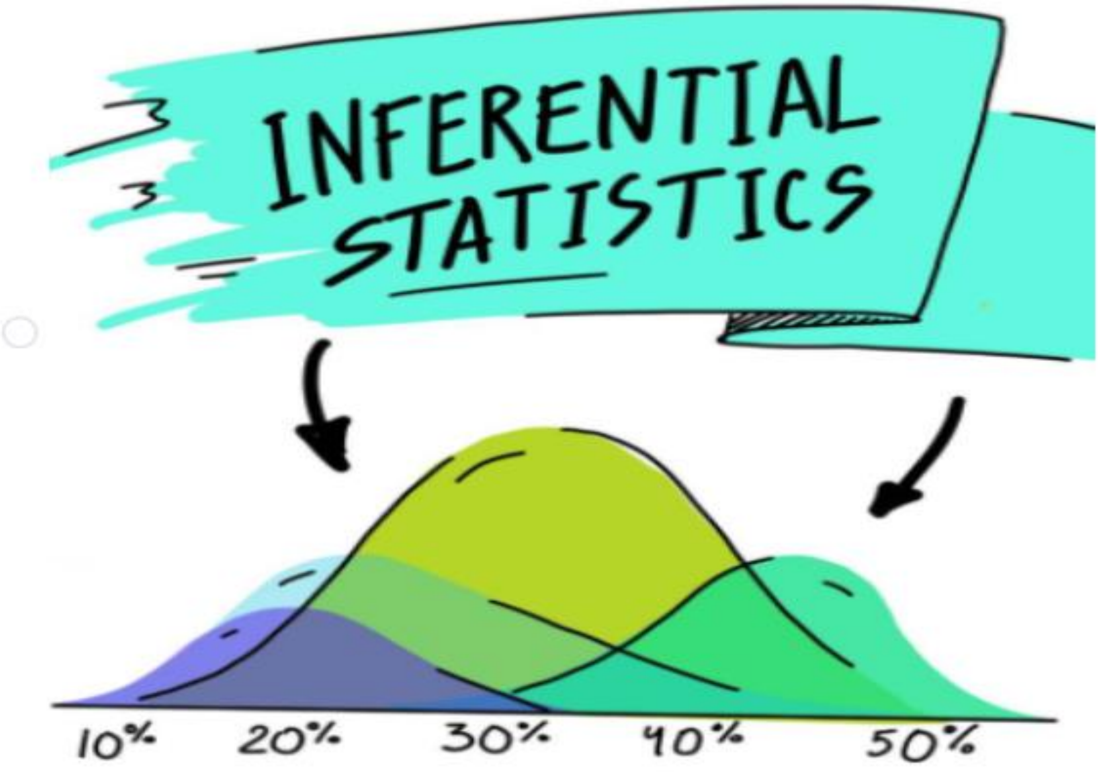


INFERENTIAL STATISTICS PROJECT

By Kurt Warren Mario Gilby On April 7th 2024

Submitted to

Great Learning
POWER AHEAD

As a part of the requirements for completion of PGP-DSBA offered in affiliation with

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Table of Contents

List of Figures

List of Tables

List of Equations

Introduction

As a part of the "Inferential Statistic" learning course we have seen and practiced various tools and techniques that can be used to infer information about a population given a sample/s from that population.

I will in the following pages use the said tools and techniques to review the problem sets given and answer the questions posed in each problem set. In the following pages of this document, I have given an overview, dataset description, objective and information given section, which give added information where needed, the questions asked section under each, lists out the specific questions ask for each dataset/sample.

The basic premise we use to solve each problem is that given a sample/s, we should be able to identify the distribution the sample data follows e.g. Normal, Binominal, Uniform etc. having Identified the distribution we can get the probability density values, using probability density functions (PDF) which help us derive the population parameters using the sample/s statistic values.

One other factor that plays a huge impact in us to be able to look at a sample statistic and infer a population parameter is the "Central Limit Theorem", the premise of which is, no matter what the distribution of the population is, if we take random independent samples from that population and plot the means of samples taking the distribution of the means of the population would be a normal distribution,

Some of the statistics that help us with inferencing the population parameters are for e.g. Z score, t score, F score etc. The equations to calculate some of these distributions using one sample are added to the Appendix A. Using the "Inferential Statistic" tool in the SciPy package of python, we are able to look at statistical problems posed in this project and find answers to the questions asked.

Having given this quick intro of what to expect in this document lets get into the problem sets and the solutions for each.

Problem 1

Overview

We have been given data about foot injuries of player and the positions at which they play, A physiotherapist is interested in studying if at all there is a relationship between foot injuries and the position at which a player plays.

Dataset

The data is in the form of a contingency table across the players that are injured and not injured, and across the positions played Striker, Forward, Attacking Midfielder and Winger.

Players\Position	Striker	Forward	Attacking Midfielder	Winger	Total
Players Injured	45	56	24	20	145
Player Not Injured	32	38	11	09	90
Column Total	77	94	35	29	235

Table 1 Contingency Table Injured and Not Injured Players vs Position Played

Objective

Using Table 1 above we can get the various probabilities of a player playing in a certain position to get a foot injury or not, using this information we will answer the questions asked below.

Questions Asked

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

Using Table 1 we have the number of players injured in the sample data is equal to **145**, The total number of players are **235**.

$$P(Injury) = \frac{Number\ of\ players\ Injured}{Number\ of\ total\ Players} = \frac{145}{235} = 0.6170\ \approx 62\%$$

Equation 1 Calculation for probability of Injury of a randomly chosen player

The probability that a randomly chosen player would suffer an injury is **0.6170** or about **62%** of the time.

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

Using Table 1 we have the number of players who play as Forward is **94** and the number of players who play as Winger is **29**, The total number of players is **235**.

$$P(Forward\ or\ Winger) = \frac{Number\ of\ player\ as\ Forward + Number\ of\ palyers\ as\ Winger}{Number\ of\ total\ Players}$$

$$= \frac{(94+29)}{235} = 0.5234\ \approx 52\%$$

Equation 2 Calculation for probability of Player in Forward or Winger position

The probability that a player is a Forward or Winger is **0.5234** or **52%** of the time.

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

Using Table 1 we have the number of Players with Injury and position Striker **45**, The number of players in Striker Position **77**.

$$P(Injury|Striker) = \frac{Number\ of\ Strikers\ Injured}{Number\ of\ Strikers} = \frac{45}{77} = 0.5844\ \approx 58\%$$

Equation 3 Calculation of probability of Player in Striker position getting Injured

The probability that a player is a Striker and gets injured **0.5844** or **58%** of the time.

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

Using Table 1 we have the number of players with injury and position Striker **45**, The number of injured players **145**.

$$P(Striker|Injury) = \frac{Number\ of\ Players\ Injured\ and\ Stricker}{Number\ of\ Players\ Injured} = \frac{45}{145} = 0.3103 \approx 31\%$$

Equation 4 Calculation of probability of injured Player being in Striker position

The probability that an injured Player is a Striker is **0.3103** or **31%** of the time.

Problem 2

Overview

We have been given some Population Parameters around the breaking strength of Gunny Bags in kg per sq. centimetres, knowing what is given to us, we should be able to construct the **Probability density function(pdf)** for the breaking strength. It is specifically asked that we provide an appropriate visual representation for our answers.

Information Given

The breaking strength of Gunny Bags is normally distributed, the mean of the said distribution **μ = 5 kg per sq. centimetres**, the standard deviation is given to be **σ = 1.5 kg per sq. centimetres**.

Using this information, we can construct and plot the **pdf** for the breaking strength of Gunny Bags, which would follow a normal distribution. Steps for doing this is given in Appendix B.

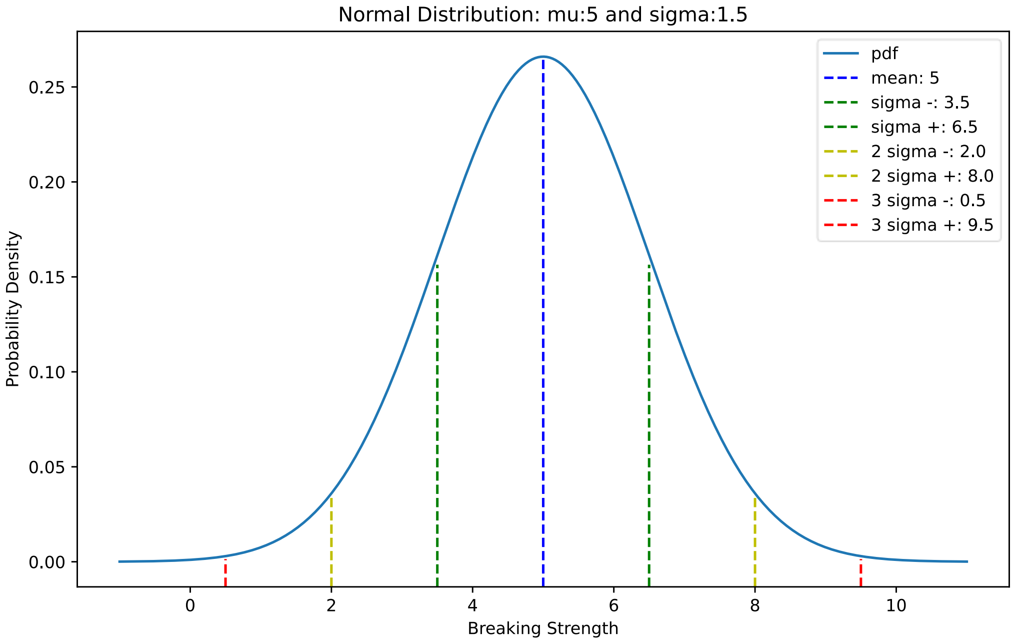


Figure 1 Normal Distribution of Breaking Strength in kg per sq. cm.

Figure 1 is a normal distribution which follows the following properties:

- 68.27% of the values lie between Mean ± Standard Deviation, in this case between the values of 3.5 and 6.5 kg per sq. cm.
- 95.45% of the values lie between Mean ± 2*Standard Deviation, in this case between the values of 2 and 8 kg per sq. cm.
- 99.73% of the values lie between Mean ± 3*Standard Deviation, in this case between the values of 0.5 and 9.5 kg per sq. cm.
- Reference: Normal distribution - Wikipedia

Objective

Using the properties that we know as per Information Given, we can answer the Questions Asked

Questions Asked

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

Using the properties of the normal distribution plotted (Figure 1), we can find where the value **3.17** lies on the x-axis, once we have this value, we can calculate the cumulative probability for all the values that are less than **3.17** this would give us the proportion of gunny bags that have a breaking strength of less than **3.17 kg per sq. cm**, Steps for doing this is given in Appendix C.

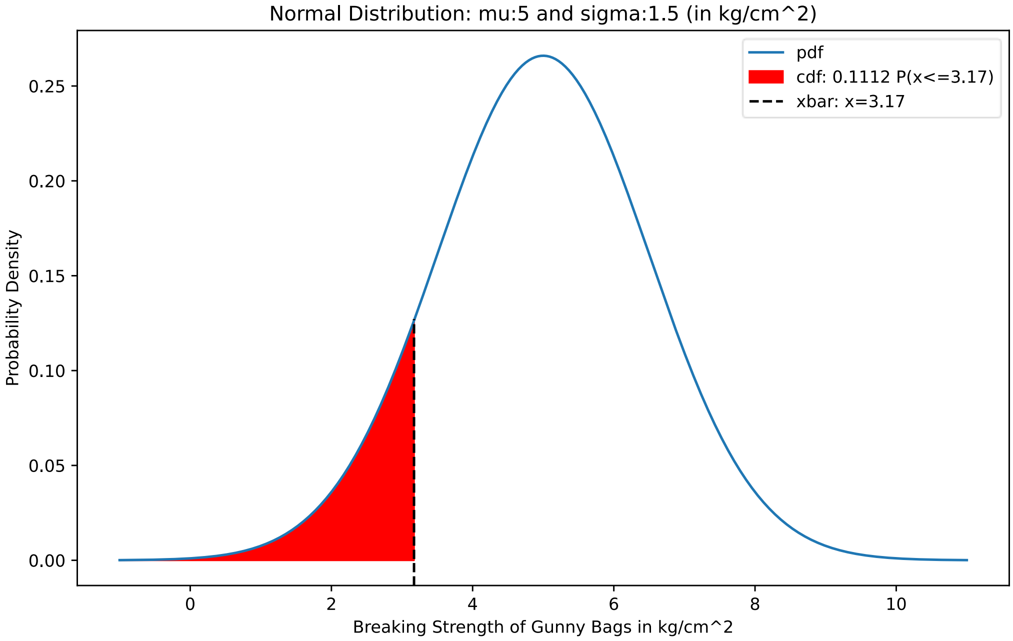


Figure 2 Normal distribution of Breaking Strength of Gunny Bags, showing where the check value of 3.17 lies (black dotted line) and the proportion that lies below the check value (red shaded area)

We see that **11.12%** (Figure 2 ***cdf value***) of Gunny Bags that have a breaking strength below **3.17 kg sq. cm.**

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

Using the properties of the normal distribution plotted (Figure 1), we can find where the value **3.6** lies on the x-axis, once we have this value, we can calculate the cumulative probability for all the values that are greater than or equal to **3.6** this would give us the proportion of gunny bags that have a breaking strength of greater than **3.6 kg per sq. cm**, Steps for doing this is given in Appendix C.

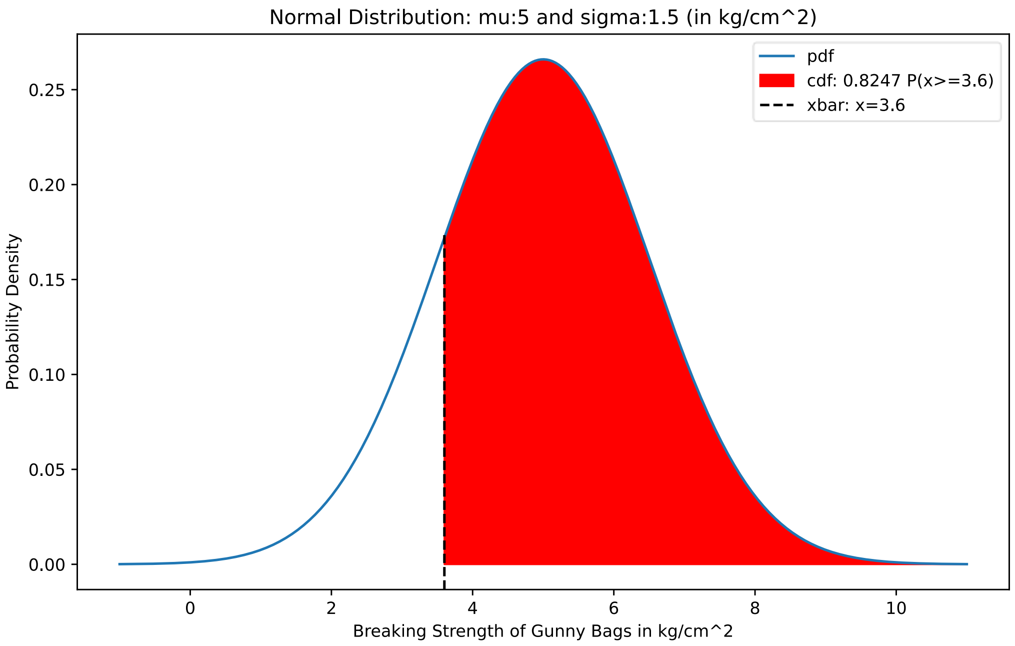


Figure 3 Normal distribution of Breaking Strength of Gunny Bags, showing where the check value of 3.6 lies (black dotted line) and the proportion that lies above the check value (red shaded area)

We see that **82.47%** (Figure 3 ***cdf value***) of Gunny Bags that have a breaking strength below **3.17 kg sq. cm.**

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

Using the properties of the normal distribution plotted (Figure 1), we can find where the value of **5 and 5.5** lies on the x-axis, once we have these values, we can calculate the cumulative probability for all the values that are between **5 and 5.5** this would give us the proportion of gunny bags that have a breaking strength of between **5 and 5.5 kg per sq. cm**, Steps for doing this is given in Appendix C.

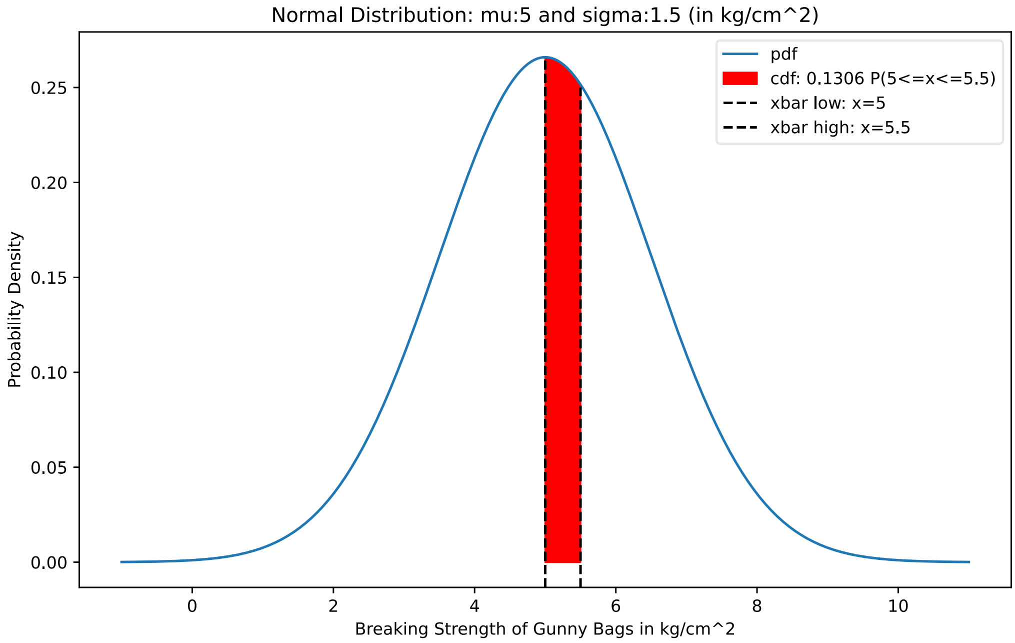


Figure 4 Normal distribution of Breaking Strength of Gunny Bags, showing where the check value of 5 and 5.5 lies (black dotted lines) and the proportion that lies between these check values (red shaded area)

We see that **13.06%** (Figure 4 ***cdf value***) of Gunny Bags that have a breaking strength between **5 and 5.5 kg sq. cm.**

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

Using the properties of the normal distribution plotted (Figure 1), we can find where the value of **3 and 7.5** lies on the x-axis, once we have these values, we can calculate the cumulative probability for all the values that are less than **3 and** greater than **7.5** this would give us the proportion of gunny bags that have a breaking strength of less than **3 and** greater than **7.5 kg per sq. cm**, Steps for doing this is given in Appendix C.

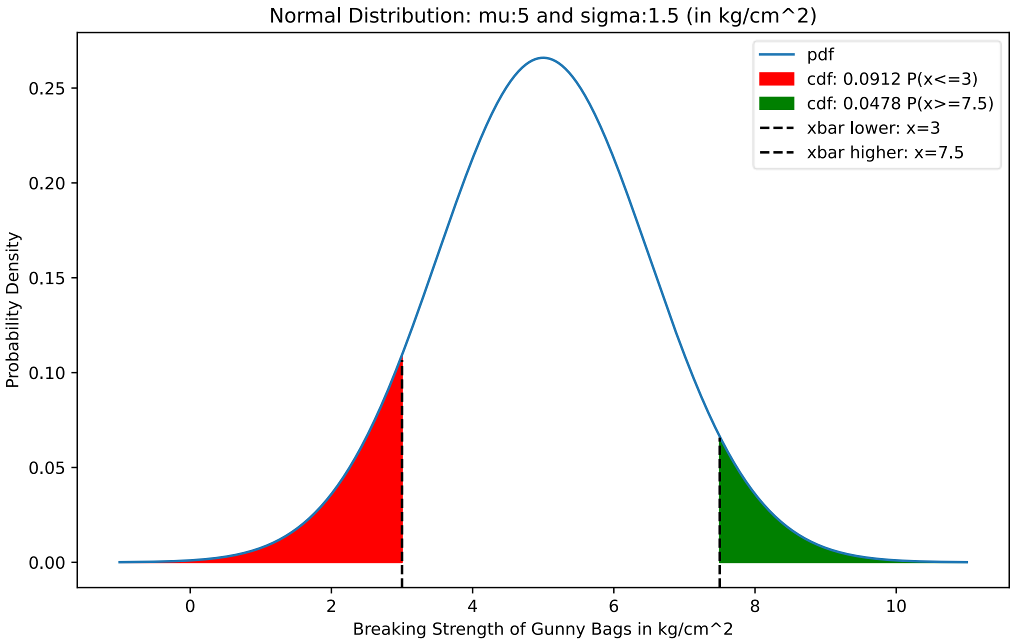


Figure 5 Normal distribution of Breaking Strength of Gunny Bags, showing where the check value of 3 and 7.5 lies (black dotted lines) and the proportion that does NOT lie between these check values (red and green shaded area)

We see that **13.09% (cdf lower +cdf upper where: cdf lower 9.12% and cdf upper 4.78%)** (Figure 5 ***cdf value***) of Gunny Bags that have a breaking strength that do **NOT** lie between **3 and 7.5 kg sq. cm.**

Problem 3

Overview

Zingaro stone printing is a company that specializes in printing images or patterns on both polished and unpolished stones. The quality of the outcome of the printing is dependent on the **Brinell's hardness index(BHI)** of the material being printed on, for optimum results the stone surface should have a **BHI** of **at least 150 or higher**, we have been given the data for a recent batch of polished and unpolished stones that the company has received from its clients, using the data provided we need to answer the Questions Asked, for all the tests that we run to answer the questions it is said that we can consider a **5% significance level.**

Dataset

The dataset provided has 75 rows/observations in it and has two columns: "Unpolished" and "Treated and Polished".

The "Unpolished" column has the **BHI** for unpolished Stones and the "Treated and Polished" column has the **BHI** for polished Stones.

Exploring the dataset, we see that:

- The mean BHI of the two datasets are as follows: Unpolished=134.11 and Polished = 147.79
- The standard deviation of the BHI for the two datasets are as follows: Unpolished=33.04 and Polished = 15.59

Objective

Given the dataset run the applicated statistical tests to be able to answer the questions asked

Questions Asked

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?

In order to test that the assumption made by Zingaro that unpolished stones may not be suitable for printing, we first need to review what has been given in by the dataset, and based on the same choose a statistical test we could do to check for the assumption.

Given are the following:

- The significance level for test that we do **α = 0.05**
- The BHI value to check against would be **μ = 150**, as anything under 150 would not be suitable for printing
- We can get the mean BHI for unpolished stones **xbar =134.11**
- We can get the std of BHI for unpolished stones **s = 33.04**
- Sigma = s/n^1/2 = **3.82**
- The number of observations is **n=75**

Since we have the mean and std of unpolished stones we can calculate the t-statistic using the Equation 6.

Let us state the Null Hypothesis and Alternate Hypothesis for this test

- Null Hypothesis for our t_test **H_0 : Mean BHI of Unpolished stones >= 150 BHI**, which would mean there are suitable for printing.
- Alternate Hypothesis for our t_test **H_A: Mean BHI of Unpolished stones < 150 BHI**, which would mean they are NOT suitable for printing.
- We are doing a one-sided test. (left side)

The **t-statistic calculated as per the Equation 6 = -4.1646**, for this value of the t-statistic and degrees of freedom of 74 (n-1), using the t.cdf function from stats(reference: [scipy.stats.t — SciPy v1.13.0 Manual]), we get the **pvalue= 0.0000417**, since the pvalue is less than the value of **∝ = 0.05**, we can reject the Null Hypothesis with a 95% Confidence level and say that :

H_A: Mean BHI of Unpolished stones < 150 BHI, which would mean they are NOT suitable for printing, Zingaro is justified in thinking that the unpolished stones may not be suitable for printing.

In other to visualize the test done here is the below Figure 6, the plot gives the t-distribution on which the test is run t-critical is the value below which if the t-statistic lies then we reject the Null Hypothesis, which in this case has happened.

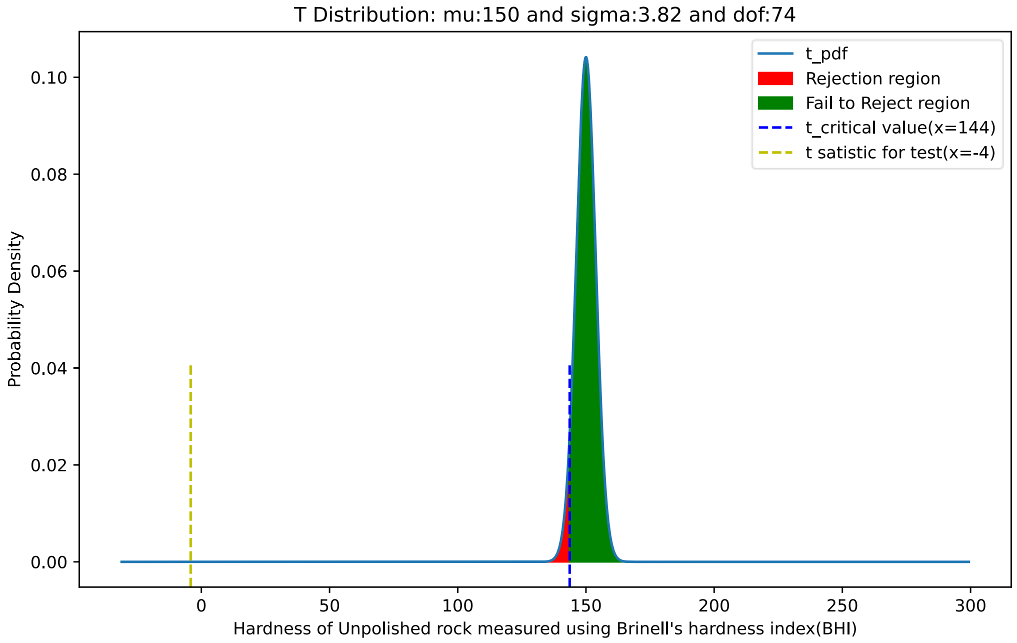


Figure 6 T-Distribution with the mean of 150 and std of 3.82 and dof of 74

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

In order to test if the mean hardness of the polished and unpolished stones the same, we will run the stats.ttest_ind test.(reference: [scipy.stats.ttest_ind — SciPy v1.13.0 Manual]).

There are few assumptions for this test:

- We will use a significance level of **0.05** for all tests
- Independence of the observations: The Assumption is that the observations here are independent of each other.
- No Significant outliers in the two samples
- Normality of the samples
- Homogeneity of variances

We test for each of these assumptions:

- It is understood from the business context that the observations are Independent.
- No Significant outliers in the two samples

  - Checking the data we see that there are four outlier values, three in the polished sample and one in the unpolished sample
  - We treat the same using the 1.5*IQR method.
- Normality of the samples
  - We run the "Shapiro-Wilk test" (reference: scipy.stats.shapiro — SciPy v1.13.0 Manual)
  - This tells us that in this case for both the samples, The data is consistent with a normal distribution.
- Homogeneity of variances
  - We run the "Levene's test "(reference: scipy.stats.levene — SciPy v1.13.0 Manual)
  - This tells us that in this case, The variance are not equal for this pair of samples

Now we run the stats.ttest_ind test.(reference: scipy.stats.ttest_ind — SciPy v1.13.0 Manual).

- This gives us a pvalue of "**0.00156**" since is lower than the significance level of **0.05**, we reject the Null Hypothesis of the Levene test which is "H_0 : There No significant difference in means, or means are equal."

This means with a 95% confidence level we can say that the mean hardness is significantly different between the polished and unpolished stones.

To visualize the two samples as a distribution and the t_test here are the two figures below (Figure 7,Figure 8).

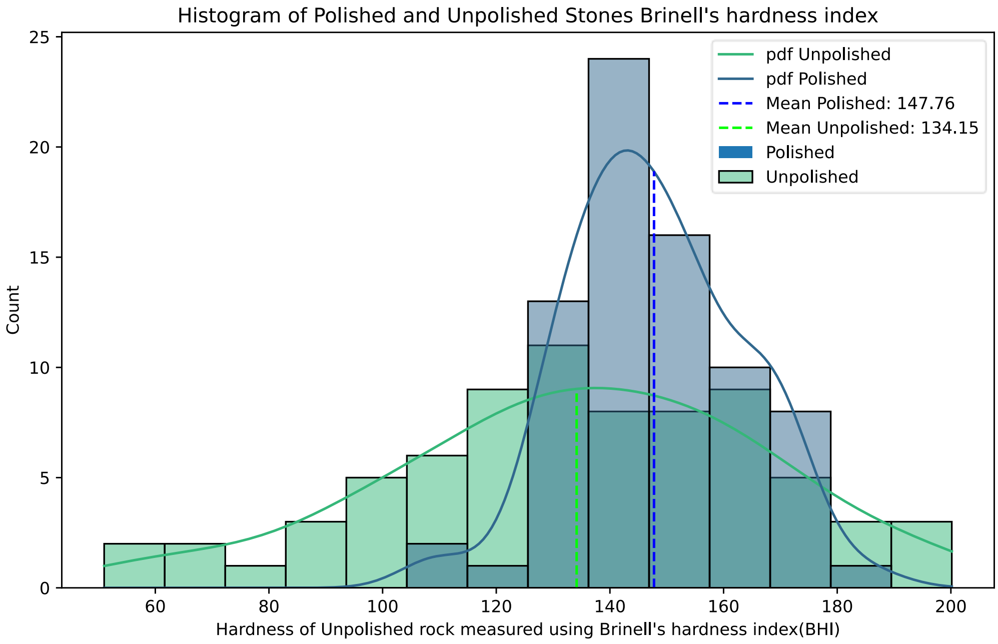


Figure 7 Histograms for polished and unpolished stones, with the KDE plot and the means plotted.

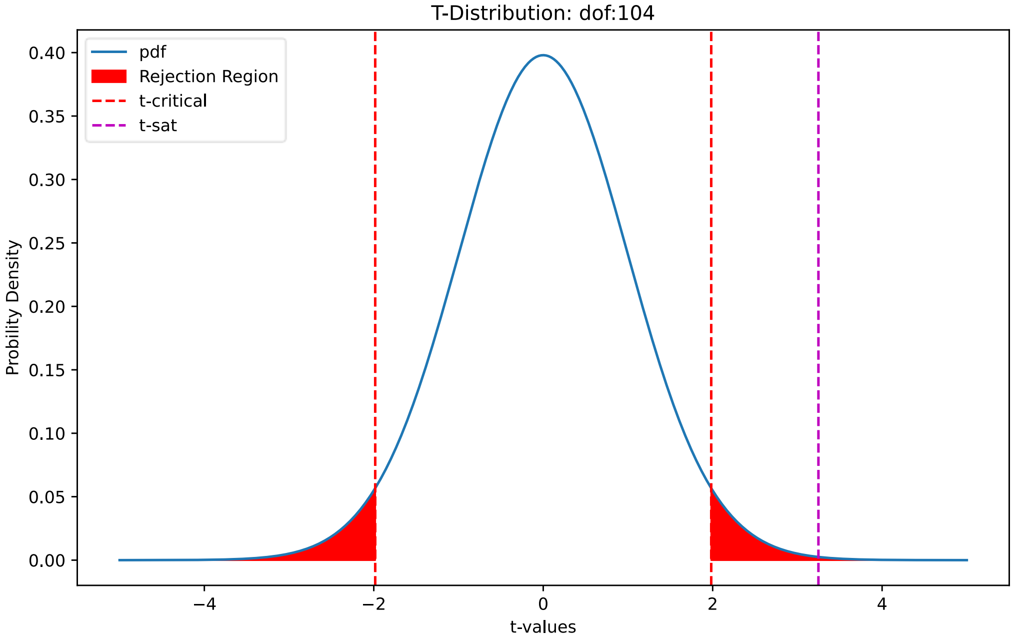


Figure 8 t-disputation for the ttest_ind, with the t-critical points and the t-sat plotted

Problem 4

Overview

The hardness of Dental Implant may dependant on a few factors, the Alloy used, the Method followed, the Dentist doing the procedure and the Temp at which the Method is done, given is a dataset with all the said variables, using this are to look at the Questions Asked.

Dataset

The data set given has three factors: Dentist, Method, Alloy, Temp, the dependant variable or variable of interest is Response, which is the indicator of the hardness of the metal implants.

Exploring the dataset this this what we find, Dentist, Method, Alloy, and Temp are all categorical variables. Dentist has 5 levels, Method has 3 levels, Alloy has two levels and Temp has 3 levels.

The Overall number of observations are 90, the number of observations for each level in Dentist are 18, the number of observations for each level in Method are 30, the number of observations for each level in Alloy are 45, the number of observations for each level in Temp are 30.

All the Questions Asked would need ANOVA tests which rely on the following assumptions:

- The populations are normal.
- The observations are independent.
- The variances from each population are equal.
- The groups must have equal sample sizes.
- Reference: 11.3: Two-Way ANOVA (Factorial Design) - Statistics LibreTexts

We will test all above assumptions before attempting to answer the questions, the level of significance used for all the test will be 0.05, This is done in Appendix D.

Take aways after testing the assumptions, use the "Response" transformed with box-cox transformation as the variable of interest, the variances is not equal when we include Dentist, Note to check with business if we can get more observations to increase the overall observations for Dentist levels which could help resolve this issue.

Objective

The dataset is a collection of factors with multiple levels, we will use this dataset to answer the below questions

Questions Asked

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

We have a look at the point plot of the levels in the Dentist factor/group.

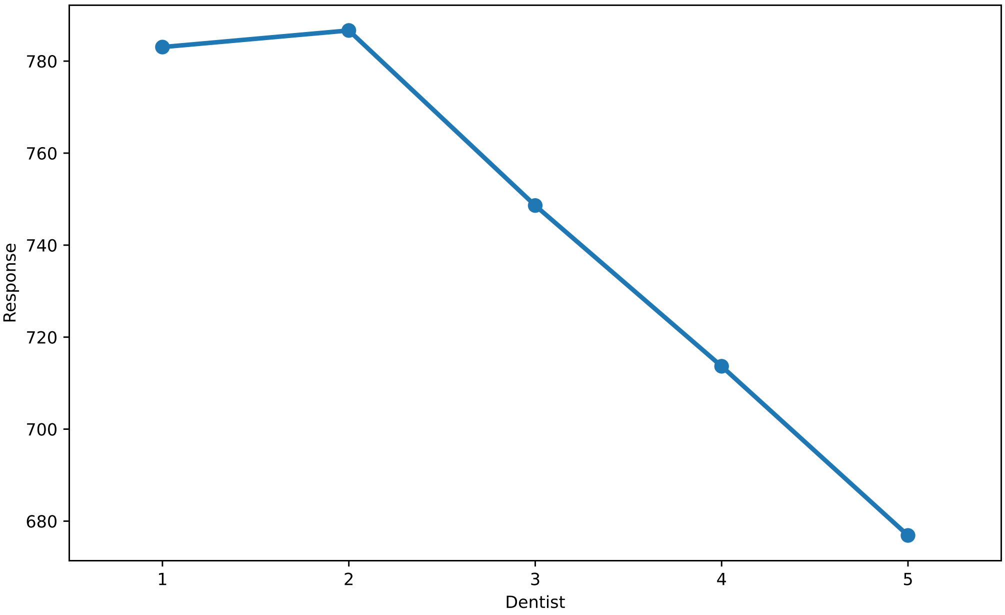


Figure 9 The point plot for Dentist Group plotting the mean response per level in Dentist

Looking at the plot it seems to vary but is it significant, we will test this using the One Way ANOVA test (reference: scipy.stats.f_oneway — SciPy v1.13.0 Manual)

H0: All the samples have the same population mean.

HA: At least one sample have different mean to the population mean.

We run the One-Way ANOVA test with the input of Dentist level 1 to 5 as separate samples

The pvalue for this is "0.2425", since the pvalue is greater alpha value of 0.05 we **Cannot** reject the Null Hypothesis.

So, in this case the H0: All the samples have the same population mean, the hardness of the implants is the same and do not vary significantly depending on dentist

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

We have a look at the point plot of the levels in the Method factor/group.

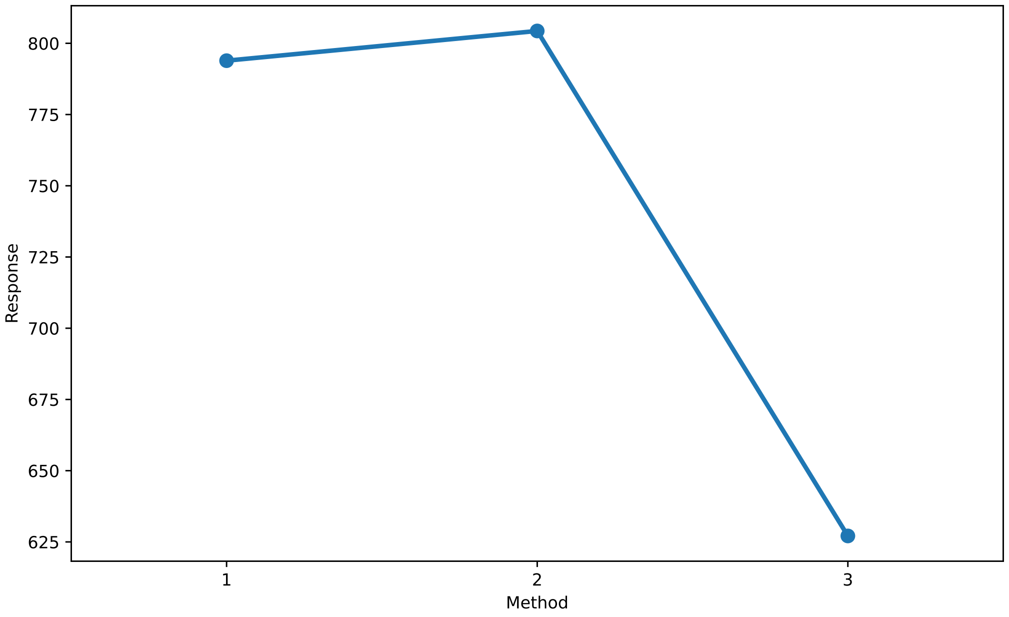


Figure 10 The point plot for Method Group plotting the mean response per level in Method

The point plot seems to show that means do vary especially for the level 3, but is it significant, we will test this using the One Way ANOVA test (reference: [scipy.stats.f_oneway — SciPy v1.13.0 Manual])

H0: All the samples have the same population mean.

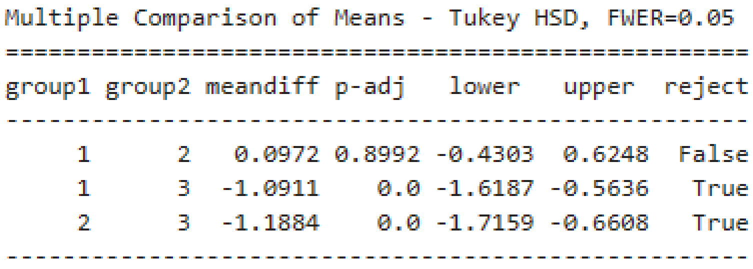
HA: At least one sample have different mean to the population mean.

We run the One-Way ANOVA test with the input of Method level 1 to 3 as separate samples

The pvalue for this is "0.0000003", since the pvalue is less than alpha value of 0.05 we reject the Null Hypothesis.

So, in this case the HA: At least one sample have different mean to the population mean, in order to find the means that vary we need to run the Tukey HSD test (reference: [statsmodels.stats.multicomp.pairwise_tukeyhsd - statsmodels 0.14.1])

Running the Tukey HSD test gives the following results.



```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
     1      2   0.0972 0.8992 -0.4303  0.6248  False
     1      3  -1.0911    0.0 -1.6187 -0.5636   True
     2      3  -1.1884    0.0 -1.7159 -0.6608   True
-----------------------------------------------------
```

Figure 11 Tukey HSD test results

The Tukey HSD test shows that, the hardness of implants does not vary between Methods 1 and 2, it varies significantly for Method 3 and on an average is lower compared to Methods 1 and 2.

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

We have a look at the point plot for Dentist for different method for Alloy 1 and Alloy 2.

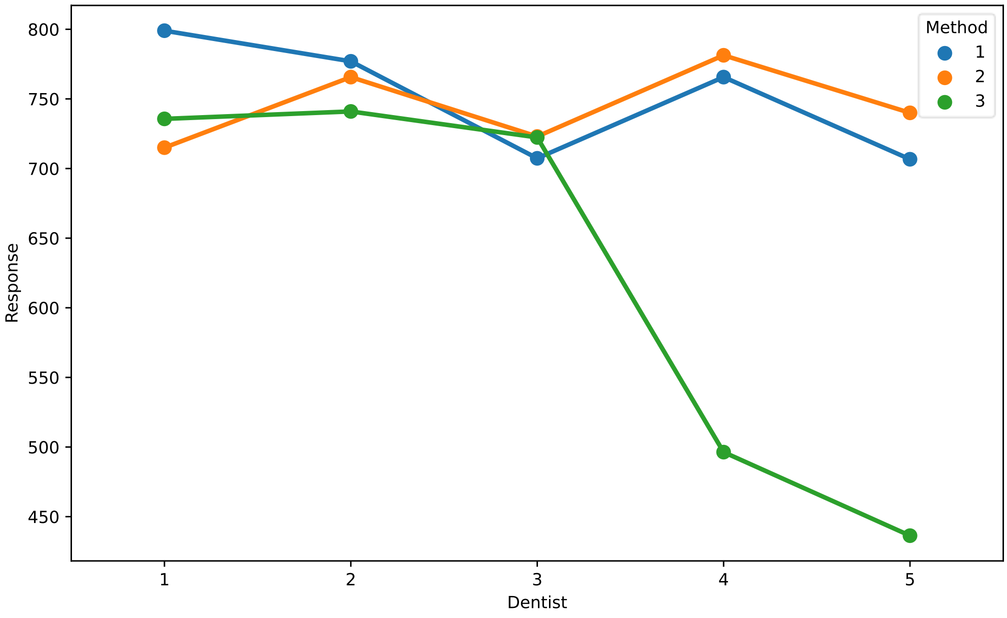
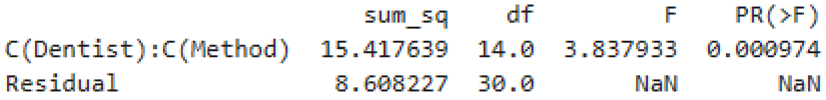


Figure 12 Point Plot of Dentist by Method for dataset which is filtered for Alloy 1

The Alloy 1 dataset point plot shows a lot of intersections which could indicate significant impart of interaction between Dentist and Method.

We test this with a two way ANOVA test (reference: statsmodels.stats.anova.anova_lm - statsmodels 0.15.0 (+243)), running this test gives the following result.



```
                        sum_sq    df         F    PR(>F)
C(Dentist):C(Method)  15.417639  14.0  3.837933  0.000974
Residual               8.608227  30.0       NaN       NaN
```

Figure 13 Alloy 1 dataset test results of interaction between Dentist and Method.

The Alloy 1 data set gives a pvalue of 0.000974 which is significant and shows a high interaction effect between Dentist and Method used.

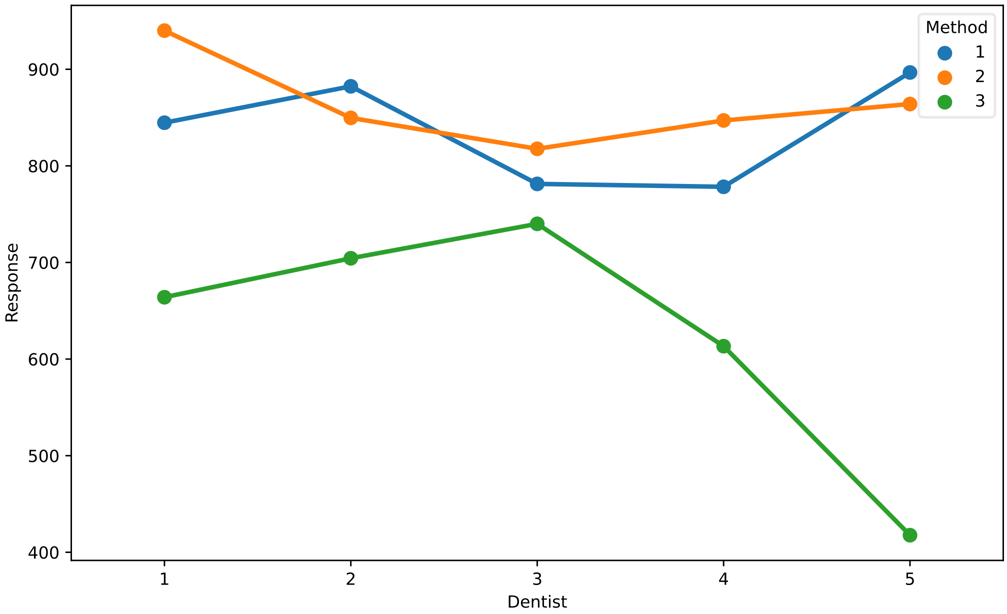
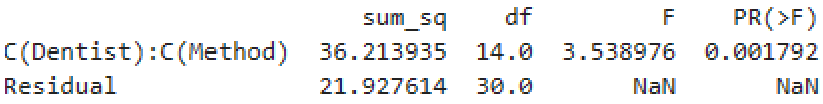


Figure 14 Alloy 2 dataset test results of interaction between Dentist and Method.

The Alloy 1 dataset point plot shows intersections of the Method 1 and 2 samples which could indicate significant impart of interaction between Dentist and Method.

We test this with a two way ANOVA test (reference: statsmodels.stats.anova.anova_lm - statsmodels 0.15.0 (+243)), running this test gives the following result.



```
                         sum_sq    df         F    PR(>F)
C(Dentist):C(Method)  36.213935  14.0  3.538976  0.001792
Residual              21.927614  30.0       NaN       NaN
```

Figure 15 Alloy 2 dataset test results of interaction between Dentist and Method.

The Alloy 1 data set gives a pvalue of 0.001792 which is significant and shows interaction effect between Dentist and Method used.

When we compare the interaction of Dentist and Method in datasets of Alloy 1 and Alloy 2 as an impact on Response or Hardness, three is an impart in both the Alloy 1 and 2 datasets but it is significantly higher for Alloy 1 compared to Alloy 2.

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

We have a look at the point plot for Dentist and Method.

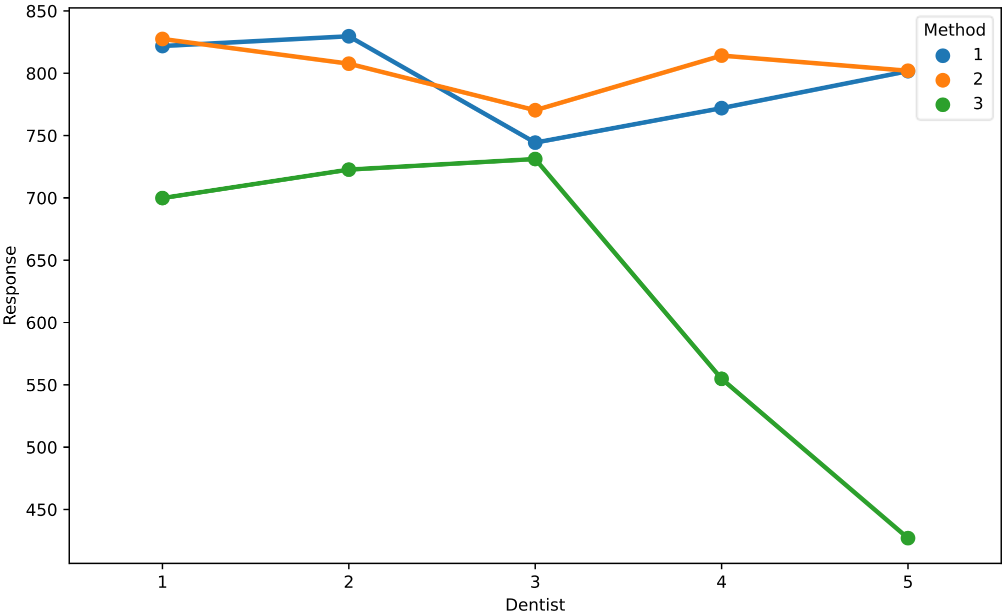
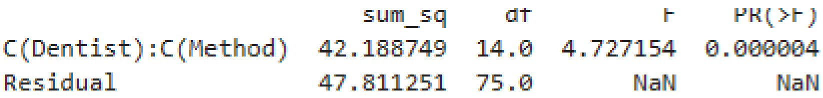


Figure 16 point plot dentist and method

Run the two way ANOVA on Dentist and Method interaction



```
                        sum_sq    df         F    PR(>F)
C(Dentist):C(Method)  42.188749  14.0  4.727154  0.000004
Residual              47.811251  75.0       NaN       NaN
```

Figure 17test result interactions between Dentist and Method

Looking at the plot and checking the interaction here is the result.

Dentist 1 and 4, gets the best response for hardness Using Method 2

Dentist 2 and 5, gets the best response for hardness Using Method 1

Dentist 3, gets the best response for hardness Using Method 3

Appendix A

Z distribution is made up of Z-scores, which is calculated as below for a sample when the population mean "μ" and standard deviation "σ" is known, "xbar" is the sample mean, "n" is the number of observations in the sample.

Equation 5 Z score formula

$$Z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

T distribution is made up of t-scores, which is calculated for a sample when the population mean "μ" is known or hypothesised, "xbar" is the sample mean, "s" is the standard error of the sample and "n" is the number of observations, "n-1" is the degrees of freedom

Equation 6 t score formula

$$t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$$

F distribution is calculated as the ratio between the "variation between samples /variation within the samples"

Equation 7 F score formula

$$F = \frac{SSB}{SSW}$$

Equation 8 SSB formula

$$SSB = \frac{\sum(\bar{x}_s - \mu)^2}{n_s - 1}$$

Equation 9 SSW formula

$$SSW = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

In this project all of the calculation are done using the python SciPy package e.g., stats.norm.pdf, stats.norm.cdf, stats.t.ppf, stats.t.pdf, stats.ttest_ind etc, the reference material for all of these can be found at: "Statistical functions (scipy.stats) — SciPy v1.13.0 Manual"

Appendix B

1. Given:
   a. μ= 5 # in kg/cm^2 units
   b. σ = 1.5 # in kg/cm^2 units
2. Calculate the min and max value for a normal distribution:
   a. Min = μ - 4 * σ
   b. Max = μ + 4* σ
3. Get a series of points between Min and Max values
   a. Here we used the linspace function of the numpy package of python
   b. X = np.linspace(Min, Max, 10000)
4. Get the corresponding pdf function values for each X.
   a. Here we used the norm.pdf function from the stats package of SciPy for python
   b. pdf = stats.norm.pdf(x, loc = μ, scale = σ)
5. Plot X, pdf using the matplotlib.pyplot.plot function from matplotlib package for python

6. References:
    a. numpy.linspace — NumPy v1.26 Manual
    b. scipy.stats.norm — SciPy v1.13.0 Manual
    c. matplotlib.pyplot.plot — Matplotlib 3.8.4 documentation

Appendix C

1. Please note that for all Problem 2 Questions the method to follow remains the same just that the Xbar values and the calculations of the **Area under the curve (AUC) or CDF,** would change depending if we are looking for values less than, greater than, in between or NOT in between the Xbar values
    a. For less than values we would take the cdf values as returned from the function, as it gives us the cumulative probability less than equal to the xbar value.
    b. For greater than values we would take the 1- cdf values as returned from the function, as cdf gives us the cumulative probability less than equal to the xbar value so 1- cdf value would give us greater than equal to the xbar.
    c. For values between two Xbars let's say Xbar upper and Xbar lower, we would get the cdf value for Xbar upper and Xbar lower, and take the difference between the upper and lower to give us the proportion between the two values.
    d. For values NOT between two Xbars let's say Xbar upper and Xbar lower, we would get the 1- cdf value for Xbar upper, say cdf upper and cdf value for Xbar lower, say cdf lower, then add cdf upper and cdf lower to give us the proportion that does NOT lie between Xbar upper and Xbar lower.
2. Example:
    a. Value less than which to find the proportion of gunny bags, Xbar = 3.17
    b. Calculate the **Area under the curve (AUC)** for Figure 1 where x-axis values are < 3.17
        i. auc_lequal_xbar = stats.norm.cdf(xbar, loc=mu, scale=sigma)
        ii. The value auc_lequal_xbar is equal to 0.1112 or 11.12%
        iii. This tells us that 11.12% of Gunny Bags have a breaking strength < 3.17 kg per sq. cm.
3. To visualize this
    a. We plot Figure 1 and add a line at xbar = 3.17
    b. We then shade the **AUC** for all values under 3.17
    c. We add the appropriate legends to show the xbar value, the auc_lequal_xbar value which is the proportion of Gunny Bags having a breaking strength < 3.17 kg per sq. cm
4. Reference:
    a. scipy.stats.norm — SciPy v1.13.0 Manual

Appendix D

- We test for normality of the overall data and the factors
    - We run the "Shapiro-Wilk test" (reference: scipy.stats.shapiro — SciPy v1.13.0 Manual)
    - Results:
        - for Response, H_A: The data is NOT consistent with a normal distribution., with alpha:0.05 and pvalue:8.080212865024805e-06
        - for Dentist and level 1, H_0: The data is consistent with a normal distribution., with alpha:0.05 and pvalue:0.1856756955385208
        - for Dentist and level 2, H_0: The data is consistent with a normal distribution., with alpha:0.05 and pvalue:0.7310513854026794
        - for Dentist and level 3, H_0: The data is consistent with a normal distribution., with alpha:0.05 and pvalue:0.5520960092544556

        - for Dentist and level 4, H_A: The data is NOT consistent with a normal distribution., with alpha:0.05 and pvalue:0.002037237398326397
        - for Dentist and level 5, H_0: The data is consistent with a normal distribution., with alpha:0.05 and pvalue:0.5022943019866943
        - for Method and level 1, H_0: The data is consistent with a normal distribution., with alpha:0.05 and pvalue:0.41838711500167847
        - for Method and level 2, H_0: The data is consistent with a normal distribution., with alpha:0.05 and pvalue:0.07601878046989441
        - for Method and level 3, H_A: The data is NOT consistent with a normal distribution., with alpha:0.05 and pvalue:0.014201466925442219
        - for Alloy and level 1, H_A: The data is NOT consistent with a normal distribution., with alpha:0.05 and pvalue:1.1945070582441986e-05
        - for Alloy and level 2, H_A: The data is NOT consistent with a normal distribution., with alpha:0.05 and pvalue:0.00040293222991749644
        - for Temp and level 1500, H_A: The data is NOT consistent with a normal distribution., with alpha:0.05 and pvalue:0.023662082850933075
        - for Temp and level 1600, H_0: The data is consistent with a normal distribution., with alpha:0.05 and pvalue:0.47178080677986145
        - for Temp and level 1700, H_A: The data is NOT consistent with a normal distribution., with alpha:0.05 and pvalue:0.0001631420018384233
    - Looking at the results there are some samples which are NOT consistent with a normal distribution
    - To try and treat these, the "Response" variable is transformed using square root, cube root, log and box-cox transformation and the skew and kurtosis is checked for each of these transformations:
        - Response, Skew: -1.0585424288338496 Kurt: 2.0960859684721944
        - Response_root, Skew: -1.5344001316794174 Kurt: 3.172720390650926
        - Response_3root, Skew: -1.6919274459587494 Kurt: 3.6500251056632145
        - Response_log, Skew: -2.004559264063631 Kurt: 4.770739579859132
        - Response_boxcox_pt, Skew: 0.09842940792303596 Kurt: 1.8352363035678874
    - Post transformation too there is some skewness, but using the box-cox transformation gives us as close to normal as we can get, so for ANOVA tests we will run them on the transformed **Response_boxcox_pt** variable as the dependant variable.
- We test if the variance in the groups is equal
    - We run the "Levene's test "(reference: scipy.stats.levene — SciPy v1.13.0 Manual)
    - The Levene's test is run on all the different levels for all the factors
    - Results:
        - Dentist H_A: The variance is NOT equal for at least one pair of samples.
        - Method H_0: The variance is equal across all samples.
        - Alloy H_0: The variance is equal across all samples.
        - Temp H_0: The variance is equal across all samples.
        - All H_0: The variance is equal across all samples.
        - Dentist-Method H_A: The variance is NOT equal for at least one pair of samples.
        - Dentist-Method-Alloy H_A: The variance is NOT equal for at least one pair of samples.
    - The variance is NOT equal for combinations where the levels from Dentist group is looked at which could be because the number of observations for Dentist is low
    - Note: would be good to go back to the business and check if we could get more observations.