

Module-3

R Practice



November 21, 2021

Jeseeka Shah

Nuid-002134289

**Introduction**

The dataset plant growth used from the R repository is used in this analysis. Hypothesis testing is conducted here on plant growth dataset using the t.test () to conduct one and two sample test. I have also calculated the p value test by importing the library infer to conduct p value calculations. Plant growth dataset has two features that are weight and group. There are 30 observations of weights for three different groups in total. Figure 1. Shows the statistical information of the dataset. And figure 2 shows the distribution of the values for each group in the dataset. Figure 3 shows the mean of each group present in the dataset in the column group. The total mean for the entire 30 weight is 5.07 (fig.1) and individual mean for ctrl,trt1 and trt2 is 5.032,4.661 and 5.526 respectively.

Table

Description automatically generated with low confidence

Figure 1. shows the mean, standard deviation, range and median of the weights

Table

Description automatically generated Text

Description automatically generated with low confidence

Figure 2. Distribution of the weights Figure 3. Mean of groups

Chart, box and whisker chart

Description automatically generated

Figure 4. Boxplot to show outlier in groups

Figure 4. shows the range of the weight that each group has along with median of each group. The is an outlier in the weights of trt1 group which can be seen in the boxplot. Also, its visible that weights of the group trt2 has range of values mostly above the mean value of the total weights. Figure 5. Shows the total distribution of weights for all groups along with the boxplot showing the median.

Chart, histogram

Description automatically generated

Figure 5. Histogram and Boxplot

Chart, histogram

Description automatically generated

Figure 6. Density of population weight

Figure 6. shows the density distribution of weights of each group present in dataset plant growth. It can be inferred that trt2 has the curve which tends very much close to normal distribution curve followed by trt1 and then on the less side is trt2.

Performing test to check null hypothesis by performing one sample and two sample tests. Hypothesis testing is performed to check the reliability of the sample data statistically. There are 2 types of hypothesis test namely null hypothesis and Alternative hypothesis. Null hypothesis is performed to test the validity of the claim made and alternative hypothesis is performed to disprove the claim made in null hypothesis.

A picture containing schematic

Description automatically generated Chart

Description automatically generated with medium confidence

μ0 = actual mean  
x¯= sample mean  
n = number of observations  
s = Sample standard deviation  
sx¯ = standard error of the mean (s/sqrt(n))

**Performing t.test () to carry out hypothesis testing**

**One Sample Test**

**Case 1**: Null hypothesis claim of sample: *H*0: µt = 0 and *H*1: µ≠ 0, confidence interval is 95 percent and one sample value.

Text

Description automatically generated

After performing the test, the result should that p value is p< 0.05. Therefore, the null hypothesis will be rejected.

***Case 2***: *H*0: µt = 5.1 and *H*1: µ≠ 5.1 where alternative is greater side of the normal distribution

Text, letter

Description automatically generated

Here, p=0.5828 which means p >0.05. Therefore, test has failed to reject the null value as there is not enough evidence to support the claim.

**Case 3:** *H*0: µt = 5.1 and *H*1: µ≠ 5.1 where alternative is less side of the normal distribution

Graphical user interface, text, application, email

Description automatically generated

Here, the one sample t-test shows that p =0.417 which is greater than 0.05 (p>0.05). Therefore, the test fails to reject the null hypothesis.

**Two sample tests**

For conducting two sample tests, I have filtered the groups present in dataset and carried the test. Where sample size of all three group is 10.

Text

Description automatically generated

Text

Description automatically generated

After checking the observation, p=0.2504 which is greater than 0.05. Therefore, we fail to reject the null hypothesis. This is Type I error (false positive) as the hypothesis is accepting the difference in the means is zero which not true.

**Hypothesis Testing for calculating P -value**

Importing the library infer to do the hypothesis testing. The **specify ()** function is used to declare the relation between two variables of dataset**. Hypothesize ()** function in which independence or point is assigned to the null value. In our hypothesis we have assumed independence between two variables weight and group. Next in the pipeline comes **generate ()** function which will create null distribution using **type= “permute**”. In type permute each replicate will have randomly assigned value for input and output respectively. After which we calculate the difference in the mean of the two different group present in the subset of the dataset.Text, letter

Description automatically generated

Above, you can notice the code part of hypothesis as well as the output generated for 1000 random samples. Below, I have considered actual group with there sample size and randomly computed the values for crtl and trt1. Similarly, steps are performed while carrying out crtl and trt2.



A picture containing text

Description automatically generated

P values for the combination can be computed using **get\_p\_values** =function of the infer library. P value for difference in mean of ctrl and trt1 is 0.2 which is greater than 0.05. hence, fail to reject the null hypothesis. P value for combination crtl and trt2 is 0.118 again which is greater than 0.05. Hence, fail to reject the null hypothesis as enough evidence is not present.

A picture containing graphical user interface

Description automatically generated

1. Above, P value for crtl and trt1 b) Below, P value for crtl and trt2

A picture containing logo

Description automatically generated

Now, visualizing the output with respect to the difference in the mean for crtl and trt1 combination on the left and crtl and trt2 combination on the right.



Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**Summary**

After the analysis and hypothesis testing, I have concluded following

* Dataset has 3 groups, trt1 has a outlier as shown in the boxplot and trt2 has weight above in the higher range compared to other two.
* The mean weight of population of dataset plant growth is 5.07.
* Individual mean of crtl, trt1 and trt2 are 5.032, 4.661 and 5.526 respectively.
* After performing the normality test of shapiro.test (PlantGrowth$weight). I inferred that data is very close to the normal distribution which is visible in the density distribution graph too.
* Performing t.test(), in case 1 - *H*0: µt = 0 and *H*1: µ≠ 0 – the results help in rejecting the null hypothesis, in case 2- *H*0: µt = 5.1 and *H*1: µ≠ 5.1 where alternative is greater and left side of the normal distribution – the result fail to reject the null hypothesis as enough evidence was not present.
* Two sample test is carried out on different groups such as crtl and trt1 and the result has failed to reject the null hypothesis as it falls in the region greater than 0.05.
* Carried out hypothesis testing using infer library where I have used the get\_p\_value functionality to get the pvalue and the result was same which indicated fail to reject the null hypothesis.

**References**

* *Getting to Know infer*. (n.d.). Cran-Infer-Library. <https://cran.rproject.org/web/packages/infer/vignettes/infer.html>
* Examples identifying Type I and Type II errors. (n.d.). Khan Academy. <https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/error-probabilities-power/v/examples-identifying-type-i-and-type-ii-errors>
* *How Hypothesis Testing Works*. (2021, October 28). Investopedia. <https://www.investopedia.com/terms/h/hypothesistesting.asp>