ADVANCED STATISTICS – PROJECT BUSINESS REPORT

Table of Contents:

Table of Figures:

Problem-1

A physiotherapist with a male football team is interested in studying the relationship between foot injuries and positions at which the players play. The data collected is summarized in the table below, Answer the questions based on the table.

	Striker	Forward	Attacking Midfielder	Winger	**Total**
Players Injured	45	56	24	20	**145**
Players Not Injured	32	38	11	9	**90**
Total	**77**	**94**	**35**	**29**	**235**

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

From the given table we can see that the, number of players injured is 145, and the total number of players is 235. Hence the probability that a randomly chosen player is injured is;

P (Injured) = 145/235 = 0.617

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

Total number of players that are forwards or wingers is 123 (94 + 29), and total number of players under study is 235. Hence, the probability that a randomly chosen player is a forward or a winger is;

P (Forward OR Winger) = (94+29)/235 = 0.523

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

Note that this is a joint probability problem, where two conditions are being satisfied simultaneously. The number of players who have a foot injury and who play in striker position is 45. The total number of players under consideration is 235. Hence the probability that a player is a striker and is injured is;

P (Striker AND Injured) = 45/235 = 0.1914

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

Note that this is a conditional probability problem, because not all players and considered. Given that the chosen player is injured we need to find the probability that he is a striker. Since there are 145 injured players. The conditional probability is;

P (Striker|Injured) = 45/145 = 0.310

P (Human Error) = P (Radiation Leak AND Human Error)/ P (Radiation Leak | Human Error)
= 0.0012/0.1 = 0.012

Problem-2

The breaking strength of gunny bags used for packaging cement is normally distributed with a mean of 5 kg per sq. centimetre and standard deviation of 1.5 kg per sq. centimetre. 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 appropriate visual representation of your answers, without which marks will be deducted)

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

Using the mean and the standard deviation of the normal distribution, we need to find the probability of a gunny bag having a breaking strength less than 3.17 kg per sq cm.(cumulative probability calculated using stats.norm.cdf function).

P (Breaking strength < 3.17) = 0.1112

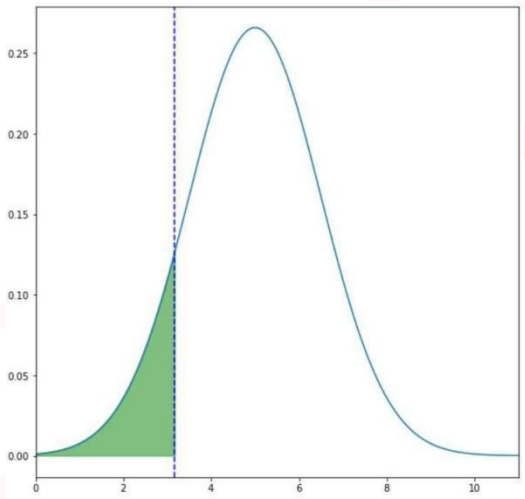

Figure 1 Green Area indicates required probability

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

Using the mean and the standard deviation of the normal distribution, we need to find the probability of a gunny bag having a breaking strength at least 3.6 kg per sq cm.

P (Breaking strength > 3.6) = 1 – 0.1753 = 0.8247

(Cumulative probability calculated using stats.norm.cdf function).

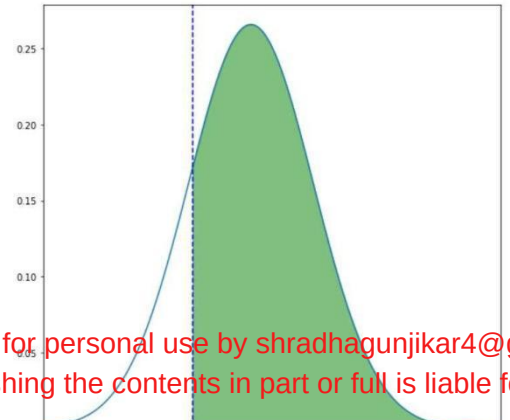

Figure 2 Required probability indicated by green shaded area.

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

Using the mean and the standard deviation of the normal distribution, we need to find the probability of a gunny bag having a breaking strength between 5kg per sq cm and 5.5 kg per sq cm.

P (5 < Breaking strength < 5.5) = 0.6306 – 0.5 = 0.1306

(Cumulative probability calculated using stats.norm.cdf function).

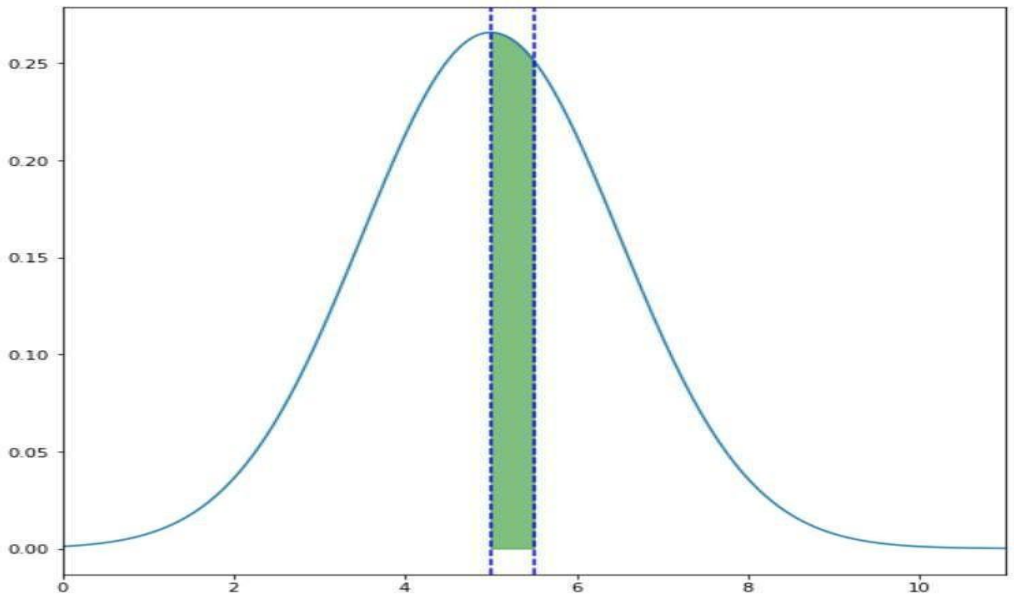

Figure 3 Required Probability indicated by the green shaded region.

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

Using the mean and the standard deviation of the normal distribution, we need to find the probability of a gunny bag having a breaking strength.

P ([Breaking strength < 3] or [Breaking strength > 7.5)

= P (Breaking strength < 3) + P (Breaking strength > 7.5)

= 0.0912 + (1 – 0.9522) = 0.0912 + 0.0478 = 0.1390

Another way of arriving at the same solution is to compute

1 – P (3 < Breaking strength < 7.5)

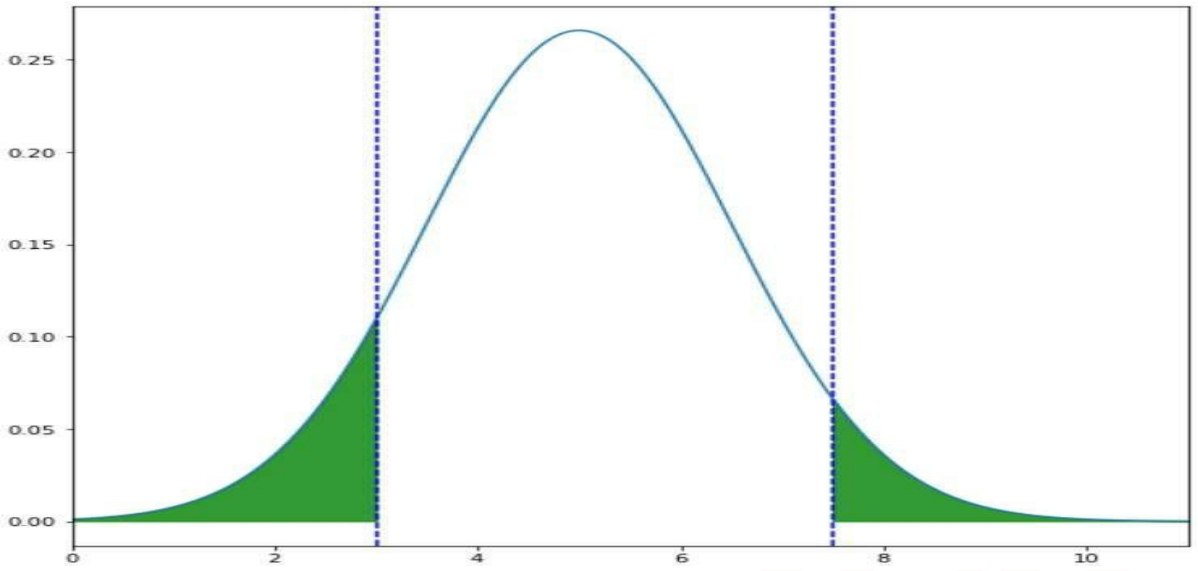

Figure 4 Green shaded area indicates the required probability

Problem-3

Zingaro stone printing is a company which pecializes in printing images or patterns on polished or unpolished stones. However, for 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 5% significance level);

3. 1 How does the hardness of implants vary depending on dentists?

The hypothesis test is formulated as below,

H_0 = The population mean BHI for unpolished stones is at least 150 or ($\mu_u \geq 150$)

H_1 = The population mean BHI for unpolished stones is at least 150 or ($\mu_u < 150$)

Looking at the Hypothesis formulation, it can be observed that this is a left tailed test with a Level of Significance 0.05 (α = 5%).

Output –

```
The critical value of t is -1.666 and the t-test statistic is -4.165

p value of the test is 0.004%
```

As observed above, since the numerical value of the test statistic is smaller than the critical value, the P-value is smaller than the level of significance. Hence, the Null Hypothesis that the unpolished stone surfaces have an average BHI of at least 150 is rejected.
Zingero is justified in concluding that the unpolished stones are not suitable for optimal level of printing.

3. 2 Is the mean hardness of the polished and unpolished stones the same?

H_0: The population mean BHI for Polished stones and Unpolished stones is equal

($\mu_p = \mu_u$) or ($\mu_p - \mu_u = 0$)

H_1: The population mean BHI for Polished stones and Unpolished stones is not equal

($\mu_p \neq \mu_u$) or ($\mu_p - \mu_u \neq 0$)

Note that this is a two-tailed test with a Level of Significance of 0.05 (α = 5%). Output -

```
The critical values of t are -1.993 and 1.993, and the t-test statistic is 3.242

p value of the test is 0.147%
```

Note that the test statistic is larger than the upper critical value and hence the P-value is smaller than the level of significance. The Null Hypothesis that the polished stone surfaces have a BHI equal to that of the unpolished stones will be rejected.
Average hardness of the polished and the unpolished stones is not equal.

Problem- 4

The hardness of metal implant in dental cavities depend on multiple factors, such as the method of implant, the temperature at which the metal is treated, the alloy used as well as on the dentists who may favour one method above another and may work better in his/her favourite method. Response is the variable of interest.

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

Solution-

Hypothesis test formulation for Alloy 1

H01: The mean implant hardness is the same across different dentists with type 1 alloy. HA1: Mean implant hardness is different for at least one pair of the dentists with type 1 alloy.

Hypothesis test formulation for Alloy 2

H02: The mean implant hardness is the same across different dentists with type 2 alloy. HA2: Mean implant hardness is different for at least one pair of the dentists with type 2 alloy.

Test of Normality (Shapiro) for Type 2 Alloy for Dentist variable

For one dentist, p-value is smaller than 0.05. We will proceed with ANOVA.

Test of Normality (Shapiro) for Type 1 Alloy for Method variable

In none of the cases, p-value is smaller than 0.05. Hence normality assumption is satisfied

Test of Normality (Shapiro) for Type 2 Alloy for Method variable

For one dentist, p-value is smaller than 0.05. We will proceed with ANOVA.

Levene test of equal variance for Type 1 and Type 2 Alloy for Method variable

Form above result for both alloy 1 and alloy 2 p value is less than 0.05 we can conclude that the variances are not equal

Levene test of equal variance for Type 1 and Type 2 Alloy for Dentist variable

Form above result for both alloy 1 and alloy 2 p value is greater than 0.05 we can conclude that the variances are equal.

Since the level of significance is 5%, we can check from the below ANOVA output arrived at by using the anova_lm function from scipy. stats.

Hypothesis test formulation for Alloy 1

H01: The mean implant hardness is the same across different dentists with type 1 alloy. HA1:

Mean implant hardness is different for at least one pair of the dentists with type 1 alloy.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	106683.688889	26670.922222	1.977112	0.116567
Residual	40.0	539593.555556	13489.838889	NaN	NaN

Since p-value is greater than 0.05, we fail to reject the null hypothesis of equality.

Hypothesis test formulation for Alloy 2

H02: The mean implant hardness is the same across different dentists with type 2 alloy. HA2:

Mean implant hardness is different for at least one pair of the dentists with type 2 alloy.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dentist)	4.0	5.679791e+04	14199.477778	0.524835	0.718031
Residual	40.0	1.082205e+06	27055.122222	NaN	NaN

Since p-value is greater than 0.05, we fail to reject the null hypothesis of equality.

As Null Hypothesis is not rejected, so no need to do post hoc test Tukey HSD

4. 2 How does the hardness of implants vary depending on methods?

Solution-

Since the level of significance is 5%, we can check from the below ANOVA output arrived at by using the anova_lm function from scipy. stats.

Hypothesis test formulation for Alloy 1

H01: The mean implant hardness is the same across different methods with type 1 alloy.

HA1: Mean implant hardness is different for at least one pair of the methods with type 1 alloy.

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	148472.177778	74236.088889	6.263327	0.004163
Residual	42.0	497805.066667	11852.501587	NaN	NaN

Since p-value is smaller than 0.05, null hypothesis is rejected. At least one method is different from the rest.

Hypothesis test formulation for Alloy 2

H01: The mean implant hardness is the same across different methods with type 2 alloy.

HA1: Mean implant hardness is different for at least one pair of the methods with type 2 alloy.

	df	sum_sq	mean_sq	F	PR(>F)
C(Method)	2.0	499640.4	249820.200000	16.4108	0.000005
Residual	42.0	639362.4	15222.914286	NaN	NaN

Since p-value is smaller than 0.05, null hypothesis is rejected. At least one method is different from the rest.

Post hoc test Tukey HSD

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-6.1333	0.9	-102.7105	90.4438	False
1	3	-124.8	0.0085	-221.3771	-28.2229	True
2	3	-118.6667	0.0128	-215.2438	-22.0895	True

Null rejected here, using Tukey HSD we get that Group 3 is different from Groups 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?

Solution –

Since the level of significance is 5%, we can check from the below ANOVA output arrived at by using the anova_lm function from scipy. stats.

Hypothesis test formulation for Alloy 1

H01: The mean implant hardness is the same across different temperature with type 1 alloy. HA1: Mean implant hardness is different for at least one pair of the temperature with type 1 alloy.

	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2.0	10154.444444	5077.222222	0.335224	0.717074
Residual	42.0	636122.800000	15145.780952	NaN	NaN

As we can see from the ANOVA output the P-value is more than the level of significance, we are Failed to Reject the Null Hypothesis

Hypothesis test formulation for Alloy 2

H01: The mean implant hardness is the same across different temperature with type 2 alloy. HA1: Mean implant hardness is different for at least one pair of the temperature with type 2 alloy.

	df	sum_sq	mean_sq	F	PR(>F)
C(Temp)	2.0	9.374893e+04	46874.466667	1.883492	0.164678
Residual	42.0	1.045254e+06	24886.996825	NaN	NaN

As we can see from the ANOVA output the P-value is more than the level of significance, we are Failed to Reject the Null Hypothesis

As Null Hypothesis is not rejected, so no need to do post hoc test Tukey HSD

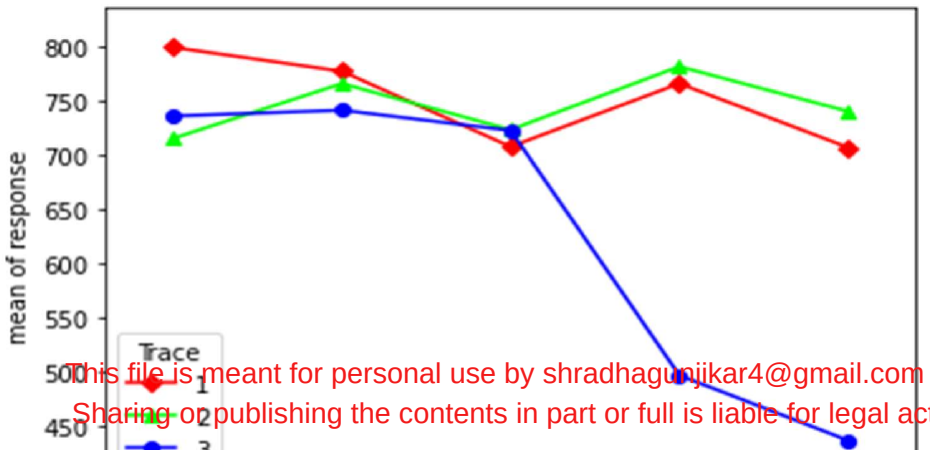

Figure 5 Interaction Plot Alloy 1

We can clearly say that mean hardness of dental implant is not same when different Dentists are using different Methods in Alloy 1. This interaction is very significant here.

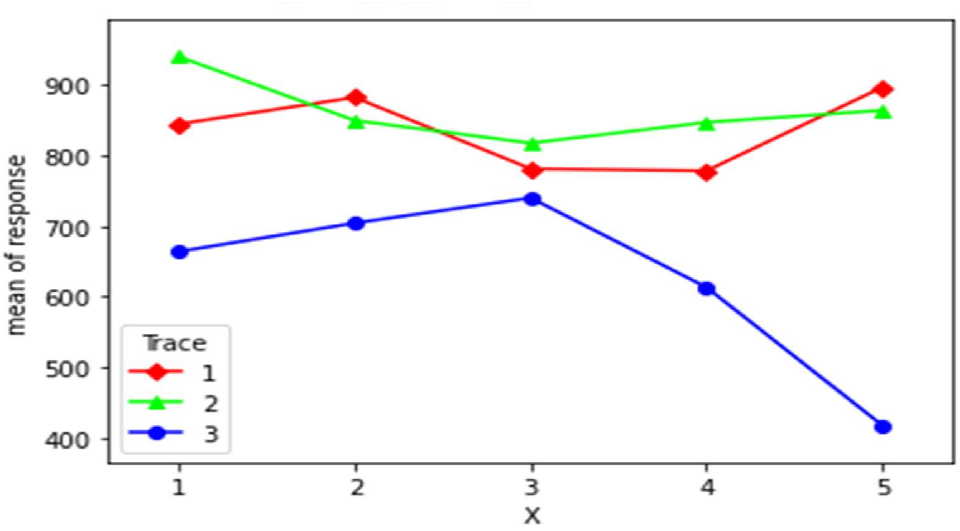

Figure 6 Interaction Plot Alloy 2

We can clearly say that mean hardness of dental implant is not same when different Dentists are using different Methods in Alloy 2. This interaction is very significant here.

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

Solution –

Since the level of significance is 5%, The two-way ANOVA can be tested using the anova_LM function resulting in the below output.

Hypothesis test formulation for Alloy 1

H0: The mean implant hardness is same across different Dentists with type 1 Alloy. HA: Mean implant hardness is different for at least one of the Dentist with type 1 Alloy.

H0: The mean implant hardness is same across different Methods with type 1 Alloy.

HA: Mean implant hardness is different for at least one of the Method type with type 1 Alloy.

H0: There is no interaction between Dentist and Method types with type 1 Alloy. HA: There is interaction between Dentist and Method types with type 1 Alloy

	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

Interaction for Alloy 1

	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

For alloy 1: Dentist, Method and their interaction all are significant

Hypothesis test formulation for Alloy 2

H0: The mean implant hardness is same across different Dentists with type 2 Alloy. HA: Mean implant hardness is different for at least one of the Dentist with type 2 Alloy.

H0: The mean implant hardness is same across different Methods with type 2 Alloy.

HA: Mean implant hardness is different for at least one of the Method type with type 2 Alloy.

H0: There is no interaction between Dentist and Method types with type 1 Alloy. HA: There is interaction between Dentist and Method types with type 1 Alloy

	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

Interaction for Alloy 2

For Alloy 2: Method and their interaction are significant for Implant hardness