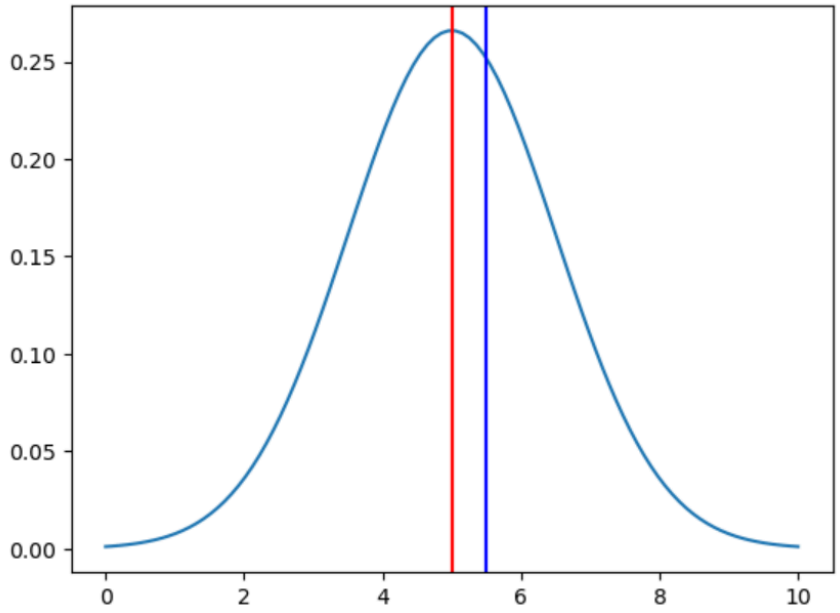
   = (3.6 - 5)/1.5

   = - 0.933

2. The cumulative area corresponding to the z-statistic calculated above (-0.933) is 0.82 - calculated by (1 - stats.norm.pdf). So, the cumulative probability that gunny bags have a breaking strength of at least 3.6 kg per sq cm - as represented by the red vertical line in the figure above is **82%**.

Problem 2.3 - What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?

Solution:



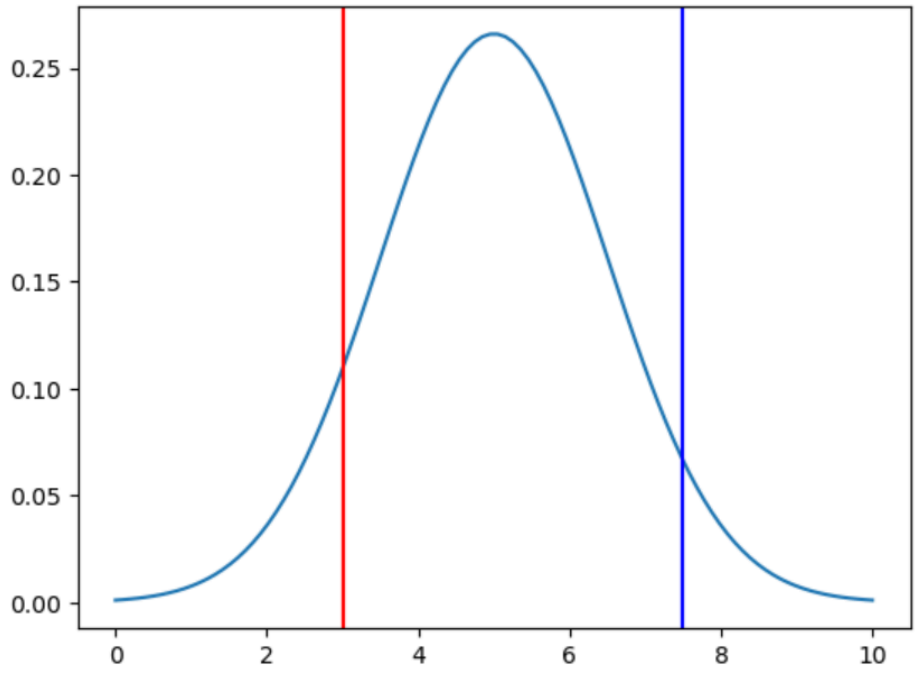
1. Z-statistic, z2, of upper limit value (5.5) is = (5.5-5)/1.5
   = 0.333

2. Z-statistic, z1, of lower limit value (5.0) is = (5-5)/1.5
   = 0

3. Cumulative area distribution between the two calculated z-statistics is 0.13 - calculated by (stats.norm.cdf(z2) - stats.norm.cdf(z1)). So, the cumulative probability that gunny bags have a breaking strength between 5 and 5.5 kg per sq cm. Is **13%** as represented by the area between the vertical red and blue lines respectively.

Problem 2.4 - What proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm.?

Solution:



1. Z-statistic, z2, of upper limit value (7.5) is = (7.5-5)/1.5
   = 1.66

4. Z-statistic, z1, of lower limit value (3.0) is = (3-5)/1.5
   = -1.33

5. Cumulative area distribution between the two calculated z-statistics is 0.139 - calculated by (1 - (stats.norm.cdf(z2) - stats.norm.cdf(z1)). So, the cumulative probability that gunny bags have a breaking strength not between 3 and 7.5 kg per sq cm. Is **13.9%.**

6. The area to the left of the vertical red line represents gunny bags that have a breaking strength less than 3 kg per sq cm. The area to the right of the vertical blue line represents gunny bags that have a breaking strength more than 7.5 kg per sq cm.

Problem 3

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).

Understanding the problem statement:

The assumed mean of the given data is - 150, considering the fact that the optimum level of printing images on stones has to have a Brinell's hardness index of at least 150.

The level of significance is - 0.05

Standard deviation of the given population is not known. The size of population is 75.

Problem 3.1 - Zingaro has reason to believe that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

"- State the null and alternate hypotheses - Conduct the hypothesis test and compute the p-value - Write down conclusions from the test results Note: Consider the level of significance as 5%."

Solution:

H0 (Null Hypothesis): Mean hardness index of unpolished stones is at least equal to 150.
H1 (Alternate Hypothesis): Mean hardness index of unpolished stones lis ess than 150.

- We have an assumed mean for the given data but we do not know the standard deviation of the same.
- We need to conduct a test to find hardness values that are less than 150. Therefore, we will conduct a one-tailed T-test. The test will be a left-tailed T-test.

  The T-statistics:

  P-value = 4.171286997419652e-05

```
t_statistic, p_value = ttest_1samp(df['Unpolished'], 150, alternative='less')
print('One sample t-test p-value:', p_value)
```

```
One sample t-test p-value: 4.171286997419652e-05
```

- Since the p-value < 0.05 (level of significance), we reject the null hypothesis.
- The alternate hypothesis is true i.e., mean hardness of unpolished stones < 150. So, the decision that the unpolished stones are not suitable for printing is justified.

Problem 3.2 - Is the mean hardness of the polished and unpolished stones the same?

- State the null and alternate hypotheses. - Conduct the hypothesis test. - Write down conclusions from the test results. Note: Consider the level of significance as 5%.

Solution:

H0 (Null Hypothesis): Mean of polished stones is equal to mean of unpolished stones.
H1 (Alternate Hypothesis): Mean of polished is not equal to mean of unpolished stones.

- We have an assumed mean for the given data but we do not know the standard deviation of the same.
- We will be comparing two independent samples whose standard deviations are unknown. For this we will apply a two-sample T-test.

The T-statistics:
P-value: - 0.0014655150194628353

```
t_statistic, p_value = ttest_ind(df['Unpolished'], df['Treated and Polished'])
print('t statistic', t_statistic)
print('p-value', p_value)
```

```
t statistic -3.2422320501414053
p-value 0.0014655150194628353
```

- Since the p-value < 0.05 (level of significance), we reject the null hypothesis.

- The alternate hypothesis is true i.e., the mean of polished stones is not equal to the mean of unpolished stones.

Problem 4

Dental implant data: The hardness of metal implants 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 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.

Problem 4.1 - How does the hardness of implants vary depending on dentists?

"- State the null and alternate hypotheses - Check the assumptions of the hypothesis test. - Conduct the hypothesis test and compute the p-value - Write down conclusions from the test results - In case the implant hardness differs, identify for which pairs it differs Note: 1. Both types of alloys cannot be considered together. You must conduct the analysis separately for the two types of alloys. 2. Even if the assumptions of the test fail, kindly proceed with the test."

Solution:

Since we have to check for hardness of implants we will run tests separately for both the alloys.

Hypotheses for Alloy 1:

H0 (Null Hypothesis): There is no difference in means among dentists in terms of hardness of implants for Alloy 1.

H1 (Alternate Hypothesis): There is at least one difference in means among dentists in terms of hardness of implants for Alloy 1.

Hypotheses for Alloy 2

H0 (Null Hypothesis): There is no difference in means among dentists in terms of hardness of implants for Alloy 2.

H1 (Alternate Hypothesis): There is at least one difference in means among dentists in terms of hardness of implants for Alloy 2.

Checking Assumptions:

Applying Shapiro-Wilk's test to data for Alloy 1:

H0 (Null Hypothesis): The population from which sample is drawn follows normal distribution.
H1 (Alternate Hypothesis): The population from which sample is drawn does not follow normal distribution.

P-value obtained after applying test = 0.00043243501568213105

p-value < 0.05 (level of significance)

Conclusion: We reject the null hypothesis and accept the alternate hypothesis i.e, the variable 'Dentist' for Alloy 1 does not follow normal distribution.

Applying Shapiro-Wilk's test to data for Alloy 2:

H0 (Null Hypothesis): The population from which sample is drawn follows normal distribution.
H1 (Alternate Hypothesis): The population from which sample is drawn does not follow normal distribution.

P-value obtained after applying test = 0.00043243501568213105

p-value < 0.05 (level of significance)

Conclusion: we reject the null hypothesis and accept the alternate hypothesis i.e, the variable 'Dentist' for Alloy 2 does not follow normal distribution.

Applying Levene's test to data for Alloy 1 and Alloy 2:

H0 (Null Hypothesis): The variances of the groups being compared for Alloy 1 and Alloy 2 are equal.
H1 (Alternate Hypothesis): The variances of the groups being compared for Alloy 1 and Alloy 2 are not equal.

P-value = 0.007858817382355401

Conclusion: p-value < 0.05. Therefore, we reject the Null Hypothesis i.e., variances of the groups being compared for Alloy 1 and Alloy 2 are not equal.

Results of hypotheses tests for Alloy 1 and Alloy 2:

- We perform one-way ANOVA for both Alloy 1 & 2 since we are testing multiple groups ('Dentist') under one variable of interest ('Response').

- P-value for Alloy 1 = 0.11656712140267628

- P-value > 0.05

Conclusion: We fail to reject the Null Hypothesis i.e., there is no difference in means among dentists in terms of hardness of implants for Alloy 1.

- P-value for Alloy 2 = 0.7180309510793431

- P-value > 0.05

Conclusion: We fail to reject the Null Hypothesis i.e., there is no difference in means among dentists in terms of hardness of implants for Alloy 2.

Problem 4.2 - How does the hardness of implants vary depending on methods?

"- State the null and alternate hypotheses - Check the assumptions of the hypothesis test. - Conduct the hypothesis test and compute the p-value - Write down conclusions from the test results - In case the implant hardness differs, identify for which pairs it differs Note: 1. Both types of alloys cannot be considered together. You must conduct the analysis separately for the two types of alloys. 2. Even if the assumptions of the test fail, kindly proceed with the test."

Solution:

Hypotheses for Alloy 1:

H0 (Null Hypothesis): There is no difference in means among methods in terms of hardness of implants for Alloy 1.

H1 (Alternate Hypothesis): There is at least one difference in means among methods in terms of hardness of implants for Alloy 1.

Hypotheses for Alloy 2

H0 (Null Hypothesis): There is no difference in means among methods in terms of hardness of implants for Alloy 2.

H1 (Alternate Hypothesis): There is at least one difference in means among methods in terms of hardness of implants for Alloy 2.

Checking Assumptions:

Applying Shapiro-Wilk's test to data for Alloy 1:

H0 (Null Hypothesis): The population from which sample is drawn follows normal distribution.
H1 (Alternate Hypothesis): The population from which sample is drawn does not follow normal distribution.

P-value obtained after applying test = 1.7407411405656603e-06

p-value < 0.05 (level of significance)

Conclusion: We reject the null hypothesis and accept the alternate hypothesis i.e, the variable 'Method' for Alloy 1 does not follow normal distribution.

Applying Shapiro-Wilk's test to Alloy 2:

H0 (Null Hypothesis): The population from which sample is drawn follows normal distribution.
H1 (Alternate Hypothesis): The population from which sample is drawn does not follow normal distribution.

P-value obtained after applying test = 1.7407411405656603e-06

p-value < 0.05 (level of significance)

Conclusion: we reject the null hypothesis and accept the alternate hypothesis i.e, the variable 'Method' for Alloy 2 does not follow normal distribution.

Applying Levene's test to Alloy 1 and Alloy 2:

H0 (Null Hypothesis): The variances of the groups being compared for Alloy 1 and Alloy 2 are equal.
H1 (Alternate Hypothesis): The variances of the groups being compared for Alloy 1 and Alloy 2 are not equal.

P-value = 0.004138452940152019

Conclusion: p-value < 0.05. Therefore, we reject the Null Hypothesis i.e., variances of the groups being compared for Alloy 1 and Alloy 2 are not equal.

Results of hypotheses tests for Alloy 1 and Alloy 2:

- We perform one-way ANOVA for both Alloy 1 & 2 since we are testing multiple groups ('Method') under one variable of interest ('Response').

- P-value for Alloy 1 = 0.004163412167505543

- P-value < 0.05

Conclusion: We reject the Null Hypothesis i.e., there is at least one difference in means among Methods in terms of hardness of implants for Alloy 1.

- P-value for Alloy 2 = 5.415871051443187e-06

- P-value < 0.05

Conclusion: We reject the Null Hypothesis i.e., there is at least one difference in means among dentists in terms of hardness of implants for Alloy 2.

To find out the difference in means among the groups for both Alloy1 and Alloy 2 we use **Tukey HSD** tests:

<u>For Alloy 1:</u>

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-6.1333	0.987	-102.714	90.4473	False
1	3	-124.8	0.0085	-221.3807	-28.2193	True
2	3	-118.6667	0.0128	-215.2473	-22.086	True

Mean 3 i.e., Method 3 has variation as compared to Method 1 and Method 2. Method 3 has a significant impact on the hardness level for Alloy 1.

For Alloy 2:

```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj    lower     upper   reject
-----------------------------------------------------
     1      2     27.0 0.8212  -82.4546  136.4546  False
     1      3   -208.8 0.0001 -318.2546  -99.3454   True
     2      3   -235.8    0.0 -345.2546 -126.3454   True
```

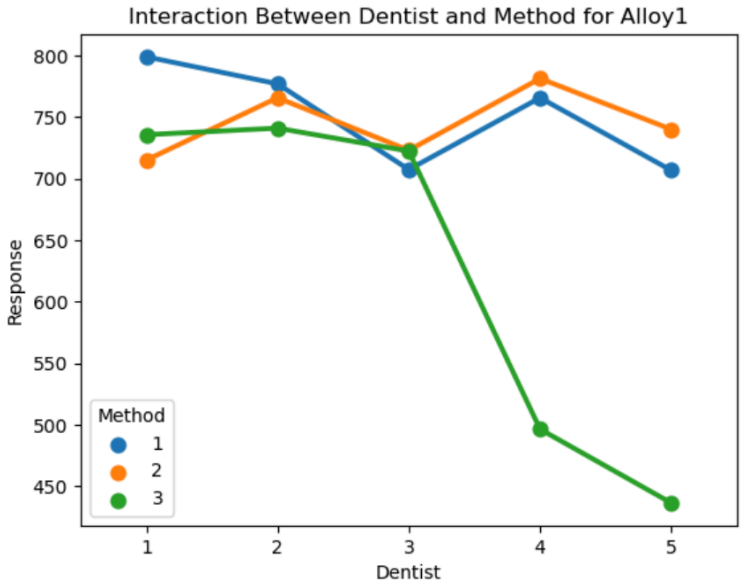
Mean 3 i.e., Method 3 has variation as compared to Method 1 and Method 2. Method 3 has a significant impact on the hardness level for Alloy 2.

Problem 4.3 - What is the interaction effect between the dentist and method on the hardness of dental implants for each type of alloy?

"- Create Interaction Plot - Inferences from the plot Note: Both types of alloys cannot be considered together. You must conduct the analysis separately for the two types of alloys."

Solution:

For Alloy 1:



Inferences:

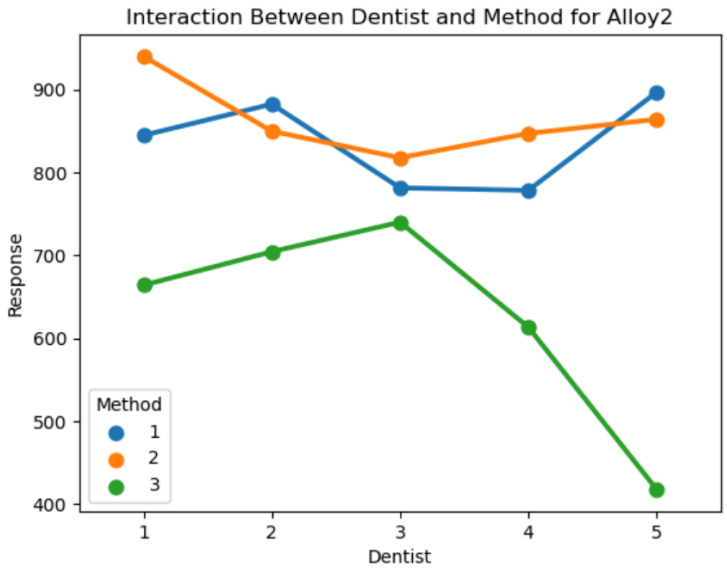
- The interaction plot shows interaction of groups under ‘Dentist’ and ‘Method’ for Alloy 1.

- Interaction is present between ‘Method’ 1, 2 and 3 and groups 1, 2 and 3 under ‘Dentist’.
- Interaction between ‘Method’ and ‘Dentist’ is absent for groups 4 and 5 under ‘Dentist’.
- The places where interactions have been mentioned will have a significant impact upon the hardness levels of implants.

	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

- We can see from the above table that interaction between Dentist and Method for Alloy 1 impacts hardness levels at all three levels.

For Alloy 2:



Inferences:

- The interaction plot shows interaction of groups under ‘Dentist’ and ‘Method’ for Alloy 2.
- Interaction is present between ‘Method’ 1, and 2 and groups 1, 2, 3, 4 and 5 under ‘Dentist’.
- Interaction between ‘Method’ and ‘Dentist’ is absent for groups belonging to Method 3.
- The places where interactions have been mentioned will have a significant impact upon the hardness levels of implants.

	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

- We can see from the above table that interaction between Dentist and Method for Alloy 2 impacts hardness levels at two levels - for Method and intersection of Dentist and Method.

Problem 4.4 - How does the hardness of implants vary depending on dentists and methods together?

"- State the null and alternate hypotheses - Check the assumptions of the hypothesis test. - Conduct the hypothesis test and compute the p-value - Write down conclusions from the test results - Identify which dentists and methods combinations are different, and which interaction levels are different. Note: 1. Both types of alloys cannot be considered together. You must conduct the analysis separately for the two types of alloys. 2. Even if the assumptions of the test fail, kindly proceed with the test."

Solution:

Hypotheses:

For Alloy 1:

H0 (Null Hypothesis): There is no difference among the means of levels belonging to factors ‘Dentist’ and ‘Method’ in terms of hardness levels of implants.

H1 (Alternate Hypothesis): There is at least one difference among the means of levels belonging to factors ‘Dentist’ and ‘Method’ in terms of hardness levels of implants.

For Alloy 2:

H0 (Null Hypothesis): There is no difference among the means of levels belonging to factors ‘Dentist’ and ‘Method’ in terms of hardness levels of implants.

H1 (Alternate Hypothesis): There is at least one difference among the means of levels belonging to factors ‘Dentist’ and ‘Method’ in terms of hardness levels of implants.

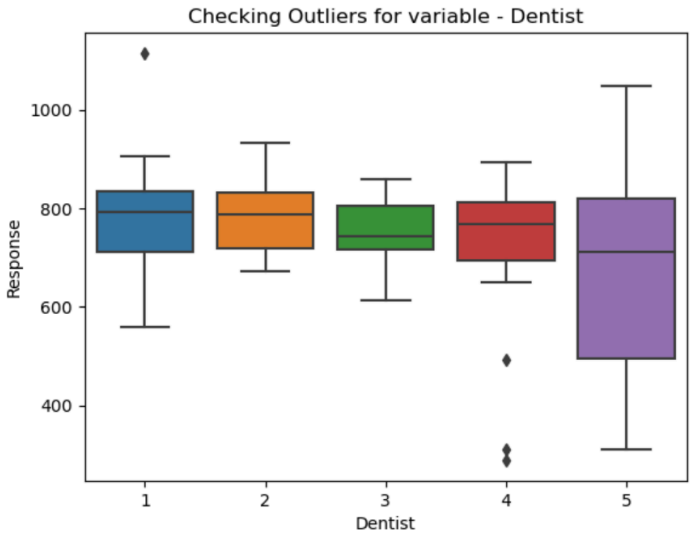
Checking Assumptions of Hypothesis tests:

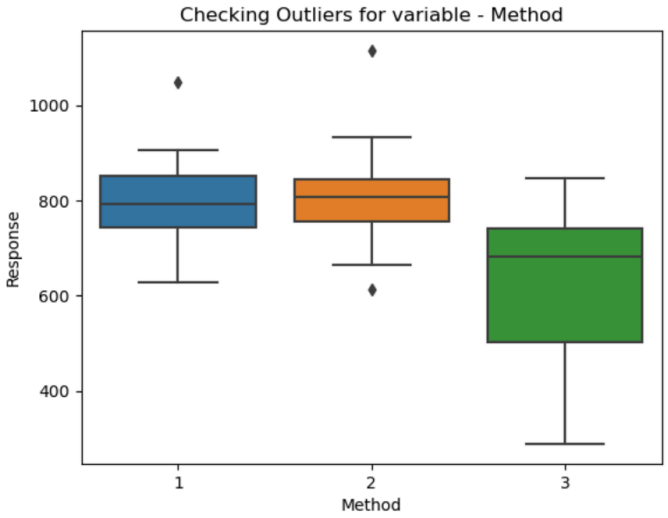
1. Independent variables - 'Dentist' and 'Method' have two or more independent, categorical groups.

   ‘Dentist’ has 5 groups - 1, 2, 3, 4 and 5.
   ‘Method’ has 3 groups - 1, 2 and 3.

2. Dependent variable ‘Response’ should be measured at continuous level - we can see that our dependent variable follows a continuous distribution i.e., it can assume any value depending on the conditions and we can apply arithmetic operations on those values.

3. There should be no significant outliers - there are outliers in group 1 and 4 belonging to variable ‘Dentist’. Groups 1 and 2 under variable ‘Method’ have outliers as well.





5. Dependent variable should be normally distributed for each combination of groups of two variables. The dependent variable does not follow a normal distribution.

Shapiro-Wilk's test for Assumption 5:

H0 (Null Hypothesis): Data for 'Response' variable follows a normal distribution.

H1 (Alternate Hypothesis): Data for 'Response' variable does not follow normal distribution.

```
w, p_value = stats.shapiro(data['Response'])
print('p-value is:', p_value)
```

```
p-value is: 8.080212865024805e-06
```

P-value: 8.080212865024805e-06

p-value < 0.05

We reject the Null hypothesis and accept the alternate hypothesis, i.e., data for 'Response' variable does not follow a normal distribution.

Hypothesis Test Results for Alloy 1:

Since we are comparing different groups of two different factors, we will use the Two-way ANOVA test.

- P- value for Two-way ANOVA test on 'Dentist' and 'Method' together for Alloy 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

- P- value for 'Method' = 0.002211 < 0.05.
- We reject the Null Hypothesis and accept the alternate hypothesis, i.e., there is at least one difference among the means of levels belonging to factor 'Method' in terms of hardness levels of implants.
- We conduct the Tukey HSD test on 'Method' to find the differing group/level.

```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2  meandiff p-adj    lower    upper   reject
-----------------------------------------------------
     1      2   -6.1333  0.987  -102.714  90.4473  False
     1      3    -124.8 0.0085 -221.3807 -28.2193   True
     2      3 -118.6667 0.0128 -215.2473  -22.086   True
-----------------------------------------------------
```

- From the results we can conclude that Method 3 is causing impact on the hardness levels on implants.

Hypothesis Test Results for Alloy 2:

- P- value for Two-way ANOVA test on 'Dentist' and 'Method' together for Alloy 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

- P-value of 'Method' = 0.000008
- P-value < 0.05
- We reject the Null hypothesis, i.e., there is at least one difference among the means of levels belonging to factor 'Method' in terms of hardness levels of implants.

- We conduct the Tukey HSD test on ‘Method’ to find the differing group/level.

```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj    lower     upper   reject
-----------------------------------------------------
     1      2     27.0 0.8212  -82.4546  136.4546  False
     1      3   -208.8 0.0001 -318.2546  -99.3454   True
     2      3   -235.8    0.0 -345.2546 -126.3454   True
-----------------------------------------------------
```

- From the results we can conclude that Method 3 is causing impact on the hardness levels on implants.