SMDB PROJECT
Advance statistics

Problem 1

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

	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**

Based on the above data, answer the following questions.

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

-The total number of players are 235.

-The total number of players Injured are 145

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

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

-The total number of players are 235.

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

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

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

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

-The total number of players which are strikers are 77

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

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

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

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

-The Total number of injured players are 145

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

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

Problem 2

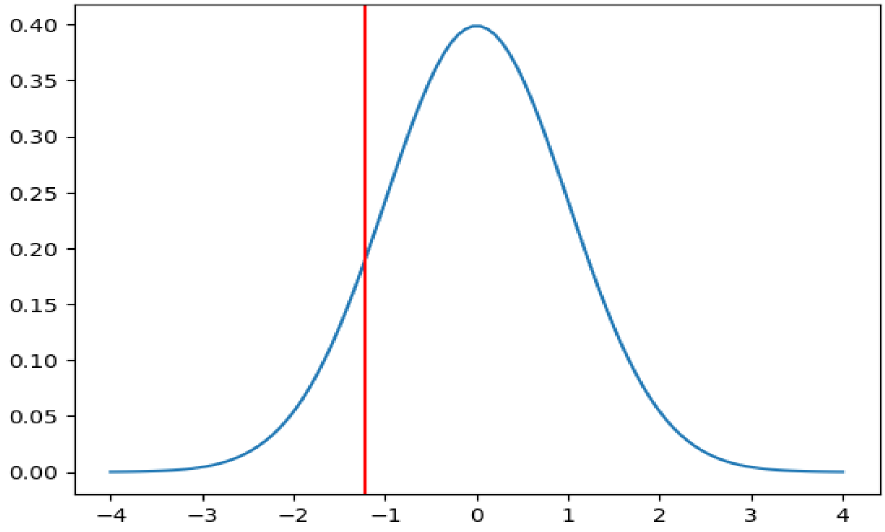
The breaking strength of gunny bags used for packaging cement is normally distributed with a mean of 5 kg per sq. centimeter and a standard deviation of 1.5 kg per sq. centimeter. The quality team of the cement company wants to know the following about the packaging material to better understand wastage 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)**

Given

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

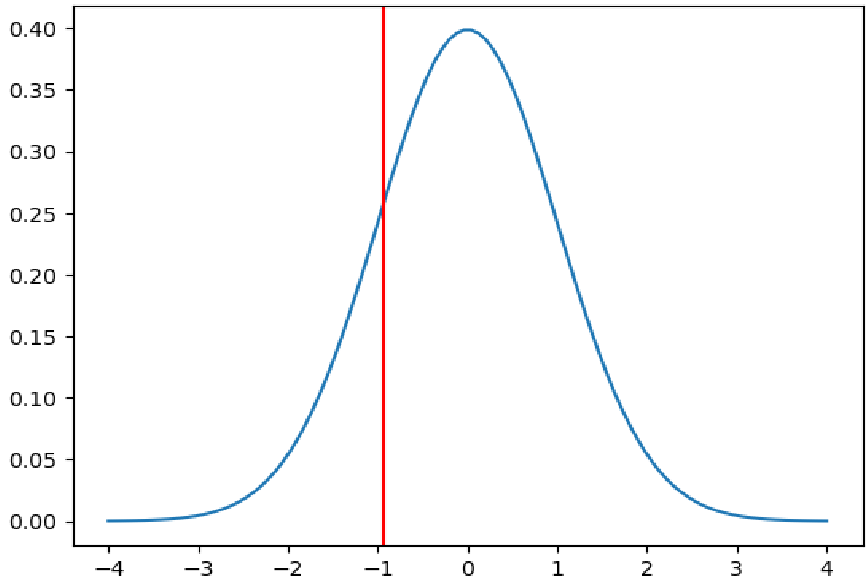
2.1 What proportion of the gunny bags have a breaking strength of less than 3.17 kg per sq cm?

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



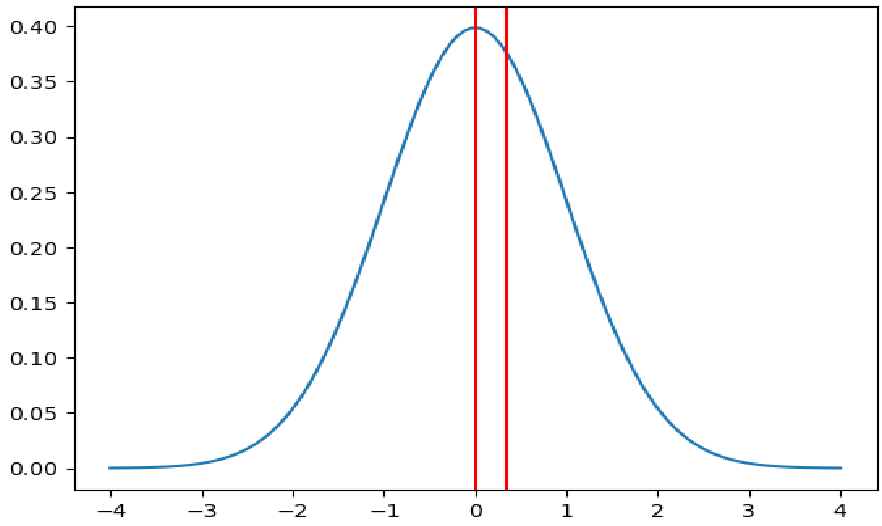
2.2 What proportion of the gunny bags have a breaking strength of at least 3.6 kg per sq cm.?

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



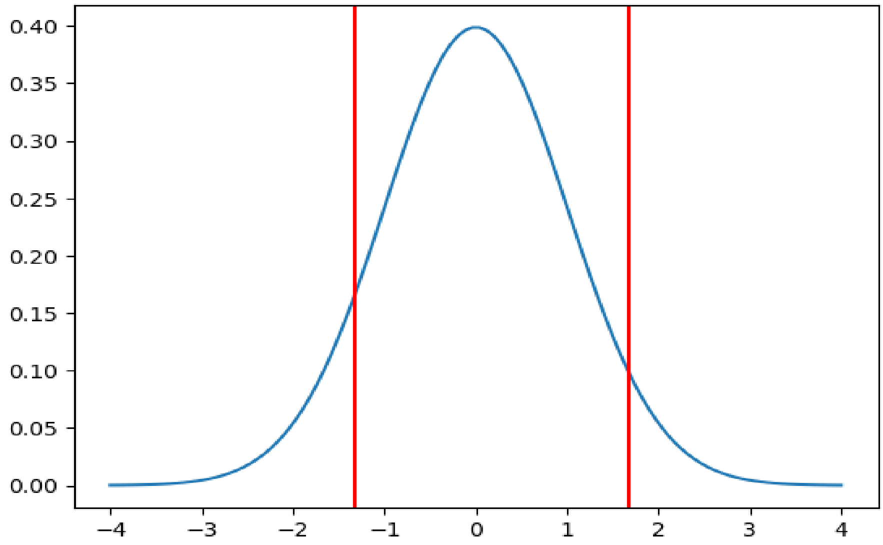
2.3 What proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.?

13.06% of gunny bags have a breaking strength between 5 and 5.5 kg per sq cm.



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

14% of proportion of the gunny bags have a breaking strength NOT between 3 and 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);

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

Dataset has two variables Unpolished and Treated and polished

Pre-analysis

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

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

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?

Step 1: Define null and alternative hypotheses

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

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

OR

Ho: u>=150

Ha: u<150

STEP 2-Decide the significance Level

Here we select alpha, α=0.05

Step 3: Identify the test statistic

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

Step 4: Calculate the p - value and test statistic

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

Values of t statistic= -4.1646296

Value of p = 4.171286995e-05

Step 5: Decide to reject or accept null hypothesis

At Level of significance: 0.05

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

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

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

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

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

Problem 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.

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

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

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

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

Hypothesis for the Anova

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

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

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

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

Testing of the null Hypothesis

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

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

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

The samples drawn from different populations are independent and random.

There should be no significant outliers.

Dependent variable should be measured at the continues level.

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

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

Number of observations in each group are same.

There is homogeneity of variance

IN OUR DATASET THE FOLLOWING ASSUMPTIONS ARE FULLFILLED.

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

Hypothesis for the Anova

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

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

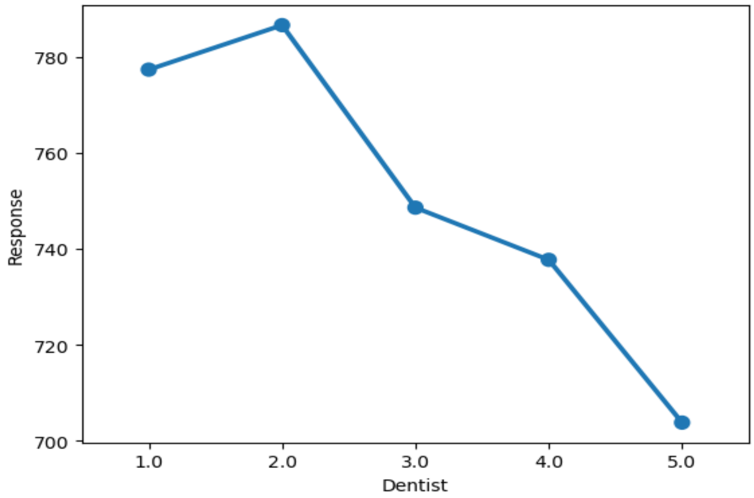
Testing of the null Hypothesis

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

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

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

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



Hypothesis for the Anova

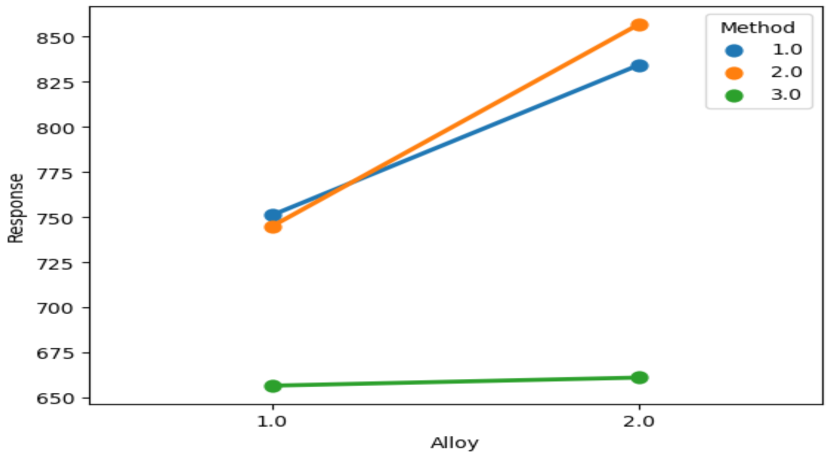
H_0: The mean response is the same for all three types of methods on the hardness of dental implant.

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

Testing of the null Hypothesis

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

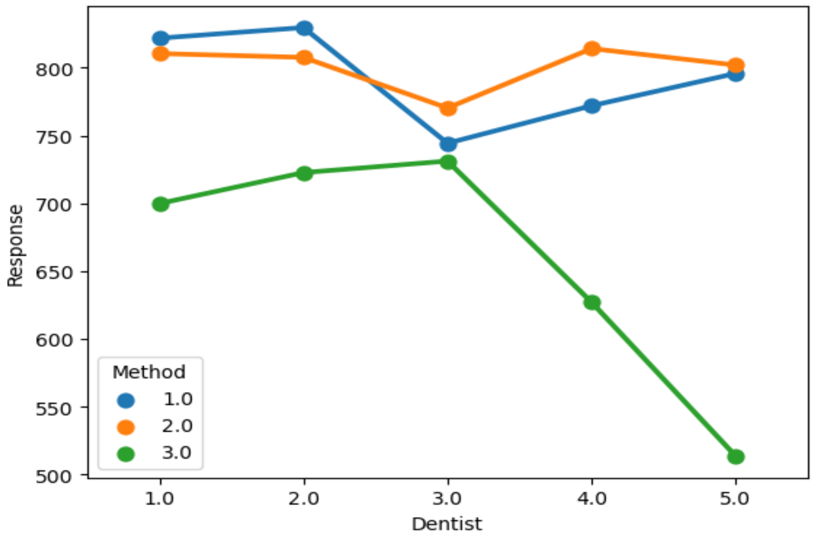
After drawing the pointplot we can clearly see there is an intereaction effect, but the sample data is not enough to conclude that. Also Anova does not help us identify which pairs of dentist differ.



Null hypothesis-

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

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



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