

# Data Analysis and Visualization

## Level 4 – Assignment 2

Lecturer: Asim Majeed

Student Name: Aleksandar Kushlev

Student ID: 2028393

## **Abstract:**

This report describes the concepts of t-Test and null hypothesis and analyses a data sample by using a t-Test to make a decision about the likelihood of variance between the two samples. This is done by developing a null and alternative hypothesis and testing them at three significance levels. Other specifics and factors influencing the data have also been elaborated on.

### **1. What is a t-Test used for?**

The t-Test is used in statistics with the purpose of finding out whether the means of two groups of data are different from each other to a degree which is statistically significant. There are two separate types of t-tests. The first type is a t-test in which the sample group members are independent of each other and the second type is a test in which the members of the group are paired.<sup>1</sup> There are some assumptions that should be made for conducting a t-Test, so that the statistical method results are accurate. The assumptions for both the independent and paired sample t-Tests are shown below in Table 1<sup>2</sup>:

|                                  |                                                                                               |
|----------------------------------|-----------------------------------------------------------------------------------------------|
| 1. Continuous                    | The variable of interest should be of any reasonable value                                    |
| 2. Normally Distributed          | The variable of interest should be spread out in a normal way (normally distributed)          |
| 3. Random Sample                 | Data drawn for the sample should be randomly selected                                         |
| 4. Enough Data                   | The sample size should be large enough                                                        |
| 5. Similar Spread Between Groups | The variable of interest is spread similarly between the two groups – homogeneity of variance |

Table 1: Assumptions for conducting a t-Test

The scenario which is the object of the current report includes a data set with independent group members, therefore, the report will focus on conducting a t-Test for independent samples which will be used in the mathematical procedures resulting into a numerical value for the t-Test which is named  $t_{obt}$ . The value of  $t_{obt}$  is proportional to the statistically significant difference between the variable in interest, that is, the higher the value of  $t_{obt}$ , the higher is the chance that the difference between the means of the two samples is statistically significant. The variable of interest to be used for conducting a t-Test in this report will be the means of the independent samples.

---

<sup>1</sup> J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, pp. 129-130

<sup>2</sup> StatsTest.com, "Paired Samples t-Test" & "Independent Samples t-Test", <https://www.statstest.com/paired-samples-t-test/>, (accessed on 29/07/2021)

The value of  $t_{obt}$  is calculated according to a formula which can be adapted for samples of equal and unequal sizes. The formulas for samples of equal and unequal sizes are shown below<sup>3</sup>:

$$t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[ \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2} \right] \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{and} \quad t_{obt} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{n}}}$$

The first formula above calculates the  $t_{obt}$  for samples of unequal sizes and the second is adapted for samples of equal sizes, where:

$t_{obt}$  = the value of  $t$  obtained

$n_1, n_2$  = the sizes of both samples (for unequal sample sizes, for equal sample sizes  $n_1 = n_2 = n$ )

$SD_1^2, SD_2^2$  = the estimates of the variances of the two populations

$\bar{X}_1, \bar{X}_2$  = the means of the two sample groups

If the absolute value of  $t_{obt}$  is large enough to be of statistical significance, the  $H_0$  (null) hypothesis can be rejected. The null hypothesis contradicts the alternative hypothesis.<sup>4</sup>

To determine whether  $t_{obt}$  is large enough so that it is more likely to reflect differences between the two samples, which are not results of chance or randomness, it must be compared with the critical value of  $t$  ( $t_{crit}$ ).<sup>5</sup> The value of  $t_{crit}$  is determined by statistical software and the degree of freedom and level of significance are needed to obtain the exact value.

The degree of freedom is determined by the total number of data points (sum of both samples) minus 2, and it is used to calculate the values for the t-Test. The level of significance is denoted as  $\alpha$  (alpha) and it shows the probability. The most common level of significance chosen is  $\alpha = 0.05$ . Each level of significance corresponds to a different value for  $t_{crit}$ .

Once the value of  $t_{crit}$  is obtained, the  $t_{obt}$  is compared against it and if  $t_{obt}$  is greater or equal to  $t_{crit}$  then we can infer that the difference in the means of both samples is large enough to be statistically significant.

## 2. Null hypothesis and alternative hypothesis.

Statistical techniques are mostly used to demonstrate that a particular hypothesis is false, rather than proving that a hypothesis is true. Therefore, to prove the truthfulness of a statement, the researcher needs to create a hypothesis which states the opposite of what the researcher wants to demonstrate to be true. If the researcher manages to prove that the hypothesis is false, then the initial statement, which the hypothesis contradicts, must be true. The hypothesis that the scientist wants to prove is called  $H_1$ , or alternative hypothesis. The hypothesis that is to be disproved or rejected is called  $H_0$ , or null hypothesis. If the researcher

<sup>3</sup> J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, pp. 132-133

<sup>4</sup> The null hypothesis, the alternative hypothesis and their relation to the t-Test will be described in the next chapter

<sup>5</sup> J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.134

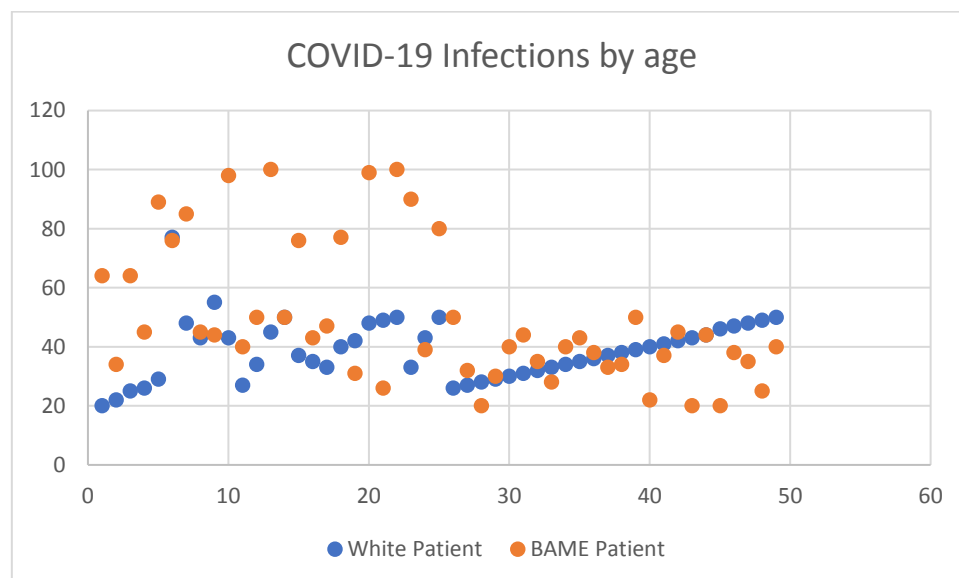
can reject the null hypothesis with the use of statistical techniques, then there would be one thing left to believe and it is the  $H_1$ , the alternative hypothesis.<sup>6</sup>

When the comparison of means of two populations are used to create a null hypothesis and the researcher wants to prove that the means of two populations are really different, then the null hypothesis for the research would be:  $H_0: \mu_1 = \mu_2$ , where  $\mu_1$  is the mean of the first population and  $\mu_2$  is the mean of the second population. If the researcher, by using the t-Test, proves that the difference between the means is large enough to be statistically significant, then the  $H_0$  can be rejected with significant certainty. If the population means are equal and the null hypothesis is true, then it can be inferred that the observed differences between sample means are the result of chance. The t-Test is helping the researcher decide whether the obtained result is due to chance or there really is difference between the populations. Therefore, the t-Test is a way to decide whether the differences are statistically significant to make a conclusion. Even if it is concluded that there is a difference, the question still remains whether it is large enough to have practical usefulness. This question is called by researcher “magnitude of effect”.<sup>7</sup>

The decision to reject or not to reject the null hypothesis is subject to error because the researcher has not studies all members of both populations and because the sample mean and sample standard deviation are subject to sampling error. There are two types of errors that can occur. The first one is when the researcher rejects a true null hypothesis and this error is called Type I error. The failure to reject a false  $H_0$  is called Type II error. The Type I error is considered more serious as it makes the researcher claim significant results when there are no differences between the populations.<sup>8</sup>

### 3. Developing a testable hypothesis.

The experiment which is the subject of this report consists of two groups of people – white and BAME. According to BBC, the infection rate varies between both groups and there is a difference between the average age of infected people between the groups. Both samples are of equal size and are independent (not paired).



Graph 1: COVID-19 infections by age (scatterplot)

<sup>6</sup> J.H. Kranzler, “The t Test” in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.121

<sup>7</sup> J.H. Kranzler, “The t Test” in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.122

<sup>8</sup> J.H. Kranzler, “The t Test” in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.123

As can be seen on the scatterplot above, the infection rate for BAME patients is spread across all age groups as compared to white patients whose infection rate is mostly spread in the range up to 60 years of age.

To prove the high likelihood that infection rates vary between both populations, the means of the above samples should be proven to be different to a statistically significant degree. For this we need to develop a null hypothesis  $H_0: \mu_1 = \mu_2$ , which states that the means of both populations are the same. If the hypothesis is proven false then we can infer that the infection rates vary between both groups of patients which is the alternative hypothesis  $H_1: \mu_1 \neq \mu_2$ .

#### 4. Testing the hypothesis.

With the sample data we have, we can test the null hypothesis by calculating the value for  $t_{obt}$  and comparing it to  $t_{crit}$  so that we can determine whether the difference between sample means is statistically significant.

The testing of the hypothesis is conducted through the following steps:

1. State the hypothesis:
  - $H_0: \mu_1 = \mu_2$
  - $H_1: \mu_1 \neq \mu_2$
2. Select  $\alpha$  (significance level)  
 $\alpha = 0.01, 0.02, 0.10$
3. Calculate the  $t_{obt}$ :

##### 3.1. Significance level $\alpha = 0.01$

| t-Test: Two-Sample Assuming Unequal Variances |               |        |
|-----------------------------------------------|---------------|--------|
|                                               | White Patient | BAME   |
| Mean                                          | 36.58         | 60.04  |
| Variance                                      | 143.51        | 547.18 |
| Observations                                  | 50.00         | 50.00  |
| Hypothesized Mean Difference                  | 0.00          |        |
| df                                            | 73.00         |        |
| t Stat ( $t_{obt}$ )                          | -6.31         |        |
| P(T<=t) one-tail                              | 0.00          |        |
| t Critical one-tail (t-crit one-tail)         | 2.38          |        |
| P(T<=t) two-tail                              | 0.00          |        |
| t Critical two-tail (t-crit two-tail)         | 2.64          |        |

**Decision:** Since the absolute value of  $t_{obt}$  (6.31) is much higher than  $t_{crit}$  in both one-tail (2.38) and two-tail (2.64) versions, we can reject the null hypothesis at significance level  $\alpha = 0.01$

### 3.2. Significance level $\alpha = 0.02$

| t-Test: Two-Sample Assuming Unequal Variances |                      |             |
|-----------------------------------------------|----------------------|-------------|
|                                               |                      |             |
|                                               | <i>White Patient</i> | <i>BAME</i> |
| Mean                                          | 36.58                | 60.04       |
| Variance                                      | 143.51               | 547.18      |
| Observations                                  | 50.00                | 50.00       |
| Hypothesized Mean Difference                  | 0.00                 |             |
| df                                            | 73.00                |             |
| t Stat ( $t_{obt}$ )                          | -6.31                |             |
| P(T<=t) one-tail                              | 0.00                 |             |
| t Critical one-tail (t-crit one-tail)         | 2.09                 |             |
| P(T<=t) two-tail                              | 0.00                 |             |
| t Critical two-tail (t-crit two-tail)         | 2.38                 |             |

**Decision:** Since the absolute value of  $t_{obt}$  (6.31) is much higher than  $t_{crit}$  in both one-tail (2.09) and two-tail (2.38) versions, we can reject the null hypothesis at significance level  $\alpha = 0.02$

### 3.3. Significance level $\alpha = 0.10$

| t-Test: Two-Sample Assuming Unequal Variances |                      |             |
|-----------------------------------------------|----------------------|-------------|
|                                               |                      |             |
|                                               | <i>White Patient</i> | <i>BAME</i> |
| Mean                                          | 36.58                | 60.04       |
| Variance                                      | 143.51               | 547.18      |
| Observations                                  | 50.00                | 50.00       |
| Hypothesized Mean Difference                  | 0.00                 |             |
| df                                            | 73.00                |             |
| t Stat ( $t_{obt}$ )                          | -6.31                |             |
| P(T<=t) one-tail                              | 0.00                 |             |
| t Critical one-tail (t-crit one-tail)         | 1.29                 |             |
| P(T<=t) two-tail                              | 0.00                 |             |
| t Critical two-tail (t-crit two-tail)         | 1.67                 |             |

**Decision:** Since the absolute value of  $t_{obt}$  (6.31) is much higher than  $t_{crit}$  in both one-tail (1.29) and two-tail (1.67) versions, we can reject the null hypothesis at significance level  $\alpha = 0.10$

## 5. Outliers in the data.

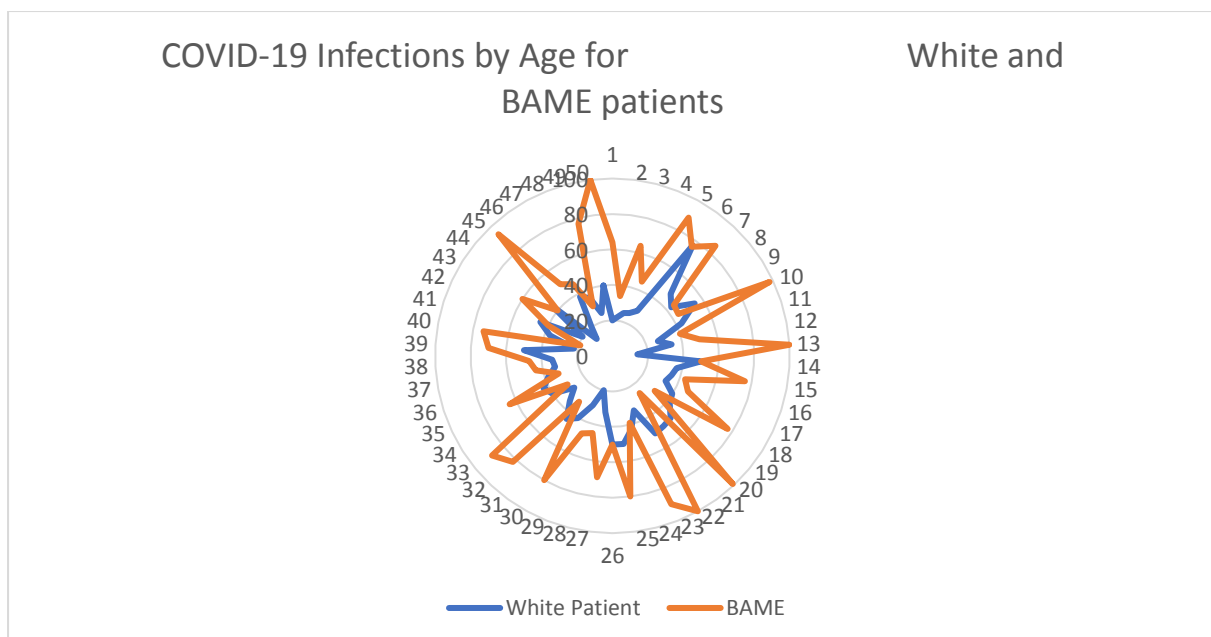
In the experimental sample there are 2 occasions in the white patients sample where the age is below 20 years (13 and 14), and one occasion in the BAME sample (19). If the t-Test is adapted with the above occasions removed from the sample, the  $t_{obt}$  increases slightly for the three significance levels but this does not affect the final decision to reject the null hypothesis as the value of  $t_{obt}$  is still much greater than  $t_{crit}$ .

| t-Test: Two-Sample Assuming Unequal Variances |               |        |
|-----------------------------------------------|---------------|--------|
| $\alpha = 0.01$                               |               |        |
|                                               | White Patient | BAME   |
| Mean                                          | 37.54         | 60.88  |
| Variance                                      | 126.00        | 522.78 |
| Observations                                  | 48.00         | 49.00  |
| Hypothesized Mean Difference                  | 0.00          |        |
| df                                            | 70.00         |        |
| t Stat                                        | -6.40         |        |
| P(T<=t) one-tail                              | 0.00          |        |
| t Critical one-tail                           | 2.38          |        |
| P(T<=t) two-tail                              | 0.00          |        |
| t Critical two-tail                           | 2.65          |        |

| t-Test: Two-Sample Assuming Unequal Variances |               |        |
|-----------------------------------------------|---------------|--------|
| $\alpha = 0.02$                               |               |        |
|                                               | White Patient | BAME   |
| Mean                                          | 37.54         | 60.88  |
| Variance                                      | 126.00        | 522.78 |
| Observations                                  | 48.00         | 49.00  |
| Hypothesized Mean Difference                  | 0.00          |        |
| df                                            | 70.00         |        |
| t Stat                                        | -6.40         |        |
| P(T<=t) one-tail                              | 0.00          |        |
| t Critical one-tail                           | 2.09          |        |
| P(T<=t) two-tail                              | 0.00          |        |
| t Critical two-tail                           | 2.38          |        |

## 6. Other features of the data and factors affecting the results.

On the graph below we can notice that the sample for white patients contains predominantly ages of up to 60 years, and the sample for BAME patients is more dispersed over the 20 – 100 years range.



Graph 2: COVID-19 Infections by Age (Radar)

Since one of the assumptions for conducting a t-Test is for similar spread (variance) between groups, this condition is not very well adhered to in the experiment as the BAME group tends to have a much higher variance (547.18) vs. the white patients group variance (143.51). This discrepancy might result into misleading results.

One solution to the above problem could be to limit the sample age to 60 or to randomize the samples in such way that there are more white patients in the 60-100 age bracket so that both samples are more uniform and have similar variances.

## **7. Conclusion.**

The t-Test conducted above for significance levels  $\alpha = 0.01, 0.02, 0.10$ , leads to the decision to reject the null hypothesis  $H_0 : \mu_1 = \mu_2$ . This results from comparing the value of  $t_{obt}$  with  $t_{crit}$  which shows that the difference between the sample means is statistically significant, therefore we can conclude that infection rates for both populations vary.

## References:

- [1] J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, pp. 129-130
- [2] StatsTest.com, "Paired Samples t-Test" & "Independent Samples t-Test", <https://www.statstest.com/paired-samples-t-test/> , (accessed on 29/07/2021)
- [3] J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, pp. 132-133
- [4] J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.134
- [5] J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.121
- [6] J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.122
- [7] J.H. Kranzler, "The t Test" in *Statistics for the terrified*, Rowman and Littlefield, New York (USA), 2018, p.123