Murali B

ADVANCED STATISTICS PROJECT

TABLE OF CONTENTS:

**PROBLEM 1 PAGE**

1.1Probability that a randomly chosen player would suffer an injury 3  
1.2 Probability that a player is a forward or a winger? 3  
1.3 Probability that a randomly chosen player plays in a striker position and has a foot injury 4  
1.4 Probability that a randomly chosen injured player is a striker? 4  
1.5 Probability that a randomly chosen injured player is either a forward or an attacking midfielder? 4

**PROBLEM 2:**

2.1Probabilities of a fire, a mechanical failure, and a human error respectively? 5-6  
2.2 Probability of a radiation leak? 6  
2.3 Probability that it has been caused by fire, mechanical failure, and human error 6

**PROBLEM 3:**

3.1Proportion of the gunny bags have a breaking strength less than 3.17 kg per sq cm 6-7  
3.2 Proportion of the gunny bags have a breaking strength at least 3.6 kg per sq cm 7-8  
3.3 Proportion of the gunny bags have a breaking strength between 5 and 5.5 kg per 8-9  
3.4 Proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 9-10

**PROBLEM 4:**

4.1Probability that a randomly chosen student gets a grade below 85 on this exam 10  
4.2 Probability that a randomly selected student scores between 65 and 87 11  
4.3 Passing cut-off so that 75% of the students clear the exam 11

**PROBLEM 5:**

5.1Do you think Zingaro is justified in thinking so? 11-13  
5.2 Is the mean hardness of the polished and unpolished stones the same? 13-14

**PROBLEM 6:**

**6.1** Using the sample data provided conclude whether the program is successful? 14-16

**PROBLEM 7:**

7.1Test whether there is any difference among the dentists on the implant hardness 17-18  
7.2 Are the assumptions fulfilled? Comment separately on both alloy types.?  18  
7.3 Clearly state your conclusion. If the null hypothesis is rejected, is it possible to identify 19  
7.4 Now test whether there is any difference among the methods on the hardness 19-20  
7.5 Now test whether there is any difference among the temperature levels 20-21  
7.6 Consider the interaction effect of dentist and method and comment on the interaction 21-22  
7.7 Now consider the effect of both factors, dentist, and method, separately on each alloy. 22-24

**CONTENTS OF FIGURE**:

**PROBLEM 1:** **PAGE**

1.1 Output 3  
1.2 Output 3  
1.3 Output 4  
1.4 Output 4  
1.5 Output 4

**PROBLEM 2:**

2.1 Output 5  
2.2 Output 6  
2.3 Output 6

**PROBLEM 3:**

3.1 Output 7  
3.1.1 Output 7  
3.2.1 Output 7  
3.3 Output 8  
3.3.1 Output 9  
3.4 Output 10

**PROBLEM 4:**

4.1 Output 10  
4.2 Output 11  
4.3 Output 11

**PROBLEM 5:**

5.1.1 Output 12  
5.1.2 Output 13  
5.1.3 Output 13  
5.2 Output 13  
5.3 **EXCEL OUTPUT** 14

**PROBLEM 6:**

6.2 Output 15  
6.3 Output 16  
6.4 Output 16  
6.5 **EXCEL OUTPUT**  16

**PROBLEM 7:**

7.2 Output 18  
7.2.1 Output 18  
7.3 Output 19  
7.3.1 Output 19  
7.4 Output 20  
7.4.1 Output 20  
7.5 Output 20  
7.5.1 Output 21  
7.6 Output 21  
7.7 Output 22  
7.7.1 Output 23  
7.7.2 Output 23  
7.7.3 Output 23

**CONTENTS OF TABLE: PAGE**Table 5.1 – Problem 5 Top 5 records of the dataset 12  
Table 6.1 - Problem 5 Top 5 records of the dataset 15   
Table 7.1 – Problem 7.1 Dataset details 17  
Table 7.1.1 – Problem 7.1 Dataset datatype details 17

**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** |

* 1. What is the probability that a randomly chosen player would suffer an injury?

**Output**: 

Fig 1.1

Here we can see total number of players injured are 14 but total number of players are 235 so we divide 145 from 235 to get the result.

* 1. What is the probability that a player is a forward or a winger?

**Output**:



Fig1.2

Here the rule is OR so we will add the attributes and we can see total numbers of forward players is 94 and winger is 29 so we add up these and divide it from 235 to get the result.

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

**Output**:



Fig 1.3

Here we can see total number of strikers injured is 45 and total number of total players injured are 145 hence we divide 45 by 145 to get the result.

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

**Output**:



Fig 1.4

So, we see injured striker players as 45 and total injured striker players are 77 hence, we divide 45 by 77 to get the result.

* 1. What is the probability that a randomly chosen injured player is either a forward or an attacking midfielder?

**Output**:



Fig 1.5

We see here forward players injured are 56 and injured attacking midfielder players are 24 so the total of number of Forward players are 94 and total number of attacking midfielder players are 35 hence we add the totals and divide it up to get the result.

**Problem 2**

An independent research organization is trying to estimate the probability that an accident at a nuclear power plant will result in radiation leakage. The types of accidents possible at the plant are, fire hazards, mechanical failure, or human error. The research organization also knows that two or more types of accidents cannot occur simultaneously.

According to the studies carried out by the organization, the probability of a radiation leak in case of a fire is 20%, the probability of a radiation leak in case of a mechanical 50%, and the probability of a radiation leak in case of a human error is 10%. The studies also showed the following.

* The probability of a radiation leak occurring simultaneously with a fire is 0.1%.
* The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%.
* The probability of a radiation leak occurring simultaneously with a human error is 0.12%.

Based on the information available, answer the questions below:

2.1 What are the probabilities of a fire, a mechanical failure, and a human error respectively?

**Output**:







Fig 2.1

For probabilities of fire: we multiply the probability of radiation leak of fire with that of probability of a radiation leak occurring simultaneously with a fire and again divide by probability of radiation leak of fire to get the result.

For probabilities of mechanical failure: we multiply the probability of radiation leak of mechanical failure with that of probability of a radiation leak occurring simultaneously with a mechanical failure and again divide by probability of radiation leak of mechanical failure to get the result.

For probabilities of human error: we multiply the probability of radiation leak of human error with that of probability of a radiation leak occurring simultaneously with a human error and again divide by probability of radiation leak of human error to get the result.

2.2 What is the probability of a radiation leak?

**Output:**



Fig 2.2

So, we multiply up all the probabilities of radiation leak along withprobability of a radiation leak occurring simultaneously by adding up with all the 3 components such as fire, mechanical failure and human error.

2.3 Suppose there has been a radiation leak in the reactor for which the definite cause is not known. What is the probability that it has been caused by:

* A Fire.
* A Mechanical Failure.
* A Human Error.

**Answer: 0** because in the problem statement it states that no 2 or more accidents will never occur together, so that probability will be 0 because we can't have all the 3 of them reacting at the same time.

**Problem 3:**

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.

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

**Output:**



Fig 3.1

So here we calculate the Z score using mu value as 5 and sigma value as 1.5 and P value and substitute in the formula to calculate normal distribution to get the result.

**Visual representation:**

Chart, histogram

Description automatically generated

Fig 3.1.1

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

**Output:**



Fig 3.2

Again, we calculate the Z score with the mu as 5 and sigma value as 1.5 and then we calculate P value using normal distribution formula by subtracting it with 1 because the data provided to calculate is at least 3.6 to get the result.

**Visualization presentation:**

Chart, histogram

Description automatically generated

Fig 3.2.1

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

**Output:**



Fig 3.3

So, we can see the data is between 5 and 5.5 kg so we calculate 2 separate Z scores for each value and then minus with each other with the help of normal distribution formula to get the output.

**Visual representation:**

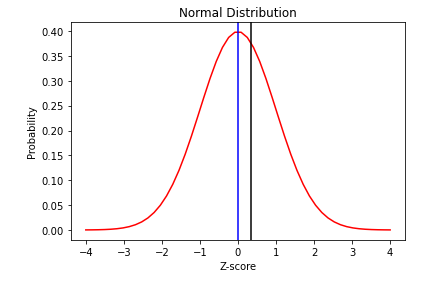


Fig 3.3.1

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

**Output:**



Fig 3.4

Again, we can see the data is not between 3 and 7.5 kg so we calculate 2 separate Z scores for each value and then minus with each other with the help of normal distribution formula and at last again the result will be subtracted to get the output.

**Visual representation:**

Chart, histogram

Description automatically generated

Fig 3.4.1

**Problem 4:**

Grades of the final examination in a training course are found to be normally distributed, with a mean of 77 and a standard deviation of 8.5. Based on the given information answer the questions below.

4.1 What is the probability that a randomly chosen student gets a grade below 85 on this exam?

**Output:**



Fig 4.1

With the help of normal distribution formula using mu as 77 and sigma as 8.5 with the data provided, we need to calculate the CDF for 85 we can obtain the desired result.

4.2 What is the probability that a randomly selected student score between 65 and 87?

**Output:**



Fig 4.2

Here we need to calculate the CDF for both 65 and 87 so we find z score of both the values with the help of mu and sigma values and again we subtract both the values using normal distribution formula to get the output.

4.3 What should be the passing cut-off so that 75% of the students clear the exam?

**Output:**

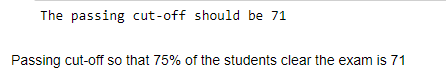


Fig 4.3

Here we need to calculate the PPF for 25% as top 75% will be starting from 25th percentile to get the desired answer.

**Problem 5:**

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);

5.1 Earlier experience of Zingaro with this client is favorable as the stone surface was found to be of adequate hardness. However, Zingaro has reason to believe now that the unpolished stones may not be suitable for printing. Do you think Zingaro is justified in thinking so?

Table

Description automatically generated

Table 5.1

Top 5 records of the dataset.



Fig 5.1.1

2 columns: 'Unpolished ', 'Treated and Polished' so Unpolished has a space so we must be careful in dealing with it.

**Step 1:** Defining the Null Hypothesis and Alternate Hypothesis

We can assume Null Hypothesis as hardness index is equal to 150 and Alternate Hypothesis as hardness index which is not equal to 150 as per the statement provided in the problem.

**Step 2:** Deciding the significance level

Here we can see level of significance as 5% given in the problem statement.

**Step 3:** Identify the test statistic

Here we can see that we have 2 samples and standard deviation is not known and sample sizes are same for both the samples.

**Step 4:** Calculate the p value and test statistic

**We basically first use the scipy.stats.ttest\_ind to calculate the t-test for the means of both the independent samples from the given values. And it will return t statistic and two-tailed p value.**

We first calculate xbar1 , xbar2 , variance1 , variance2 , number of values, degree of freedom value so using all these values we get numerator and denominator values then with the help of two sample t-test formula we can derive both p value and test statistic values.

**Output:**



Fig 5.1.2

We can see t-stat values are -3.242 and p value as 0.0014

**Step 5:** Deciding whether to reject or accept the null hypothesis

Text, Word

Description automatically generated

Fig 5.1.3

Here we see p-value is less than the significance level hence we reject null hypothesis. With the above result we can prove the problem statement things to justify it.

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

**Output:**

Text

Description automatically generated

Fig 5.2

So, we can observe that the mean of Unpolishes stones are 134.11 and treated and Polished stones are 147 hence with the above date we can say that both the means are not same.

Scatter chart

Description automatically generated

Fig 5.3

Problem 5 solution worked in excel output

**Problem 6:**

Aquarius health club, one of the largest and most popular cross-fit gyms in the country has been advertising a rigorous program for body conditioning. The program is considered successful if the candidate can do more than 5 push-ups, as compared to when he/she enrolled in the program. Using the sample data provided can you conclude whether the program is successful? (Consider the level of Significance as 5%)

Note that this is a problem of the paired-t-test. Since the claim is that the training will make a difference of more than 5, the null and alternative hypotheses must be formed accordingly.

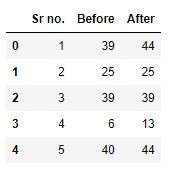


Table 6.1

Top 5 records of the dataset



Fig 6.2

3 columns are there: Sr no, Before, after but we will work only on Before and After columns values.

**Step 1:** Defining the Null Hypothesis and Alternate Hypothesis

We can assume Null Hypothesis as push-ups is greater than equal to 5 and Alternate Hypothesis push-ups is not equal to 5 as per the statement provided in the problem.

**Step 2:** Deciding the significance level

Here we can see level of significance given again as 5% given in the problem statement.

**Step 3:** Identify the test statistic

We can observe that both the sample sizes are same and standard deviation is not known.

**Step 4:** Calculate the p value and test statistic

Using the values of dbar, standard deviation, number of values, degree of freedom then we calculate the values with the help of two paired t-test formula.

**Again, we use the SciPy. stats. ttest\_rel to calculate the t-test of both the samples values. And will finally return the t statistic value and two-tailed p value.**

**Output:** 

Fig 6.3

Here we can see t-statistic value as 19.323 and p-value value as 1.146

**Step 5:** Deciding whether to reject or accept the null hypothesis

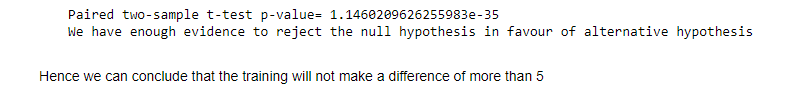


Fig 6.4

Here we see p-value is less than the significance level hence we reject null hypothesis. Hence, we can conclude that the training will not make a difference of more than 5.

Problem 6 output solved using Excel

Graphical user interface, application, table, Excel

Description automatically generated

Fig 6.5

**Problem 7:**

Dental implant data: The hardness of metal implant 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 on 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.

1. Test whether there is any difference among the dentists on the implant hardness. State the null and alternative hypotheses. Note that both types of alloys cannot be considered together. You must state the null and alternative hypotheses separately for the two types of alloys.?

**Output:**

Table

Description automatically generated

Table 7.1

Text

Description automatically generated

Table 7.1.1

The Hypothesis for alloy 1 is:

Null Hypothesis: The hardness of metal implant in dental cavities depends on factors like method , temperature and the alloy type is same.

Alternate Hypothesis: At least one method of treating the metal based on factors like method, temperature and the alloy type is different.

The Hypothesis for alloy 2 is:

Null Hypothesis: The alloy used on the dentists who may favor one method above another and may work better in their favorite method.

Alternate Hypothesis: At least one method may not favor above another and will not work better in their favorite method.

7.2 Before the hypotheses may be tested, state the required assumptions. Are the assumptions fulfilled? Comment separately on both alloy types.?

**Output:**

Table

Description automatically generated

Fig 7.2

Table

Description automatically generated

Fig 7.2.1

Required assumptions: Whether the hardness of metal implant depends on Temp, alloy and Dentist or not.

As we can see P value of Temp, Alloy and Dentist shows as less than 0.05 hence these values have a significant effect, and the assumptions is not fulfilled.

7.3 Irrespective of your conclusion in 2, we will continue with the testing procedure. What do you conclude regarding whether implant hardness depends on dentists? Clearly state your conclusion. If the null hypothesis is rejected, is it possible to identify which pairs of dentists differ?

**Output:**

Text

Description automatically generated

Fig 7.3

Table

Description automatically generated

Fig 7.3.1

We can conclude that implant hardness does not depends on dentists when combined with method and alloy or not as P value is less hence there is no significant cause on dentists with respect to implant hardness. So, hypothesis is rejected.

The pair of dentists who differ are dentists who use method of implant combined with alloy type and temperature as they differ from dentists who apply their own method.

7.4 Now test whether there is any difference among the methods on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which pairs of methods differ?

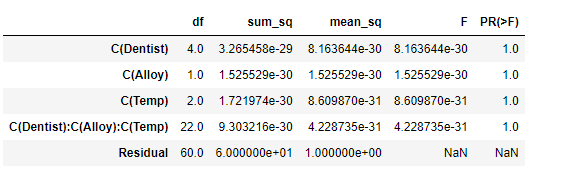
**Output:** 

Fig 7.4

Table

Description automatically generated

Fig 7.4.1

We can conclude that P value is less than 0.05 hence Null hypothesis is rejected and method which is being used by dentist along with alloy type and temperature differs from methods which is being applied as their favorite method.

7.5 Now test whether there is any difference among the temperature levels on the hardness of dental implant, separately for the two types of alloys. What are your conclusions? If the null hypothesis is rejected, is it possible to identify which levels of temperatures differ?

**Output:**

Text, letter

Description automatically generated

Fig 7.5

Text

Description automatically generated

Fig 7.5.1

No. There is no significant difference among the temperature levels on the hardness of dental implant. So, the null hypothesis is rejected.

We can conclude that the levels of temperature do not differ.

7.6 Consider the interaction effect of dentist and method and comment on the interaction plot, separately for the two types of alloys?

**Output:**

Chart

Description automatically generated

Fig 7.6

Chart, line chart

Description automatically generated

Fig 7.6.1

From the plot, we can see that only Alloy 2 has an interaction between Dentist and Method and Alloy 1 has no interaction between dentist and Method.

7.7 Now consider the effect of both factors, dentist, and method, separately on each alloy. What do you conclude? Is it possible to identify which dentists are different, which methods are different, and which interaction levels are different?

Output: Chart, rectangle, box and whisker chart

Description automatically generated

Fig 7.7

Chart, line chart, box and whisker chart

Description automatically generated

Fig 7.7.1

Chart, rectangle

Description automatically generated

Fig 7.7.2

Chart, line chart, box and whisker chart

Description automatically generated

Fig 7.7.3

We can conclude that dentist 1, 2, 3, 4 are different and method 1 and 2 are different and interaction levels between dentist and method on different alloy is different. Only dentist 5 and method 3 has an interaction.