

One sample t-test

A one sample t-test is used to determine if the mean value of a group is compared with a certain numeric value, and that numeric value depend upon the problem at hand. In other words, we can say that in order to draw inference for the whole population we take a small sample and run this test. If the test successfully rejects the null hypothesis then we can draw that inference for the population with certain level of confidence level.

Data Description

R have a inbuild dataset called CO2. This data set has 5 columns and 84 rows that are representing information about grass carbon dioxide consumption. For our analysis we are concerned with two columns only, one it the type of grass and second is the carbon dioxide uptake for samples taken in each of the category of grass. We have two categories, one is called Quebec and second is called Mississippi.

But we are only going to compare uptake value of Quebec grass.

Hypothesis Statement

Null Hypothesis H0: The mean uptake value of grass is equal to 30.

Alternative Hypothesis H1: The mean uptake value of grass is greater than 30.

Code

First, I have loaded a in-built dataset called CO2 in variable df. Then I loaded dplyr package in order to filter the dataset. Since it is a one tail test, we are only considering the filtered value of category Quebec. The t.test function is used to run the hypothesis, the first argument is the data itself, the second argument tells us about how we are comparing the mean value of dataset to the value of mu. The mu argument tells us the threshold to which we want our result to be compared.

```
1 df<- co2 #loading the dataset
2
3 library(dplyr) # loading library for filtering data
4
5 a <- df%>% filter(Type=="Quebec") %>% select(uptake) # filtering out uptake value of Quebec
6 result<- t.test(a,alternative="greater",mu=30) # running test in comparison mean value to be 30
7 result # printing the result.
8
```

Result Interpretation

```
one sample t-test

data:  a
t = 2.3734, df = 41, p-value = 0.0112
alternative hypothesis: true mean is greater than 30
95 percent confidence interval:
 31.03082      Inf
sample estimates:
mean of x
 33.54286

> |
```

We see that the p-value is 0.011, meaning that the probability of mean value of uptake carbon dioxide for Quebec grass being equal to 30 is 0.011. In other words, probability of null hypothesis being true is 0.011 which is lower than 0.05 at 95% confidence level.

Hence, we reject the null hypothesis (that mean is equal to 30) towards the alternative hypothesis (mean is greater than 30).

Two sample t-test

A two-sample t-test is conducted to have a conclusion about the difference of mean value of two different groups. The two groups must be independent of each other. However, the type of values measured must be same. For example, we can run a hypothesis test to determine the difference of mean between height of male from one country to height of male from another country.

Data Description

The dataset I have used for this test is called CO2 which is a inbuilt R dataset. The information in it is pertaining to two types of grass and their consumption of CO2. There are 5 columns and 84 rows, but we will only be using two columns, viz. the grass type and CO2 uptake. We have two categories for grass, one is called Quebec and another one is called Mississippi. I have divided the data set in those two categories and filtered their CO2 uptake value for our hypothesis testing.

Hypothesis Statement

H0: The mean difference of two sample of grass is equal.

H1: The mean difference of two sample is greater than zero.

Code

```
1 df <- CO2 #loading the dataset
2
3 library(dplyr) #loading the package
4
5 Quebec <- df %>% filter(Type=="Quebec") %>% select(uptake) #filtering uptake value for category 1
6
7 Mississippi <- df %>% filter(Type=="Mississippi") %>% select(uptake) # filtering uptake value for category 2
8
9 test <- t.test(Quebec$uptake, Mississippi$uptake, alternative = ("greater"),var.equal = F ) # running the tow sample t-test
10
11 test #pringint the value of the test
12
```

First, I have to group the uptake value based on two categorical values we have in Type column. Hence, I created a array of values that have uptake value of Quebec type of grass and second array that have uptake value of Mississippi type of grass.

Second, I used t.test function where the first two arguments are the data that we have to run our hypothesis on. The third argument is used to define our alternative hypothesis, in this case it is stating the alternative hypothesis is testing for mean value to be greater than 0. The last argument used whether to take variance of two sets equal or not. I have given it a false argument, so the variance is considered different because the data is independent of each other.

Interpretation

```
> test <- t.test(Quebec$uptake, Mississippi$uptake, alternative = ("greater"),var.equal = F ) # running the tow sample t-test
>
> test #pringint the value of the test

      Welch Two Sample t-test

data:  Quebec$uptake and Mississippi$uptake
t = 6.5969, df = 78.533, p-value = 2.225e-09
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 9.465352      Inf
sample estimates:
mean of x mean of y
33.54286  20.88333

> |
```

Here we can see p-value is very small. This means the probability of null hypothesis being true is very small. In other words, the probability of the mean difference being equal to zero is too small to be a deciding factor. Also, by looking at the t-value of 6.5 we can say that this value falls on extreme edges of standard normal distribution chart. The result become more promising because it is a two tailed test, hence the critical value of 0.05 will be considered as 0.025 on left and 0.025 on right.

Paired t-test

A paired t-test is again performed for comparison of mean value of two groups where each one of them have continuous data. For example, weight for women vs weight for men. Here the two set are independent of each other because weight of women does not affect the weight of men.

However, paired t-test comes into picture when the two set of data have dependencies among them. For example, first weight of men is calculated and then a medicine is given to those men for certain time period and weight is recorded again.

To illustrate on this, I am going to do a paired t-test on a paired dataset.

Data description

A manager wants to install a music system in the workplace, but he is not sure if it will uplift the mood of the employees or will distract them. Here is the data of 15 employees that is showing the ratings given by employee before and after the music session. Let's run a paired test and see how it effected the environment.

Before = (21,35,40,38,23,27,28,39,22,35,28,20,39,28,34)

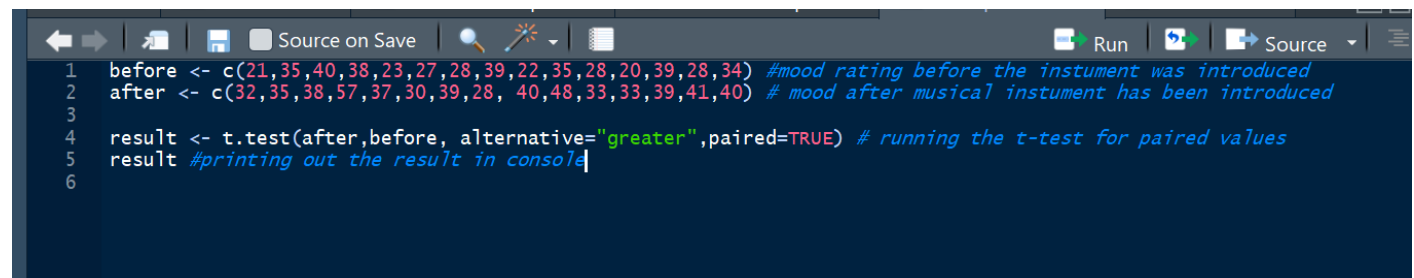
After = (32,35,38,57,37,30,39,28, 40,48,33,33,39,41,40)

Hypothesis statement

Null Hypothesis H0: There is no difference in mood of employees.

Alternative Hypothesis H1: There is an upliftment of mood in employees.

Code



```
1 before <- c(21,35,40,38,23,27,28,39,22,35,28,20,39,28,34) #mood rating before the instument was introduced
2 after <- c(32,35,38,57,37,30,39,28, 40,48,33,33,39,41,40) # mood after musical instument has been introduced
3
4 result <- t.test(after,before, alternative="greater",paired=TRUE) # running the t-test for paired values
5 result #printing out the result in console
6
```

First, I have created two vector dataset that are not independent. Then I have used t.test function where first two arguments takes the data and the third argument takes the comparison type, which in our case is mean greater than zero, and the last argument is paired which I have set to True because the dataset is dependent and a dependent data is assumed to have same properties, such as variance.

Result Interpretation

```
> result  
  
      Paired t-test  
  
data:  after and before  
t = 3.4985, df = 14, p-value = 0.001773  
alternative hypothesis: true difference in means is greater than 0  
95 percent confidence interval:  
 3.740653      Inf  
sample estimates:  
mean of the differences  
      7.533333  
  
> |
```

The p-value is 0.002 which is less than 0.05 at 95% confidence level, which the function takes by default. Based on that **we reject the null hypothesis** that there is no difference in mean value of mood of employees. In other words, we had enough evidence to reject the fact that there is no change in mood.