Inferential Statistics Project Report

by
Kanimozhi S

Contents

List of Tables

List of figures

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**

Table 1 : Problem 1 dataset

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

P(Injury) = Number of Injured Players / Total Number of Players

P = 145/235

Probability of Random player would suffer an Injury = 0.617

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

P(Forward or Winger) = Players who are Forward or Winger/Total Number of Players

P = 94+29/235

Probability that a player is a forward or a winger = 0.523

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

P(Striker and Injured) = Number of Strikers who are Injured/Total Number of Players

P = 45/ 235 , Probability that a random player plays in a striker position and has a foot injury = 0.191

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

P(Striker | Injured)=Striker and Injured/Injured

P = 45/145 , Probability that a randomly chosen injured player is a striker = 0.310

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

Given:

The breaking strength of gunny bags follows a normal distribution with

Mean μ= 5 kg/ sq.cm

Standard Deviation σ= 1.5 kg/sq.cm

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

Formula : $Z = X - \mu / \sigma$

μ= 5 kg/ sq.cm

σ= 1.5 kg/sq.cm

X = 3.17 kg/sq.cm

Z = 3.17 - 5 / 1.5

$P(Z < -1.22) \approx 0.1112$

$P(X<3.17) \approx 0.1112$

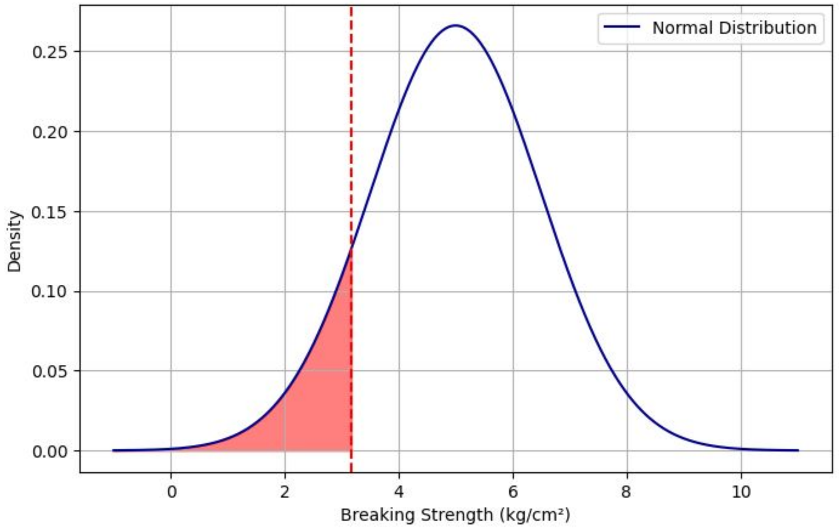


Fig 1: P(X< 3.17)

The proportion of gunny bags that have a breaking strength of less than 3.17 kg per sq cm is **0.1112 (or 11.12%)**. This means that about **11.12%** of the gunny bags are expected to have a breaking strength below this threshold, which could contribute to wastage or pilferage.

FYI: Used python to calculate the whole 2nd question and code is mentioned ipynb file

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

Formula : $Z = X - \mu / \sigma$

μ= 5 kg/ sq.cm, σ= 1.5 kg/sq.cm

X = 3.6 kg/sq.cm

Z = 3.6 - 5 / 1.5 = - 0.93

$P(X \geq 3.6) = 1 - P(Z < -0.93)$

$P(Z < -0.93) \approx 0.1753$

$P(X \geq 3.6) = 1 - 0.1753 = 0.8247$

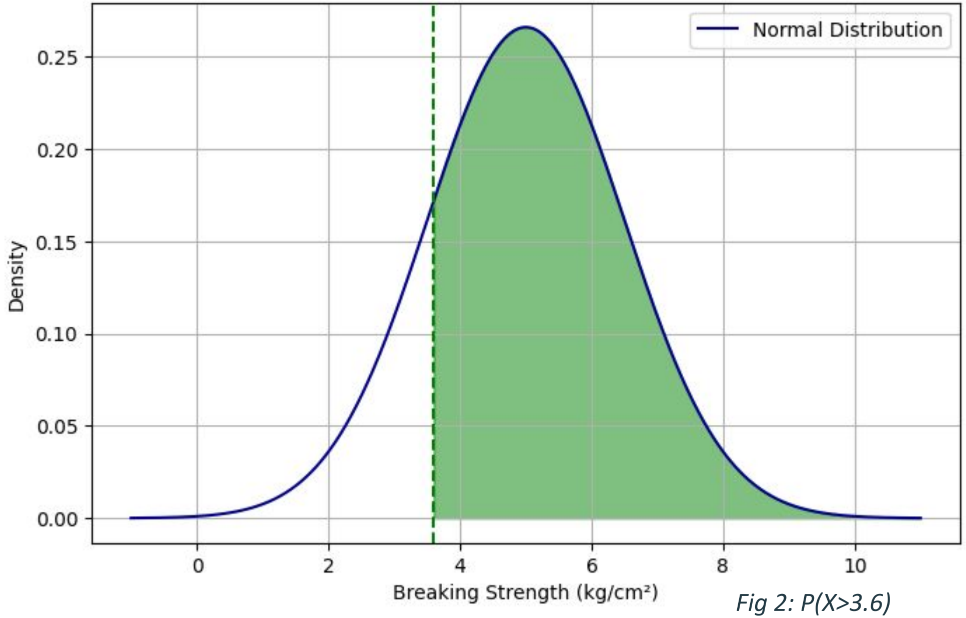


Fig 2: P(X>3.6)

The proportion of gunny bags that have a breaking strength of at least 3.6 kg per sq cm is approximately **0.8247 (or 82.47%)**.

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

Find the proportion where **5 ≤ X ≤ 5.5**

Formula : Z= X−μ / σ

- **For 5:**
  Z low = 5 – 5 / 1.5 = 0
- **For 5.5:**
  Z high = 5.5 – 5 / 1.5 ≈ 0.33

So, **proportion between 5 and 5.5**:

= 0.1306

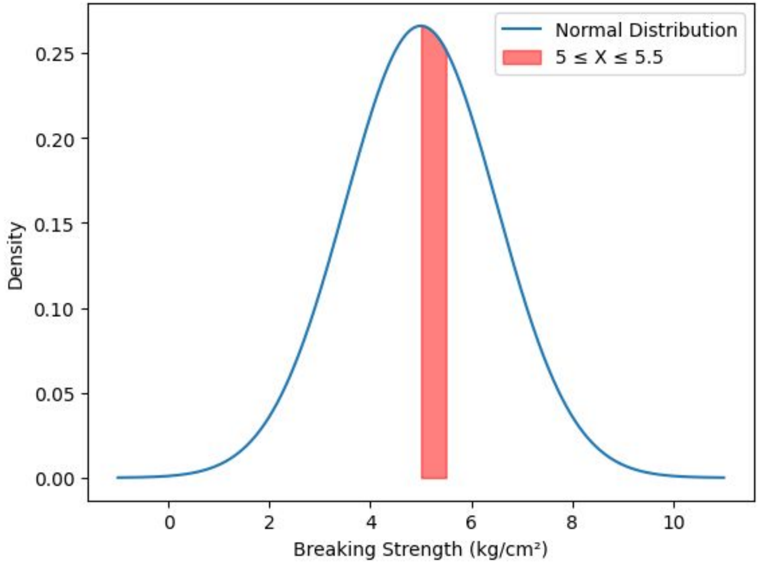


Fig 3: Proportion (5 ≤ X ≤ 5.5)

About 13.06 % of the gunny bags have a breaking strength between 5 and 5.5 kg/cm².

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

Find the proportion where **X < 3** or **X > 7.5**

Formula : Z= X−μ / σ

For 3 (left tail) Z low = 3 – 5 / 1.5 = – 1.33

For 5 (right tail) Z high = 7.5 – 5 / 1.5 = 1.67

P(X<3 or X>7.5)=Left tail + right tail = **0.1393**

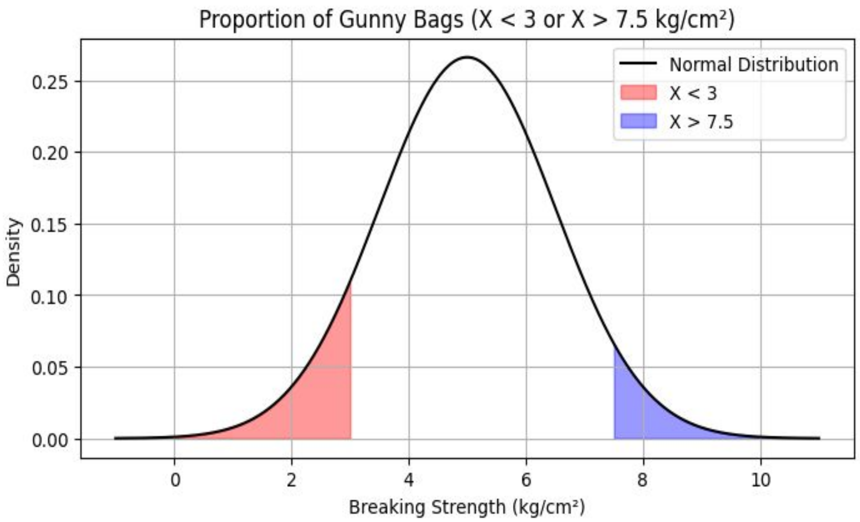


Fig 4: Proportion (X < 3 or X > 7.5)

About 13.9% 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);

Dataset Overview:

The dataset consists of two Variables:

VARIABLE	DESCRIPTION	TYPE
Unpolished	BHI (Raw stone)	**Continuous**
Treated and Polished:	BHI (Polished stone)	**Continuous**

Table 2 : 3.1 Variable type table

Each group contains **continuous numerical data** representing Brinell hardness index values. measured in **float64 and** there are **no categorical variables**, making the dataset suitable for statistical testing.

There are 75 rows and 3 columns in the dataset and first five rows of the datasets are below

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

Table 3: 3.1 Top 5 rows of the dataset

No missing or null values in either column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 75 entries, 0 to 74
Data columns (total 2 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Unpolished            75 non-null     float64
 1   Treated and Polished  75 non-null     float64
dtypes: float64(2)
memory usage: 1.3 KB
```

Table 4: 3.1 Basic information of dataset

Statistical summary:

	count	mean	std	min	25%	50%	75%	max
Unpolished	75.0	134.110527	33.041804	48.406838	115.329753	135.597121	158.215098	200.161313
Treated and Polished	75.0	147.788117	15.587355	107.524167	138.268300	145.721322	157.373318	192.272856

Table 5: 3.1 Numerical summarization of the dataset

Observations:

Treated and Polished stones have a **higher and more consistent** hardness (lower st deviation = 15.58), and the mean is close to the quality threshold (147.79 ≈ 150).

Unpolished stones show **more variability** (st deviation = 33.04), with a lower mean average (134.11), indicating less reliability in meeting the required BHI threshold.

Outlier Values:

Unpolished

48.406838

Treated and Polished

192.272856

107.524167

107.579388

Outliers were detected using the IQR method. However, all values appear to be valid and representative of natural variability in stone hardness. Thus, no clipping or removal was applied.

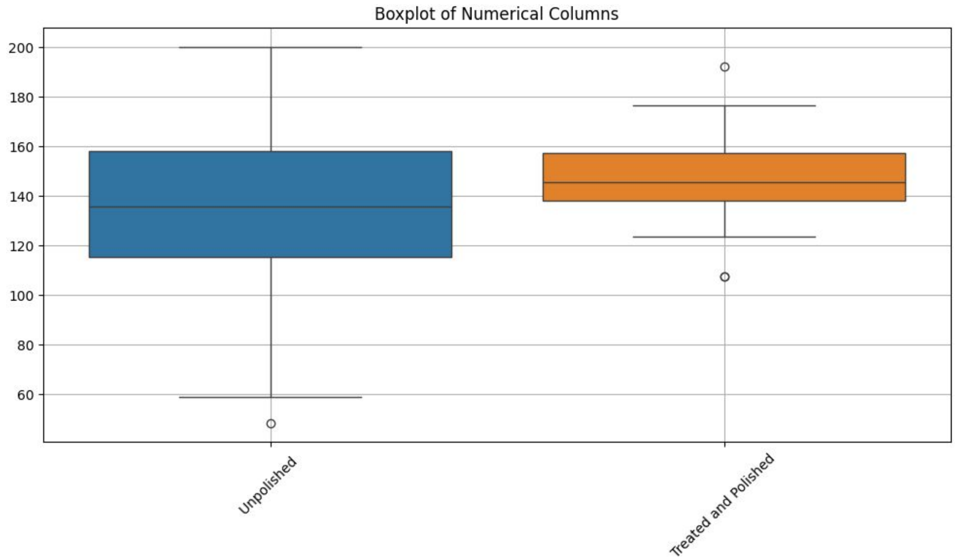


Fig 5: Outliers in Zingaro company dataset

Hardness Distribution – Treated & Polished Stones

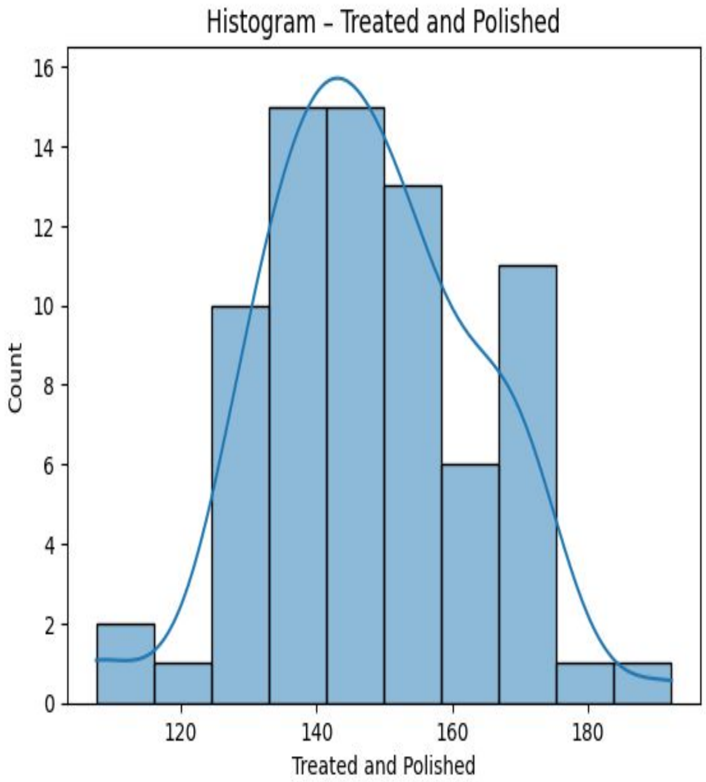


Fig 6: Hardness distribution

The histogram shows the distribution of Brinell Hardness Index values for treated and polished stones. The curve is approximately bell-shaped, indicating that the data is **roughly normally distributed**. This supports the use of **parametric tests** like the t-test for further analysis.

The Shapiro-Wilk test returns a p-value of 0.3314

Since p-value > 0.05, This means the data is **approximately normally distributed**, and it is appropriate to proceed with the **t-test**.

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?

Objective:

To determine whether the unpolished stones supplied to Zingaro meet the minimum required Brinell's hardness index of 150, which is essential for printing suitability.

Let μ be the mean Brinell's hardness index of the unpolished stones supplied to Zingaro.

Statistical Hypotheses:

We will test the null hypothesis

H_0: $\mu = 150$

against the alternative hypothesis

H_1: $\mu < 150$

Checking T-test Assumptions:

Continuous data: Yes, hardness index is measured on a continuous scale.

Normality: Sample size is 75 (which is large enough for t-test by Central Limit Theorem).

Random sample: Yes, the sample is considered random.

Unknown population standard deviation: Yes,

In this analysis, the company's operational concern is whether the unpolished stones are **not suitable for printing** that is, whether their mean hardness falls **below the minimum required standard** of 150 on the Brinell's hardness index. While a two-sided hypothesis test (H_0: $\mu = 150$, H_1: $\mu \neq 150$) would detect any difference from 150, it would also include cases where the mean is **higher** than 150, which does not represent a business risk.

Therefore, a **one-sided (left-tailed) test** was chosen, with the alternative hypothesis set as H_1: $\mu < 150$. This approach directly tests the company's concern: whether the average hardness is **less than** the required standard, indicating that the stones are potentially unsuitable for printing. This makes the test more relevant and statistically powerful for the given business context.

A **one-sample, left-tailed t-test** was conducted at a 5% significance level ($\alpha = 0.05$).

Output:

Mean Hardness (Unpolished): 134.11

t-statistic: -4.16

p-value: 0.00004

Conclusion :

- Since the p-value is much less than 0.05, we **reject the null hypothesis**.
- There is strong statistical evidence that the mean Brinell’s hardness index of unpolished stones is **less than 150**.
- This means the unpolished stones supplied in this batch are, on average, **not suitable for printing** as per Zingaro’s requirements.

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

Objective

To test whether there is a significant difference in the mean Brinell's hardness index between polished and unpolished stones supplied to Zingaro.

Let μ_1 be the mean Brinell's hardness index of polished stones, and μ_2 be the mean for unpolished stones.

Statistical Hypotheses

We will test the null hypothesis

H_0: $\mu_1 = \mu_2$

against the alternative hypothesis

H_1: $\mu_1 \neq \mu_2$

Checking T-test Assumptions:

Continuous data - Yes, Checked in EDA

Normally distributed populations - Yes, we have checked the normality distribution in EDA

Independent populations - Yes

Unequal population standard deviations - As the sample standard deviations are different, the population standard deviations may be assumed to be different.

Random sampling from the population - Yes,

t-test is appropriate, We are going with **Two Independent Sample T-test for Equality of Means - Unequal Std Dev**

Two Independent Sample T-test for Equality of Means - Unequal Std Dev was performed at a 5% significance level

(α = 0.05).

Output:

Mean Hardness (Polished): 147.79

Mean Hardness (Unpolished): 134.11

t-statistic: 3.24

p-value: 0.00159

Conclusion : Based on the results of the two independent sample t-test , the p-value is **0.00159**, which is much less than the significance level of 0.05. **We reject the null hypothesis.**

Interpretation:

There is strong statistical evidence that the mean Brinell's hardness index of polished stones is **significantly different** from that of unpolished stones.

Since the mean hardness for polished stones (**147.79**) is higher than that of unpolished stones (**134.11**), it is clear that **polishing the stones has a statistically significant positive effect on their hardness**.

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.

Data Overview:

The dataset consists of 5 variables

Although variables like **Dentist, Method, and Alloy** are represented numerically (e.g., values from 1 to 5), they are treated as categorical variables in the analysis. This is because these numbers function as labels identifying distinct groups, rather than representing measurable quantities or continuous scales.

Temp is also considered as a categorical variable and not continuous because it represents discrete, predetermined levels chosen during the experiment. It is therefore treated as a categorical variable. So, **Only Response is considered as a continuous variable and rest as discrete variables**

Variable	Description	Type	Data Type
Dentist	Identifier of dentist	Categorical	int
Method	Type of implant method used	Categorical	int
Alloy	Type of alloy material used	Categorical	int
Temp	Firing temperature	Categorical	int
Response	Hardness of Implants	Numerical	int

Table 6: 4.1 Data description table

There are 90 rows and 5 columns in the dataset and first five rows of the datasets are below

	Dentist	Method	Alloy	Temp	Response
0	1	1	1	1500	813
1	1	1	1	1600	792
2	1	1	1	1700	792
3	1	1	2	1500	907
4	1	1	2	1600	792

Table 7: 4.1 Top five rows of the dataset

No missing or null values in the columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90 entries, 0 to 89
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Dentist   90 non-null     int64
 1   Method    90 non-null     int64
 2   Alloy     90 non-null     int64
 3   Temp      90 non-null     int64
 4   Response  90 non-null     int64
dtypes: int64(5)
memory usage: 3.6 KB
```

Table 8: 4.1 Basic information of the dataset

Statistical summary:

	count	mean	std	min	25%	50%	75%	max
Dentist	90.0	3.000000	1.422136	1.0	2.0	3.0	4.0	5.0
Method	90.0	2.000000	0.821071	1.0	1.0	2.0	3.0	3.0
Alloy	90.0	1.500000	0.502801	1.0	1.0	1.5	2.0	2.0
Temp	90.0	1600.000000	82.107083	1500.0	1500.0	1600.0	1700.0	1700.0
Response	90.0	741.777778	145.767845	289.0	698.0	767.0	824.0	1115.0

Table 9: 4.1 Statistical summary

Observations:

The Response variable, representing implant hardness, is the only continuous variable in the dataset. It ranges from **289 to 1115**, with a **mean of 741.78** and **standard deviation of 145.77**, indicating high variability.

All other variables (Dentist, Method, Alloy, and Temp) are categorical and serve as grouping factors for the factorial ANOVA analysis.

Outlier check

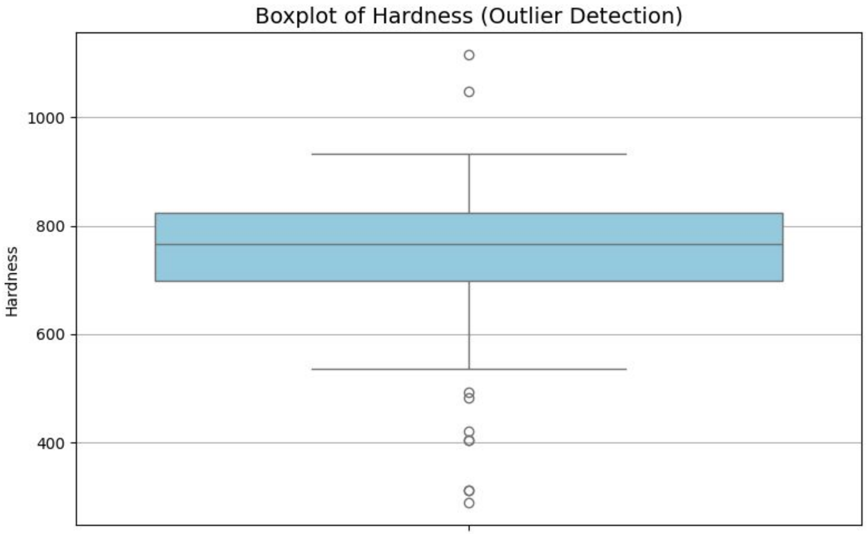


Fig 7: Outlier in dental hardness data

Outlier values:

Response:

1115, 1048, 493, 483, 421, 405, 405, 312, 312, 289

Outlier detection using the IQR method revealed 10 outliers in the Response variable for Alloy 1. However, all values appear plausible and are retained in the dataset. No other variables had outliers. The data quality is clean with no missing or duplicated rows.

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

Objective

To test whether there is a significant difference in the **mean implant hardness (Response)** across different **dentists** for **Alloy 1 and Alloy 2 separately** .

Let μ_1, μ_2, μ_3, μ_4, and μ_5 be the mean hardness values for dentists 1 to 5 respectively.

Statistical Hypotheses

We will test the **Null hypothesis**:

H0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

(The mean hardness is the same across all dentists)

against the **Alternative hypothesis**:

Ha: At least one μ differs

(The mean hardness differs for at least one dentist)

Checking Assumptions for both alloys:

Checked normality using shapiro wilk test

Output:

Shapiro-Wilk Test - Alloy 1: Statistic = 0.830, p-value = 0.00001

Shapiro-Wilk Test - Alloy 2: Statistic = 0.888, p-value = 0.00040

Conclusion: The Shapiro-Wilk test for Alloy 1 returned a p-value < 0.05, indicating a violation of the normality However, since the sample sizes are balanced and ANOVA is robust to such deviations, the analysis proceeds as per project instructions.

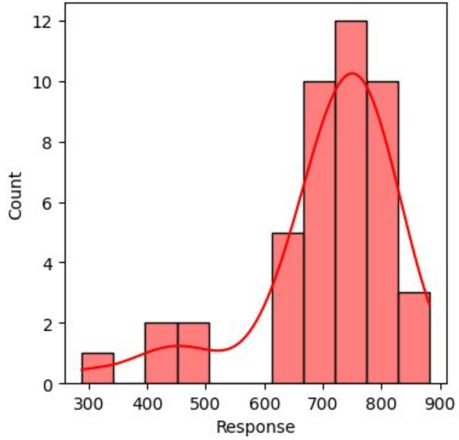
Levene's Test Output:

Levene's Test for Homogeneity of Variance (Alloy 1 - Dentist):

Statistic = 1.385, p-value = 0.25655

Levene's Test (Alloy 2): Statistic = 1.446, p-value = 0.23687

P value is larger than 0.05 , Levene's test showed equal variances, we proceed with the ANOVA as all of the assumptions are satisfied. This approach also aligns with the project guidelines.



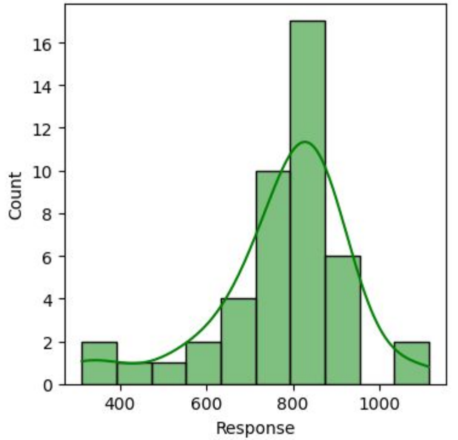


Fig 8: normality check for both alloys

One-way Anova test for Alloy 1

Output:

One-Way ANOVA (Alloy 1 - Dentist): F = 1.977, **p-value = 0.11657**

Conclusion:

A one-way ANOVA was conducted to test whether the mean implant hardness varies by dentist for Alloy 1. The test produced an F-statistic of **1.977** with a **p-value of 0.1166**. Since the p-value is greater than 0.05, we **fail to reject the null hypothesis**.

Therefore, we conclude that there is **no significant difference** in mean hardness between dentists for Alloy 1.

Now, we are doing a **Tukey HSD** test to compares all pairs of dentists to see if their mean implant hardness values differ.

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2  meandiff p-adj    lower    upper  reject
----------------------------------------------------
     1      2   11.3333 0.9996 -145.0423  167.709  False
     1      3  -32.3333 0.9757  -188.709 124.0423  False
     1      4  -68.7778 0.7189 -225.1535  87.5979  False
     1      5 -122.2222 0.1889 -278.5979  34.1535  False
     2      3  -43.6667 0.9298 -200.0423  112.709  False
     2      4  -80.1111 0.5916 -236.4868  76.2646  False
     2      5 -133.5556 0.1258 -289.9312  22.8201  False
     3      4  -36.4444 0.9626 -192.8201 119.9312  False
     3      5  -89.8889 0.4805 -246.2646  66.4868  False
     4      5  -53.4444 0.8643 -209.8201 102.9312  False
----------------------------------------------------
```

Table 10: 4.1 Tukey HSD output for alloy 1

As per the output
None of the pairwise comparisons are statistically significant (all `p-adj > 0.05`).

This confirms that **no specific dentist differs significantly** from another in terms of mean hardness supporting the **ANOVA result**.

One way Anova test for Alloy 2

Output:

ANOVA – Alloy 2 (Dentist): F-statistic = 0.525, p-value = 0.71803

Conclusion:

For **Alloy 2**, the ANOVA produced an **F-statistic of 0.525** with a **p-value of 0.718**
the **p-value is greater than 0.05**, indicating that the differences in mean hardness across dentists are **not statistically significant**.

Further, **Tukey HSD post-hoc tests** were performed to compare all possible pairs of dentists. These tests revealed **no significant pairwise differences** for either alloy, reinforcing the ANOVA results.

Tukey HSD - Alloy 2 (Dentist):
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-4.1111	1.0	-225.5687	217.3465	False
1	3	-36.5556	0.9895	-258.0131	184.902	False
1	4	-70.0	0.8941	-291.4576	151.4576	False
1	5	-90.1111	0.7724	-311.5687	131.3465	False
2	3	-32.4444	0.9933	-253.902	189.0131	False
2	4	-65.8889	0.9132	-287.3465	155.5687	False
2	5	-86.0	0.8008	-307.4576	135.4576	False
3	4	-33.4444	0.9925	-254.902	188.0131	False
3	5	-53.5556	0.9574	-275.0131	167.902	False
4	5	-20.1111	0.999	-241.5687	201.3465	False

Table 11: 4.1 Tukey HSD output for alloy 2

Conclusion: There is **no evidence** to suggest that the dentist has a statistically significant impact on implant hardness for either alloy.

The output revealed that **none of the pairwise differences** were statistically significant, with all adjusted p-values greater than 0.05

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

Objective

To test whether there is a significant difference in the mean implant hardness (**Response**) across different **methods** for **Alloy 1** and **Alloy 2** separately.

Let μ_1, μ_2, and μ_3 be the mean hardness values for Methods 1, 2, and 3 respectively.

Statistical Hypotheses

We will test the **Null Hypothesis**:

H0: $\mu_1 = \mu_2 = \mu_3$
(The mean hardness is the same across all methods)

against the **Alternative Hypothesis**:

Ha: At least one μ differs
(The mean hardness differs for at least one method

We have already done normality and levene checks in 4.1 problem and all assumptions are satisfied

One way Anova test for Alloy 1

Output:

F-statistic: 6.26, **p-value**: 0.0042
Since p-value < 0.05, we **reject the null hypothesis**. There is a statistically significant difference in implant hardness across methods for **Alloy 1**.

Conclusion:

The analysis clearly shows that **Method 3 yields significantly lower hardness values** than both **Method 1 and Method 2** for implants made from **Alloy 1**. There is **no significant difference between Method 1 and Method 2**, suggesting that those methods produce similar outcomes in terms of implant hardness.

This result has potential implications for **material processing protocols**, indicating that **Method 3 may not be optimal** for achieving desired implant strength when using Alloy 1.

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

Table 12: 4.2 Output-Tukey Hsd test for alloy 1

The Tukey HSD test revealed the following:

- **Method 1 vs Method 3** → significant difference (p = 0.0085)
- **Method 2 vs Method 3** → significant difference (p = 0.0128)
- **Method 1 vs Method 2** → no significant difference (p = 0.987)

One way Anova test for Alloy 2

Output: The **p-value (0.004163)** is well below the 0.05 significance level, leading us to **reject the null hypothesis**.

This confirms that **implant hardness varies significantly by method** for **Alloy 2**.

Conclusion:

For **Alloy 2**, the method of processing significantly impacts implant hardness. Specifically, **Method 3 leads to significantly lower hardness** than the other two methods.

Tukey HSD Results – Alloy 2 (Method):
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

Table 13: 4.2 Output-Tukey Hsd test for alloy 2

For **both Alloy 1 and Alloy 2**, the method of preparation has a **statistically significant impact** on implant hardness.

- In both alloys, **Method 3 consistently produces significantly lower hardness**.
- **Method 1 and Method 2** produce comparable results and appear more reliable.

These results suggest that **Method 3 is suboptimal** and may compromise implant durability.

Method 3 produces **significantly lower implant hardness** compared to both **Method 1 and Method 2**.

There is **no significant difference** between **Method 1 and Method 2**.

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

Interaction Plot for Alloy 1

Observation:

- The lines corresponding to different dentists are **not parallel**.
- Some lines **cross each other**, and there is noticeable divergence, especially between Method 2 and Method 3.

Inference:

- There is evidence of an **interaction effect** between **dentist and method** in Alloy 1.
- This suggests that the effect of the method on hardness **depends on the dentist** applying it.
- For instance, while Method 3 generally reduces hardness, some dentists show more drastic drops than others, indicating **inconsistent outcomes across operators**.

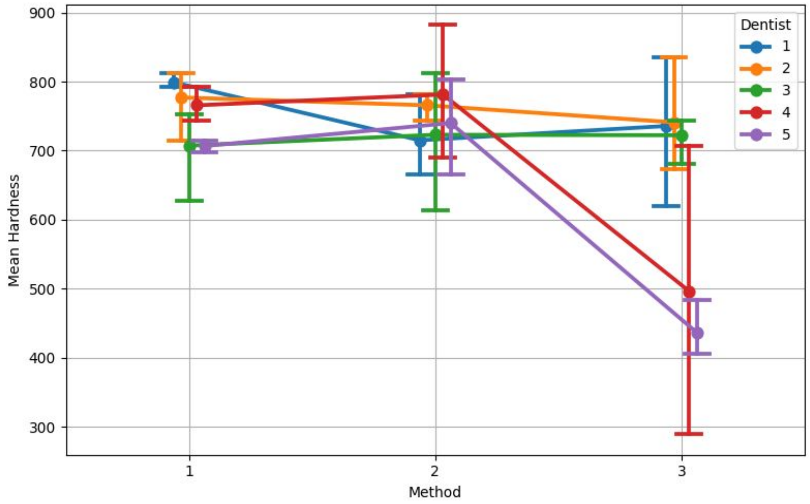


Fig 9: Interaction plot for Alloy 1

Interaction Plot for Alloy 2

Observation:

- The divergence and crossing of lines is even more **pronounced**.
- Particularly for Method 3, certain dentists (e.g., Dentist 4 and 5) show **sharp decreases in hardness**, while others remain more stable.

Inference:

- Visual evidence from the interaction plots suggests an **interaction effect between dentist and method** for **both alloys**.
- This indicates that **different dentists respond very differently** to the methods used, especially with Method 3, which produces **high variability in hardness**.
- Such interaction may compromise reliability unless techniques are standardized or dentist-specific adaptations are made.

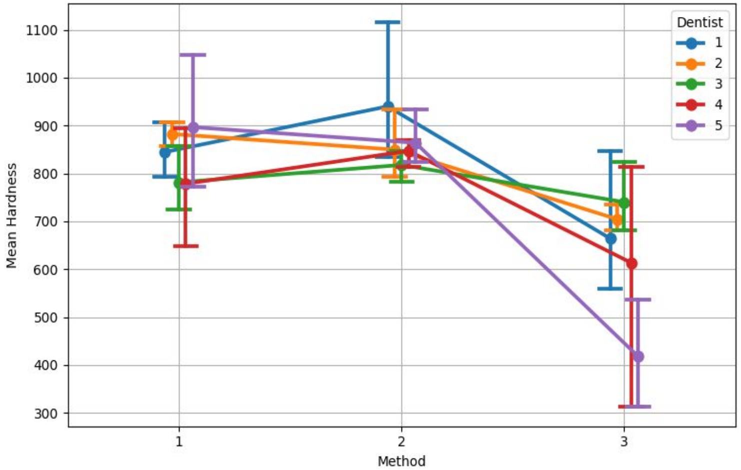


Fig 10: Interaction plot for Alloy 2

Conclusion:

- For **both alloys**, there is a **clear interaction** between dentist and method.
- The effect is **more prominent in Alloy 2**, suggesting higher variability in outcome depending on who applies the method.

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

Objective

To test whether the **combined effect** of **dentist** and **method**, including their **interaction**, has a statistically significant impact on the mean implant hardness (**Response**) for **Alloy 1** and **Alloy 2** separately.

Let μ_{dm} denote the **mean implant hardness** for the combination of **Dentist *d*** and **Method *m***.

We are interested in testing:

- The **interaction effect** between Dentist and Method

Statistical Hypotheses

We will test the **Null Hypothesis**:

H_0: All μ_{dm} values are equal (no interaction effect)

(The mean hardness is the same across all dentist-method combinations)

against the **Alternative Hypothesis**:

H_a: At least one μ_{dm} differs

(The mean hardness differs for at least one dentist-method combination)

Assumption Checks

The assumptions required for Two-Way ANOVA, **normality of residuals** and **homogeneity of variances** , were already evaluated in **Problems 4.1 and 4.2** for both **Alloy 1** and **Alloy 2** using the `Response` variable.

- **Normality (Shapiro-Wilk Test):**
  - **Alloy 1:** W = 0.830, p = 0.00001
  - **Alloy 2:** W = 0.888, p = 0.00040
    Since both p-values are less than 0.05, the assumption of normality is **violated** for both alloys.
- **Homogeneity of Variance (Levene's Test):**
  - **Alloy 1:** F = 1.385, p = 0.25655
  - **Alloy 2:** F = 1.446, p = 0.23687
    Both p-values are greater than 0.05, indicating that the assumption of equal variances is **satisfied** for both alloys.

While the normality assumption is violated, the Two-Way ANOVA is known to be **robust to moderate violations**, particularly with balanced group sizes. Therefore, we proceed with the ANOVA analysis.

Two way Anova for Alloy 1

Two way anova test returned the below

Output:

	sum_sq	df	F	PR(>F)
C(Dentist):C(Method)	185941.377778	8.0	3.398383	0.006793

The p-value (**0.0068**) for the interaction effect between dentist and method is well below the 0.05 significance level, leading us to **reject the null hypothesis**.

This indicates that the effect of method on implant hardness **varies depending on the dentist**.

Tukey HSD Test - Alloy 1

A Tukey HSD test revealed that several dentist-method combinations, particularly **Dentist 5 using Method 3**, differed significantly in mean hardness compared to others.

	group1	group2	meandiff	p-adj	lower	upper	reject
10	1_1	4_3	-302.6667	0.0070	-551.4950	-53.8383	True
13	1_1	5_3	-362.6667	0.0007	-611.4950	-113.8383	True
26	1_2	5_3	-278.6667	0.0173	-527.4950	-29.8383	True
38	1_3	5_3	-299.3333	0.0079	-548.1617	-50.5050	True
46	2_1	4_3	-280.6667	0.0160	-529.4950	-31.8383	True
49	2_1	5_3	-340.6667	0.0016	-589.4950	-91.8383	True
56	2_2	4_3	-269.3333	0.0243	-518.1617	-20.5050	True
59	2_2	5_3	-329.3333	0.0025	-578.1617	-80.5050	True
68	2_3	5_3	-304.6667	0.0065	-553.4950	-55.8383	True
76	3_1	5_3	-271.0000	0.0229	-519.8283	-22.1717	True
83	3_2	5_3	-286.6667	0.0128	-535.4950	-37.8383	True
89	3_3	5_3	-286.0000	0.0131	-534.8283	-37.1717	True
91	4_1	4_3	-269.3333	0.0243	-518.1617	-20.5050	True
94	4_1	5_3	-329.3333	0.0025	-578.1617	-80.5050	True
95	4_2	4_3	-285.0000	0.0137	-533.8283	-36.1717	True
98	4_2	5_3	-345.0000	0.0013	-593.8283	-96.1717	True
103	5_1	5_3	-270.3333	0.0234	-519.1617	-21.5050	True
104	5_2	5_3	-303.6667	0.0067	-552.4950	-54.8383	True

Table 14: 4.4 output for tukey HSD Test - alloy 1

Two way Anova for Alloy 2

Two way anova test returned the below

Output:

	sum_sq	df	F	PR(>F)
C(Dentist):C(Method)	197459.822222	8.0	1.922787	0.093234

The p-value (**0.0932**) for the interaction effect is above the 0.05 significance level, so we **fail to reject the null hypothesis**.

This means the effect of method on implant hardness is **consistent across dentists**.

Tukey HSD Test – Alloy 2

Despite the non-significant interaction, the Tukey HSD test identified several significant pairwise differences (*p* < 0.05), mostly involving **Dentist 5 using Method 3 (5_3)**.

This suggests that while the interaction effect is not statistically significant overall, **Dentist 5 using Method 3 may still be a performance concern** worth further investigation.

	group1	group2	meandiff	p-adj	lower	upper	reject
13	1_1	5_3	-427.0000	0.0049	-767.8958	-86.1042	True
26	1_2	5_3	-522.3333	0.0003	-863.2292	-181.4375	True
49	2_1	5_3	-464.6667	0.0017	-805.5625	-123.7708	True
59	2_2	5_3	-432.0000	0.0043	-772.8958	-91.1042	True
76	3_1	5_3	-363.6667	0.0279	-704.5625	-22.7708	True
83	3_2	5_3	-400.0000	0.0105	-740.8958	-59.1042	True
94	4_1	5_3	-360.6667	0.0302	-701.5625	-19.7708	True
98	4_2	5_3	-429.3333	0.0046	-770.2292	-88.4375	True
103	5_1	5_3	-479.0000	0.0011	-819.8958	-138.1042	True
104	5_2	5_3	-446.3333	0.0028	-787.2292	-105.4375	True

Table 15: 4.4 output for tukey HSD Test - alloy 2

Conclusion

Our two-way ANOVA and Tukey HSD analysis reveal two distinct operational patterns across alloys:

- **In Alloy 1**, the relationship between dentist and method significantly influences implant hardness (p = 0.0068). This indicates that performance is **not only method-driven but also highly dependent on the practitioner**. The significant differences identified via Tukey HSD, particularly involving **Dentist 5 using Method 3**, highlight opportunities to tailor method allocation based on individual dentist performance profiles.

- **In Alloy 2**, the interaction effect is not statistically significant (p = 0.0932), suggesting a more **uniform performance of methods across dentists**. However, the Tukey HSD test still flags **Dentist 5 with Method 3** as a recurring outlier — signaling that even in a stable system, **individual skill or procedural execution may impact outcomes**.

THANK YOU