Name: Pratik Premkishor Mantri

NUID:001353214

Module3

Appendix:

|  |  |
| --- | --- |
| Index | Page no |
| Introduction | 2 |
| Data analysis | 2-9 |
| References | 9 |

Introduction to One-sample T-test in R:

**What is one-sample T-test:**

One sample T-test is nothing but used to compare **the mean** of a sample to a known standard mean( the theoretical mean)

Generally the theoretical mean comes from our previous experiment which we lets say done five years ago.

The T-test can be done only if the data is **Normally distributed.** We can check that using qqplot (quartile-quartile plot).

The research questions could be:

* Whether the mean is equal to theoretical mean
* Whether t is less than theoretical mean
* Whether it is greater than theoretical mean.

In statistics, we can define the corresponding null hypothesis as follow:

1. H0: m=μH0:m=μ
2. H0: m≤μH0:m≤μ
3. H0: m≥μH0:m≥μ

The corresponding *alternative hypotheses* are as follow:

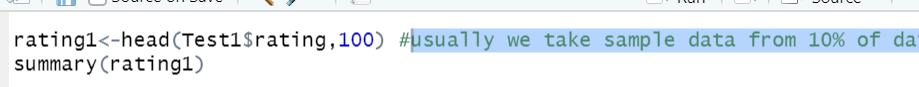
1. Ha: m≠ μHa: m≠μ (different)
2. Ha: m> μHa: m>μ (greater)
3. Ha: m< μHa: m<μ (less)

Here we are doing the T-test for a mean on ‘Rating’ variable from the dataset ‘Wine-tasting.’

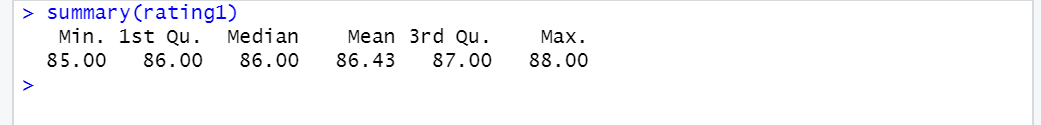
We will be using 100 samples data. **Usually, we take sample data from 10% of data population.**

Here, we will be using an example data set containing the rating of wine.

We want to know, if the average rating of the wine differs from 76.



The 1st 100 sample data of rating can be seen in the R file.



* **Min.**: the min. value
* **1st Qu.**: The first quartile. 25% of values are lower than this.
* **Median**: the median value. Half the values are lower than this and half are higher.
* **3rd Qu.**: the third quartile. 75% of values are higher than this.
* **Max.**: the max. value.

**Visualize data by box plot:**

Logo, company name

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generated

**Preliminary test for checking normality:**

We initially have two variables (Rating and price) in which we can do T-taste but we will select the one with **higher normal distribution**:

**Graphical user interface, text, application, chat or text message

Description automatically generated**

Chart

Description automatically generated

**From the above observation we can say that the data in ‘Rating’ variable is more normally distributed than that of “price” variable.**

**Computing one sample T-test:**

Text

Description automatically generated

Output:

In the result above :

* **t** is the **t-test statistic** value (t = 131.26),
* **df** is the degrees of freedom (df= 99),
* **p-value** is the significance level of the **t-test** (p-value =<2.2e-16).
* **conf.int** is the **confidence interval** of the mean at 95% (conf.int = [86.27233,86.58767]);
* **sample estimates** is the mean value of the sample (mean = 86.43).

Our assumption of mean 76 is false and hence the hypothesis has been rejected and alternative hypothesis has been selected.

Graphical user interface, text, application, email

Description automatically generated

Here due to non-normality concerns we can perform this test to addition ways to ensure about the results are not biased. We will perform t.test as well as non-parametric wilcox.test function. We have also added log function to make the data more normalized.

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**From the above all output we can clearly say that both results agree with our hand calculation which we have done earlier.**

**2) Calculation of P-value:**

p value is probability of obtaining the results which are as extreme as the one which are observed. There are two concepts in P-value analysis. **“Likely”** and “**unlikely”.** If the **p value >0.05**(which is level of confidence), then we can say its likely that we will **reject the Null hypothesis** **and p value is significant to reject the null hypothesis**. If the **(P ≤ 0.05**) then we can say it **“unlikely” to reject the null hypothesis and it is not significant to reject the null hypothesis**

.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**In the above output box**

Ptwo=8.640725e-11 gives us value for two side T-test.

Pvalue is = 4.420362e-11 gives us value for one-side T- test.

Df=degree of freedom

**P > 0.05** then the null hypothesis is **true**. **(P ≤ 0.05)** means test hypothesis is **false** or would be rejected. A **P > 0.05** means that **no effect** was recorded.

Pvalue is = 4.420362e-11<0.5(level of significance)

Here we can see that Pvalue is so small than the level of significance. When p<0.05 then it is to reject the hypothesis because the hypothesis is False.

**3) Looking for maximum price of wine between old wines (French, Italy), new wines (except French, Italy):**

As per directed in the assignment I am using **choice of method,** Firstly I am arranging the ‘Test1’ data frame in ascending of prices. Then I have seen is there any ‘NA’ values presents or not. If any I have omitted these may be null values or ‘blank values.

Text

Description automatically generated

Output:

Chart, bar chart

Description automatically generated

From the above GG-plot we can definitely say that ‘maximum price of wine belongs to Germany which is around $750. From the above graph is crystal clear that from the old wines (France and Italy) France is on the 2nd highest expensive in terms of wines, but Italy does not look that much expensive. There are many wines from ‘new wine category which are more expensive than the Italy (which belongs to Italy).

**References:**

K, M. (2002). One-Sample t-test in R. Retrieved March 21, 2021, from <http://www.sthda.com/english/wiki/wiki.php?title=one-sample-t-test-in-r>

P, M. (2009). Test: Comparing group means. Retrieved March 21, 2021, from <https://uc-r.github.io/t_test#onesample>