

Titan Case Study

Group Assignment

* Swetha Jegannathan
* Priyank Mehar
* Naveen Kumar Sambangi
* Umapathi Raja Pedaprolu

**Table of Contents**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | [Project Objective ..............................................................................](#page3). | | | 3 |
| 2 | [Assumptions ....................................................................................](#page4). | | | 4 |
| 3 | [Step by step approach .......................................................................](#page4). | | | 4 |
|  | [3.](#page4)1 | [Exploratory Data Analysis .............................................................](#page4). | | 4 |
|  | [3.1.1 Number of Rows and Columns:.................................................](#page4). | | | 4 |
|  | [3.1.2 Features and their Types: ........................................................](#page4). | | | 4 |
|  | [3.1.3 Check for Missing Values: ........................................................](#page4). | | | 4 |
|  | [3.1.](#page4)4 | | [Dataset Summary: .................................................................](#page4). | 4 |
|  | [3.](#page5)2 | [Descriptive Statistics ....................................................................](#page5). | | 5 |
|  | [3.2.1 Measures of Central Tendency:.................................................](#page5). | | | 5 |
|  | [3.2.](#page6)2 | | [Measures of Dispersion: ..........................................................](#page6). | 6 |
|  | [3.](#page7)3 | [Data Visualization – Histogram and Boxplot: ....................................](#page7). | | 7 |
|  | [3.3.](#page7)1 | | [Key Observations: ..................................................................](#page7). | 7 |
|  | [3.](#page7)4 | [Hypothesis formation ...................................................................](#page7). | | 7 |
|  | [3.](#page8)5 | [Hypothesis Testing Method............................................................](#page8). | | 8 |
|  | [3.](#page9)6 | [Confidence Interval ......................................................................](#page9). | | 9 |
|  | [3.](#page10)7 | [Need of Larger Sample Size: .......................................................](#page10). | | [1](#page10)0 |
| 4 | [Conclusion and Recommendation ......................................................](#page12). | | | [1](#page12)2 |
| 5 | [Appendix – Source Code ..................................................................](#page13). | | | [1](#page13)3 |
|  | [5.](#page13)1 | [Exploratory Data Analysis ...........................................................](#page13). | | [1](#page13)3 |
|  | [5.](#page14)2 | [Descriptive Statistics ..................................................................](#page14). | | [1](#page14)4 |

**1 Project Objective**

The Titan Insurance Company has just installed a new incentive payment scheme for its lift policy sales force. It wants to have an early view of the success or failure of the new scheme. Indications are that the sales force is selling more policies but sales always vary in an unpredictable pattern from month to month and it is not clear that the scheme has made a significant difference.

Life Insurance companies typically measure the monthly output of a salesperson as the total sum assured for the policies sold by that person during the month. For example, suppose salesperson X has, in the month, sold seven policies for which the sums assured are £1000, £2500, £3000, £5000, £10000, £35000. X's output for the month is the total of these sums assured, £61,500. Titan's new scheme is that the sales force receive low regular salaries but are paid large bonuses related to their output (i.e. to the total sum assured of policies sold by them). The scheme is expensive for the company but they are looking for sales increases which more than compensate. The agreement with the sales force is that if the scheme does not at least break even for the company, it will be abandoned after six months.

The scheme has now been in operation for four months. It has settled down after fluctuations in the first two months due to the changeover. To test the effectiveness of the scheme, Titan have taken a random sample of 30 salespeople measured their output in the penultimate month prior to changeover and then measured it in the fourth month after the changeover (they have deliberately chosen months not too close to the changeover). The outputs of the salespeople are shown in Table 1

**Questions:**

a. Describe the five per cent significance test you would apply to these data to determine whether new scheme has significantly raised outputs?

b. What conclusion does the test lead to?

c. What reservations have you about this result?

d. Suppose it has been calculated that in order for Titan to break even, the average output must increase by £5000. If this figure is alternative hypothesis, what is:

(i) The probability of a type 1 error?

(ii) The probability of a type 2 error?

(iii) The power of the test?

e. What sample size would make the probabilities of type 1 and type 2 errors equal?

|  |  |  |
| --- | --- | --- |
| Sales\_Force | Output (£000) | |
| **Old Scheme** | **New Scheme** |
| 1 | 57 | 62 |
| 2 | 103 | 122 |
| 3 | 59 | 54 |
| 4 | 75 | 82 |
| 5 | 84 | 84 |
| 6 | 73 | 86 |
| 7 | 35 | 32 |
| 8 | 110 | 104 |
| 9 | 44 | 38 |
| 10 | 82 | 107 |
| 11 | 67 | 84 |
| 12 | 64 | 85 |
| 13 | 78 | 99 |
| 14 | 53 | 39 |
| 15 | 41 | 34 |
| 16 | 39 | 58 |
| 17 | 80 | 73 |
| 18 | 87 | 53 |
| 19 | 73 | 66 |
| 20 | 65 | 78 |
| 21 | 28 | 41 |
| 22 | 62 | 71 |
| 23 | 49 | 38 |
| 24 | 84 | 95 |
| 25 | 63 | 81 |
| 26 | 77 | 58 |
| 27 | 67 | 75 |
| 28 | 101 | 94 |
| 29 | 91 | 100 |
| 30 | 50 | 68 |

**2 Assumptions**

The sample size of the data set is 30 (≥ 30) . Titan have taken a random sample of 30 salespeople measured their output in the penultimate month prior to changeover and then measured it in the fourth month after the changeover (they have deliberately chosen months not too close to the changeover).Central Limit Theorem states that irrespective of the shape of the original population, the sampling distribution of the mean will approach a normal distribution as the size of the sample increases and becomes large

We also assume that the sample estimate will be reflective of the reality.

**3 Step by step approach**

We shall follow step by step approach to arrive to the conclusion as follows:

* 1. Exploratory Data Analysis
  2. Descriptive Statistics
  3. Data Visualization
  4. Hypothesis formation
  5. Selection of appropriate Hypothesis Testing method
  6. 95% Confidence Intervals
  7. Need of Larger Sample Size
  8. Conclusion and Recommendation.

1. **Exploratory Data Analysis**

Please refer Appendix for related code.

**3.1.1 Number of Rows and Columns:**

• The number of rows in the dataset is 30

• The number of columns (Features) in the dataset is 3

**3.1.2 Features and their Types:**

• Both the features, i.e. Old Scheme and New Scheme are continuous variables.

|  |  |  |
| --- | --- | --- |
| **Feature Code** | **Type** | **Continuous/** |
|  |  | **Categorical** |
| **Sales Force** | Integer | Continuous |
| **New Scheme** | Integer | Continuous |
| **Old Scheme** | Integer | Continuous |

**3.1.3 Check for Missing Values:**

• The data was checked for Missing Values using R function columns(is.na), and found **no missing values**.

**3.1.4 Dataset Summary** (Values in £000)

• ***Old Scheme:*** The minimum value is 28 and the maximum value is110. The average value is 68.03

• ***New Scheme:*** The minimum value is 32 and the maximum value is 122. The average value is 72

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Code | Minimum | Maximum | Average |
| Old Scheme | 28 | 110 | 68.03 |
| New Scheme | 32 | 122 | 72 |

**3.2 Descriptive Statistics**

Please refer Appendix for related code.

In this step, the features are explored in detail. The goal is to describe or summarize data in ways that are meaningful and useful for insights generation. It provides simple summaries about the sample and the measures. Together with simple graphics analysis, it forms the basis of virtually every quantitative analysis of data.

Given both the features – ‘Old Scheme’ and ‘New Scheme’ are continuous in nature; the following measures are relevant to understand the central tendency and spread of the variable.

|  |  |  |
| --- | --- | --- |
| Measures of Central Tendency | Measures of Dispersion | Visualization Method |
| Mean | Range | Histogram |
| Median | 1st Quartile | Box Plot |
| Mode | 3rd Quartile |  |
| Minimum | Interquartile Range |  |
| maximum | Variance |  |
|  | Standard Deviation |  |

**3.2.1 Measures of Central Tendency:**

|  |  |  |
| --- | --- | --- |
| Measures of Central Tendency | Old Scheme | New Scheme |
| Mean | 68.03 | 72.03 |
| Mode | 84 | 84 |
| Median | 74 | 74 |
| Maximum | 110 | 122 |
| Minimum | 28 | 32 |

page5image7880544

**3.2.1 Measures of Dispersion:**

|  |  |  |
| --- | --- | --- |
| Measures of Dispersion | Old Scheme | New Scheme |
| Range | 82 | 90 |
| 1st Quartile | 54 | 55 |
| 3rd Quartile | 81.5 | 85.75 |
| Inter Quartile Range(IQR) | 27.5 | 30.75 |
| Variance | 418.44 | 578.99 |
| Standard Deviation | 20.45 | 24.06 |

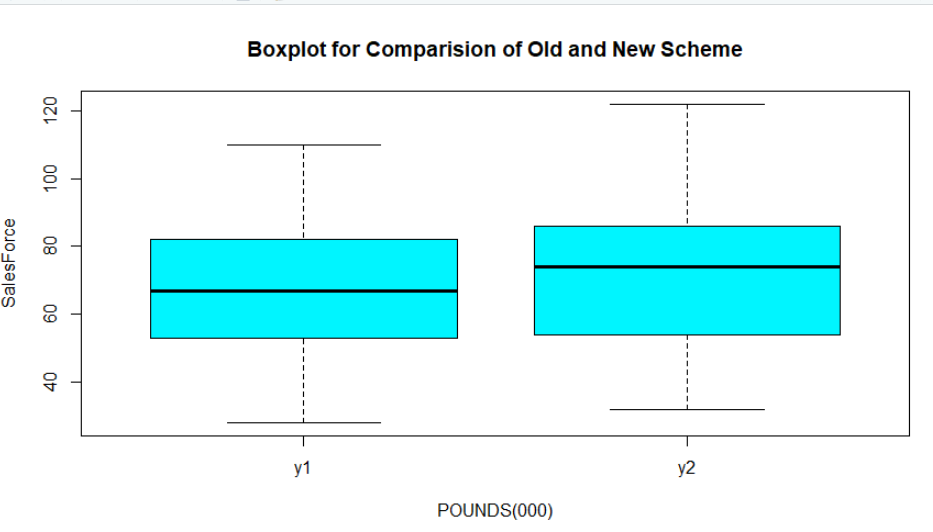
The data for ‘Range’, ‘1st Quartile’ & ‘3rd Quartile’ have been obtained using the R function ‘summary’ (The output is provided in the Appendix Measures of Central Tendency naming convention to be changed ).

The data for ‘Inter Quartile Range (IQR)’ has been computed and is the difference in value between 3rd Quartile (75 percentile) & 1st Quartile (25 percentile).

The variances have been obtained using R function ‘var’:  
The standard deviations have been obtained using R function ‘sd’:

**3.3 Data Visualization – Histogram and Boxplot:**





**3.3.1 Key Observations:**

* The Mean Output of New Scheme(72) is higher than the Mean Output of Old Scheme(68).
* No Outliers were observed in the data range for Old Scheme & New Scheme

**3.4 Hypothesis formation**

For the hypothesis formulation, we have to define the Null Hypothesis & Alternative Hypothesis.

• ***Null Hypothesis***o It is a hypothesis that says there is no statistical significance between the two variables. The null hypothesis is formulated such that the rejection of the null hypothesis proves the alternative hypothesis is true

• ***Alternative Hypothesis***

o It is one that states there is a statistically significant relationship between two variables. The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis

In Titan case, the management wants to have an early view of the success or failure of the new scheme. The scheme is expensive for the company but they are looking for sales increases which more than compensate. The agreement with the sales force is that if the scheme does not at least break even for the company, it will be abandoned after six months. a random sample of 30 salespeople measured their output in the penultimate month prior to changeover and then measured it in the fourth month after the changeover (they have deliberately chosen months not too close to the changeover).The Mean Output of New Scheme(72) is higher than the Mean Output of Old Scheme(68).

***Null Hypothesis (H0): μ1 - μ2 = 0 (i.e. they are the same)***

***Alternative Hypothesis (Ha): μ1 - μ2 ≠ 0 (i.e. they are not the same)***

Where,

* ***μ1:*** Sales Force Output in Old Scheme
* ***μ2:*** Sales Force Output in New Scheme  
  By formulation of above hypotheses, we assume that the old and new outputs of Salesforce show no significant difference to each other.

**3.5 Hypothesis Testing Method**

Based on the details shared for the Titan project, we can assume the following:

* Two Schemes
* No other influences considered

It seems to be a dependent sample case and the Paired T test will be applicable for the project.

Let us calculate the p-value using the R function ‘t.test’.

|  |
| --- |
| **t.test(New\_Scheme,Old\_Scheme,mu=0,paired = TRUE,var.equal = TRUE)**  Paired t-test  data: New\_Scheme and Old\_Scheme  t = 1.5559, df = 29, p-value = 0.1306  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -1.257949 9.257949  sample estimates:  mean of the differences  4 |

The p-value for the two-tailed test is 0.13, which is greater than level of significance α (0.05). Therefore, the Null Hypothesis (H0) will not be rejected and needs to be retained.

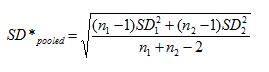
**95% confidence interval for the difference between the means of the two population:**

**Inference:** The 95% confidence interval of difference in population mean between both the schemes is between **–1.25 on the lower end 9.25 on the upper end**. This implies that, with 95% confidence, we can say that the difference in sample mean outputs of the Salesforce will be within the above range. For the 30 Sales force outputs we have taken in this sample, the difference in mean outputs is 4 which falls in the range.

**3.6 Need of Larger Sample Size:**

Steps to follow:

* Get the difference between two sample means (4 as calculated earlier)
* Calculate pooled Standard Deviation using following formula:



* Execute the power T Test, with current parameters, and decide if larger size is needed.
* Calculate the samples number (in case Power of Test is insignificant)

|  |
| --- |
| **## Pooled SD**  **pooledsd<- (((30-1)\*(20.45^2)+(30-1)\*(24.06^2))/(30+30-2))^0.5=22.32** |

**4 Conclusion and Recommendation**

a. Describe the five per cent significance test you would apply to these data to determine whether new scheme has significantly raised outputs?

* We are applying the Paired T test to these data as average of the difference between a series of paired observations is zero.

b. What conclusion does the test lead to?

* From the Preliminary Data Analysis, we confirm that the Mean Sales Force Output of New Scheme (72) is higher than the Mean Sales Force Output of Old Scheme (68).
* When the Data Set explored further with the help of descriptive statistics and Visualization, we learnt that

o New Scheme has more variance   
o No outliers observed in both the samples.  
o Both the samples have nearly Normal distribution; however New Scheme is slightly more skewed towards right.

* **Hypothesis Testing**

The p-value for the two-tailed test is 0.13, which is greater than level of significance α (0.05). Therefore, the Null Hypothesis (H0) will be retained.

c. What reservations have you about this result?

New Scheme is better (As the mean difference of the sample is 4) however we are retaining the Null hypothesis (p-value 0.13 is more than the significant value 0.05) as there is no evidence from the sample.

d. Suppose it has been calculated that in order for Titan to break even, the average output must increase by £5000. If this figure is alternative hypothesis, what is :

(i) The probability of a type 1 error?

Probability of type 1 error is 0.6499(65%) so we don't have enough evidence from the sample the average output is increased by 5000 Pounds.

(ii) The probability of a type 2 error?

Probability of type 2 error is 0.75 (1-0.25=>0.75)

(iii) The power of the test?

0.246 approximated to 25%

e. What sample size would make the probabilities of type 1 and type 2 errors equal?

We need sample size 406 to get probability of type 1 and type 2 error are equal

**5 Appendix–Source Code**

**## Titan Case Study Group Assignment R CODE**

**## getwd()**

**## Read the File Titan Insurance**

**TitanData<-read.csv("TitanInsurance.csv")**

**##Attach the File Titan Insurance**

**attach(TitanData)**

**## Measures of Central Tendency and Dispersion**

**summary(TitanData)**

**##Range**

**range(Old\_Scheme)**

**110-28**

**range(New\_Scheme)**

**122-32**

**##Interquartile Range(Q3-Q1)**

**IQR(Old\_Scheme)**

**IQR(New\_Scheme)**

**##Variance**

**var(Old\_Scheme)**

**var(New\_Scheme)**

**## Standard Deviation**

**sd(Old\_Scheme)**

**sd(New\_Scheme)**

**# Histogram for before and after payment scheme introductions**

**par(mfrow=c(2,2))**

**hist(Old\_Scheme,main='OldScheme',xlab = "POUNDS(000)",ylab = "SalesForce",col = "turquoise")**

**hist(New\_Scheme,main='NewScheme',xlab = "POUNDS(000)",ylab = "SalesForce",col = "turquoise1")**

**##Boxplot for before and after payment scheme introductions**

**y1<-Old\_Scheme**

**y2<-New\_Scheme**

**library(reshape)**

**data<-data.frame(y1,y2)**

**meltData <-melt(data)**

**boxplot(data=meltData, value~variable, na.rm=TRUE,main='Boxplot for Comparision of New and Old Schemes',xlab = "POUNDS(000)",ylab = "SalesForce",col = "turquoise1")**

**## Power T Test**

**power.t.test(n=30, delta = 4, sd=22.32, sig.level = 0.05,type = "two.sample",**

**alternative = "two.sided" )**

**##Pooled Test**

**# Calculation to see need of Larger Sample Size:**

**# Power of the test**

**delta=mean(Old\_Scheme)-mean(New\_Scheme)**

**## Pooled SD**

**pooledsd<- (((30-1)\*(20.45^2)+(30-1)\*(24.06^2))/(30+30-2))^0.5**

**##Pooled sample of two tail test as the variances and means are equal**

**t.test(New\_Scheme,Old\_Scheme, mu=0,paired = TRUE,var.equal = TRUE)**

**##Pooled sample of one tail test as the variances and means are equal**

**t.test(New\_Scheme,Old\_Scheme,alternative = c("greater"), mu=0,paired = TRUE,var.equal = TRUE)**

**##Pooled sample of average output must increase by £5000.**

**t.test(titandata$New\_Scheme,titandata$Old\_Scheme,mu=5,alternative = c("greater"),var.equal=TRUE, paired=TRUE)**

**##Based on the above result(P value :06499)**

**##Probabilty of type 1 error is 0.6499 so we don't have enough evidence from the**

**#sample the average output is increaed by 5000 Pounds.**

**##Probabilty of type 2 error is 1-power(1-0.25==>0.75)**

**##Power of the Test where the average output must increase by £5000**

**power.t.test(n = 30, delta = 4, sd = 22.33, sig.level = 0.05,power = NULL,type = c("paired"),**

**alternative = c("one.sided"),strict = FALSE, tol = .Machine$double.eps^0.25)**

**##We need sample size 406 to get probability of type 1 and type 2 error are equal**

**power.t.test(power = 0.95, delta = 4, sd = 22.33, sig.level = 0.05,**

**type = c("paired"), alternative = c("two.sided"),strict = FALSE, tol = .Machine$double.eps^0.25**)

*#======================================================================= #  
 THE-END*