

Data Analysis and Visualization

Assessment 2 Submission Date:03/09/2021

Table of Contents

1. Introduction.	3
2. What do you mean by t-Test?	4
3. What is Hypothesis? How to develop null and alternative Hypotheses? ...	4
4. Hypotheses that are relevant to the analysis.	5
5. The methods and procedures for presenting the results.....	5
6. Type 1 error and type 2 error.....	9
7. Outliers in dataset and their impact.	10
8. Factors affecting hypotheses test.	13
9. Results – Investigation.	13

1. Introduction.

In this document, data from BBC(UK) News shows that the infectious disease rate of Covid-19 varies between White and BAME patients. Fifty patients from both communities had chosen to test the data's accuracy using a T-test two-sample assuming unequal variances hypothesis. Multiple factors which affect the Hypothesis, the methods and recommendations for better hygienic practices are the subject of debate. Outliners in data sets are under investigation, and their effects; eliminating outliers can aid in better hypothesis testing—the methods for detecting Gaining an understanding and removing them from Datasets.

The dataset has been present when it became known that the Covid-19 infection rate differed between White and Black and minority ethnic groups in the United Kingdom. The dataset includes 50 patients from both the community and the hospital. The age of the participants who provided the data ranged from 20 to 100 years old. These people's data are under investigation to determine whether or not they are telling the truth. Several data sets have worked to form hypotheses. Finally, the Hypothesis is tested on both groups of citizens to see if the infection rate changes between them.

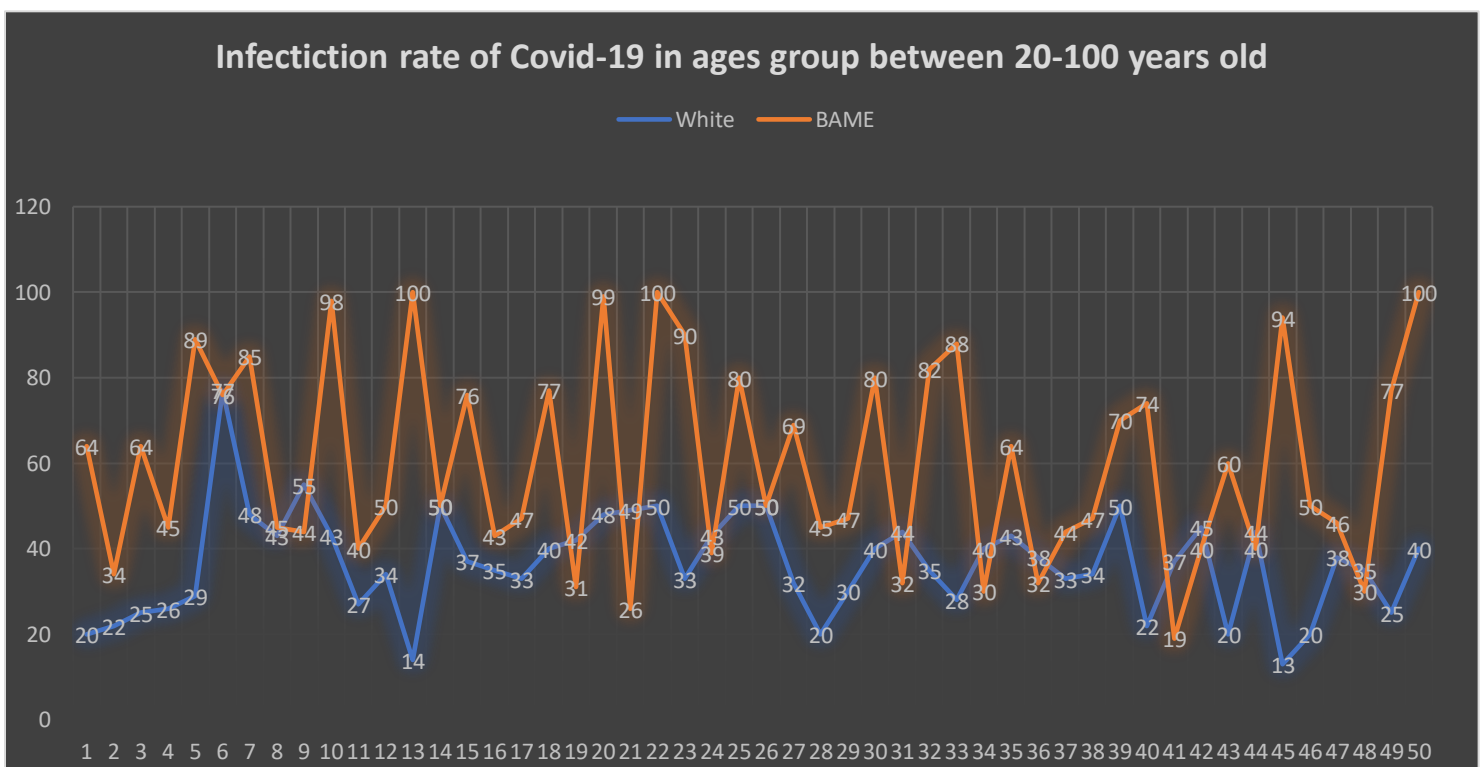


Figure 1. Visualization White and BAME by ages.

Graphic representation of patients' demographics from the white and BAME populations, with the ages changing. The line graph depicts the age differences, and it stands out that the infection rate varies depending on age.

2. What do you mean by t-Test?

It's a sort of inferential statistic used to see if two groups have a statistically significant difference. It proves the problem mathematically by assuming that the means of the two populations are identical ($H_0: \mu_1 = \mu_2$). If the t-test rejects the null Hypothesis ($H_0: \mu_1 = \mu_2$), the groups are very likely to differ [2].

The t-test proves the statement of the problem mathematically by taking a sample from each of the two sets and assuming a null hypothesis that the two means are identical. Then, specific values are computed and compared to the normal values using the appropriate formulae, and the presumed null Hypothesis is accepted or rejected accordingly [2].

To begin, you must identify the Hypothesis you will test and the acceptable risk of reaching an incorrect result. When comparing two populations, for example, you may assume that their means are the same and decide on a sufficient likelihood of finding that a difference exists if this is not the case. Then, using your data, you compute a test statistic and compare it to a theoretical value from a t-distribution. Finally, you either reject or fail to reject your null Hypothesis depending on the outcome [2].

The Null Hypothesis has rejected the t-test, which refuses or fails to leave it [2].

3. What is Hypothesis? How to develop null and alternative Hypotheses?

Hypothesis: a hypothesis is an assumption or explanation for anything based on known facts but has yet to be proven.

In other words, it's a statement that, although based on facts (or some acceptable assumptions), explains why or how something works but hasn't a thorough examination. A hypothesis may, for example, be a text as follows:

Consumers' willingness to trust a financial adviser affects their opinion of the financial advisor's competence and expertise.

We're remarking on the relationship between perceived skill and trustworthiness in this case [3].

Although the connection between the two appears plausible and logical, it has not had a firm foundation (at least not in a scientifically rigorous way in every possible context). As a result, this is a theory (loosely speaking). However, in research, a hypothesis must meet a few more requirements to be considered a legitimate research hypothesis or scientific Hypothesis. Let's look at these factors in more detail [3].

As you can see, the Hypothesis is quite detailed in that it defines the factors at play (expertise and confidence), the people involved (a client and an adviser), and the sort of relationship that should be expected (positive correlation). Consequently, there is no ambiguity or misunderstanding about who is engaged in the statement, and the desired outcome is evident [3].

To be considered a scientific hypothesis, a hypothesis must be testable. It's not a hypothesis if it is not possible to test it. Testability refers to the ability to collect observable evidence in a scientifically rigorous manner to determine if the Hypothesis is supported or not. Put another way, and there must be a means to back up the assertion [3].

4. Hypotheses that are relevant to the analysis.

T-test: two-sample assuming unequal variances:

That has completed on two separate populations with different variability. Thus, there are two sorts of tests: one with two tails and one with one behind. As a result, the dataset has two populations from separate communities, and the operation is run on them to determine the data's accuracy. Because the variances of the two sets of data differ, this test takes place.

5. The methods and procedures for presenting the results.

Null hypotheses H_0 : This is a type of Hypothesis that shows that the characteristics of two groups are the same.

$H_0 = \text{Mean1 and Mean2} = 0$

Alternative hypotheses are H_a : Indicates that the features of one population do not correspond to those of another. With enough data, the null Hypothesis may be rejected, and an alternative hypothesis can be favored [1].

H_a = Mean1 and Mean2 are not equal to zero [1].

Stating the Hypotheses.

H_o = In the 50 patients, White and BAME have no difference in the ratio of the infection rate.

H_a = In the 50 patients, White and BAME have a difference in the ratio of the infection rate.

Testing the hypotheses to the various significance level.

- $p=0.10$

The hypotheses are calculated for various significance levels, as the significance level is 0.10.

1. If the t Stat < t Crit two tails, this is statistically significant and null hypotheses cannot be rejected, but in this case, 6.312 is not less than 1.665, the value of Null hypotheses will be rejected. No evidence for equality between the two location means. Therefore, the infection rate between white and BAME patients cannot be considered equal.

2. It may be observed that the p-value of the two tails, 1.909 is greater than the value of significance Alpha, which is 0.10. The null hypotheses cannot be rejected.

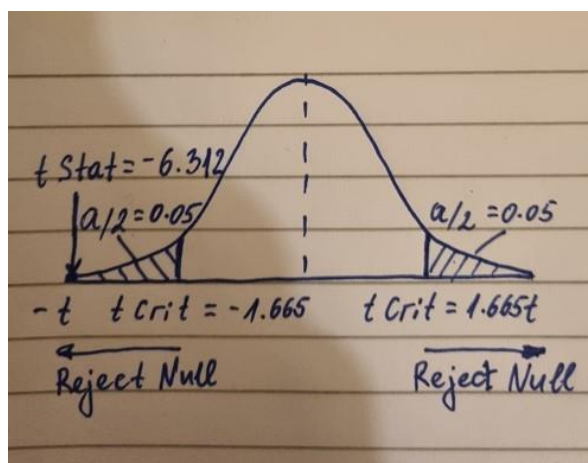


Figure 2. Graphical representation for $p=0.10$

t-Test: Two-Sample Assuming Unequal Variances		
	White	BAME
Mean	36.58	60.04
Variance	143.5138776	547.182041
Observations	50	50
Hypothesized Mean Difference	0	
df	73	
t Stat	-6.312037424	
P(T<=t) one-tail	9.54795E-09	
t Critical one-tail	1.293256413	
P(T<=t) two-tail	1.90959E-08	
t Critical two-tail	1.665996224	

Figure 3. Test for $p=0.10$

- $p=0.02$

The same method is put into practice to calculate the significance value. The following is the value derived from the result:

1. That if $t \text{ Stat} < t \text{ Crit two-tail value}$. It is possible to reject the null hypotheses. Here the value of $t \text{ Stat}$ is 6.312, which is more significant than $t \text{ Crit}$ is 2.378. So null hypotheses can be rejected.
2. It can be seen that $p\text{-value two tails } 1.909$ is greater than the value of significance Alpha who is 0.02. The null hypotheses cannot be rejected. They have statistically significant difference.

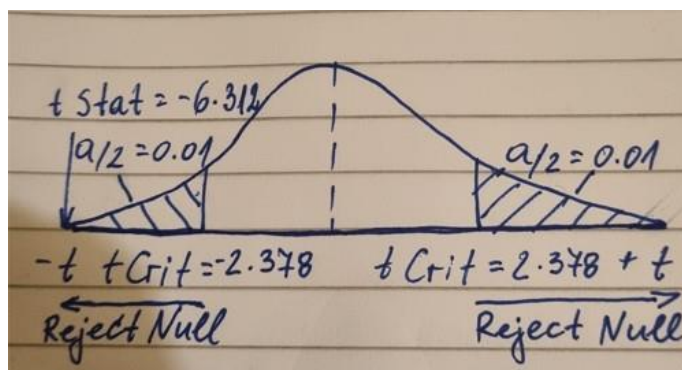


Figure 4. Graphical representation for $p=0.02$

t-Test: Two-Sample Assuming Unequal Variances		
	White	BAME
Mean	36.58	60.04
Variance	143.5138776	547.1820408
Observations	50	50
Hypothesized Mean Difference	0	
df	73	
t Stat	-6.312037424	
P(T<=t) one-tail	9.54795E-09	
t Critical one-tail	2.091096978	
P(T<=t) two-tail	1.90959E-08	
t Critical two-tail	2.378521878	

Figure 5. Test for $p=0.02$

- $p=0.01$

1. That if the t Stat value is less than t Crit two-tail value. The null hypotheses will be rejected. For example, here, the value of t Stat is 6.312, greater than 2.644, and the null hypotheses will be rejected.
2. It is seen that p-value two tails 1.909 is greater than the value of significance Alpha who is 0.01. The null hypotheses cannot be rejected.

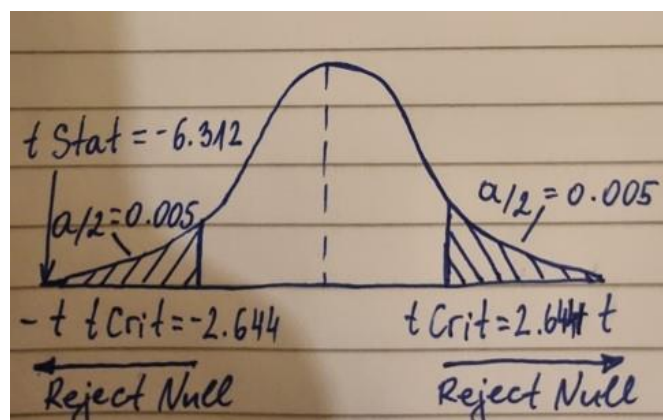


Figure 6. Graphical representation for $p=0.01$

t-Test: Two-Sample Assuming Unequal Variances		
	White	BAME
Mean	36.58	60.04
Variance	143.5138776	547.1820408
Observations	50	50
Hypothesized Mean Difference	0	
df	73	
t Stat	-6.312037424	
P(T<=t) one-tail	9.54795E-09	
t Critical one-tail	2.378521878	
P(T<=t) two-tail	1.90959E-08	
t Critical two-tail	2.644868782	

Figure 7. Test for $p=0.01$

6. Type 1 error and type 2 error.

In hypothesis testing, the chances of making these mistakes are inherent since making a statistical judgement always entails uncertainty.

The chance of making a Type I error is determined by the level of significance, or alpha, but the possibility of making a Type II error is defined by beta. These risks can be mitigated by planning your study strategy properly [7].

Type I and Type II Error		
Null hypothesis is...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β

Figure 8. Type I and Type II error [7]

Once the null hypothesis is rejected whenever it's completely true, this is known as a Type I error. It includes asserting statistical significance for outcomes that were acquired by happenstance or due to extraneous factors.

The chance of committing this error is determined on the significance level (alpha or) you choose. This is a figure you come up with at the outset of your study to assess the statistical probability of receiving your findings (p value).

Your results are statistically significant and support the alternative hypothesis if the p value of your test is less than the level of significance. If your p value is larger than the significance level, your results are statistically non-significant. When the null hypothesis is not rejected when it is false, this is known as a Type II error. That's not the same as “having to accept” the null hypothesis, as hypothesis testing can only tell you if it should be rejected.

A Type II error, on either side, occurs when you are unaware of an effect while one exists. In reality, your research might not have had enough statistical power to detect a meaningful effect.

When an actual effect occurs, a test's power affects how well it can detect it. A peak power of 80% or above is typically considered acceptable [7].

7. Outliers in dataset and their impact.

In essential words, an outlier is a data point in a data graph or dataset that is very high or extremely low compared to the nearest data point and the rest of the adjacent co-existing values. Outliers are exceptional values in a dataset or graph that deviate significantly from the main pattern of values. They disrupt the exam, resulting in an unsatisfactory outcome. In our requirements, we have to identify the ratio of the truthfulness of the findings—fifty different white and BAME patients between 20 and 100 years. In the data set can find three exceptional values less than 20 years. The means and standard deviation have a highly significant difference. So the presence of outliers is already detected and found, but just in case, to be sure, we can do some tests and graphs to find out [5].

1. Sorting the Data is the first step. Sorting data can help you find outliers in your data. The data set's horizontal and vertical sequence will provide unusually low or high data values [4].
2. Create graphs based on the information provided: Scatter plots and Box plots are examples of charts that may quickly show distinct and placed in different locations [4].
3. Determining the z-score. It's possible to utilize the z-score to identify data out of the ordinary in a normal distribution. The Z-score may be figured out in every observation by deducting the mean from the value and dividing it by the standard deviation. The outlier may be detected since the z-score is high for higher values and low for lower ones [4].

Patient No	White Patient	BAME	Patient No	White Patient	BAME
1	20	64	26	50	50
2	22	34	27	32	69
3	25	64	28	20	45
4	26	45	29	30	47
5	29	89	30	40	80
6	77	76	31	44	32
7	48	85	32	35	82
8	43	45	33	28	88
9	55	44	34	40	30
10	43	98	35	43	64
11	27	40	36	38	32
12	34	50	37	33	44
13	14	100	38	34	47
14	50	50	39	50	70
15	37	76	40	22	74
16	35	43	41	37	19
17	33	47	42	45	40
18	40	77	43	20	60
19	42	31	44	44	40
20	48	99	45	13	94
21	49	26	46	20	50
22	50	100	47	38	46
23	33	90	48	35	30
24	43	39	49	25	77
25	50	80	50	40	100

Figure 9. Outliners in dataset

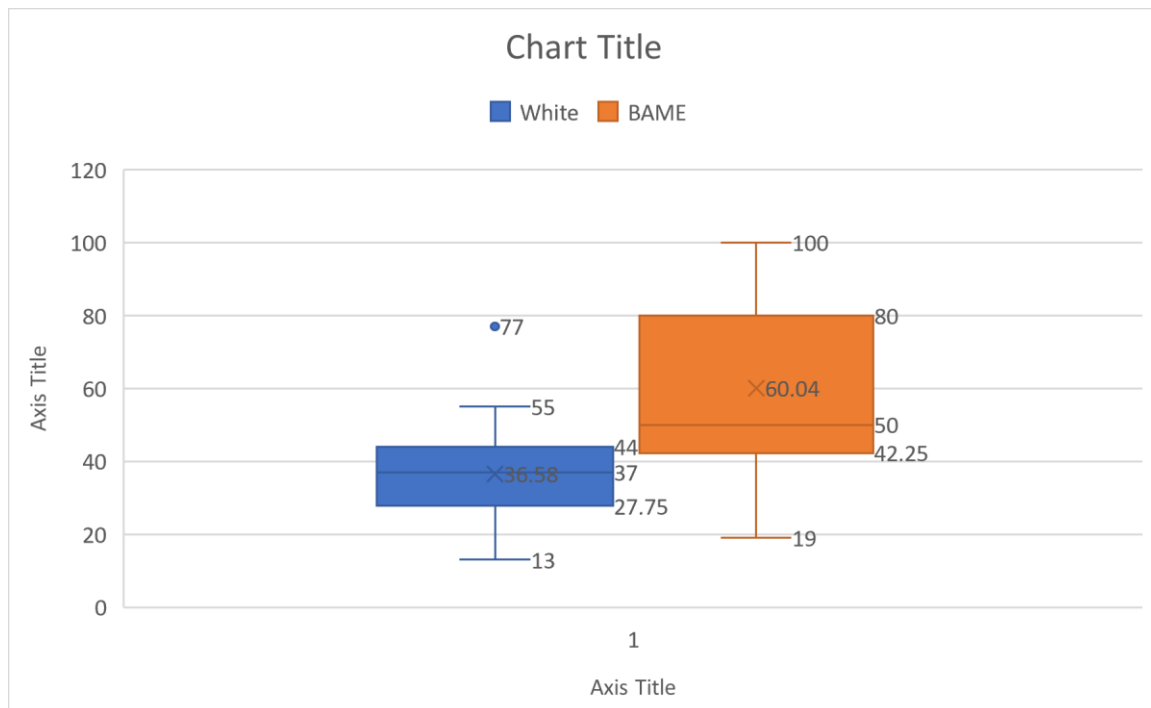


Figure 10. Graph showing outliners

8. Factors affecting hypotheses test.

1. The sample size (n). If the population size is larger, the test's power will be higher; but, if the representative sample is less and the data trend is unknown, the test's power will be lower. As a result, the sample size must be substantial. The test is ineffective because the fewer data could perhaps make clear the correlation that the particular data set is pursuing. In the case studied, there are only 50 samples are much less, and the test is ineffective because the fewer data cannot clarify the trend and pattern that the particular data set is attempting to follow. The null values are further rejected in the testing procedure as a result of this poor hypothesis testing.

2. Level of significance (α). The test's power increases as the significance level rises. When the significance level has been elevated, the acceptability zone shrinks. As a result, the Null Hypothesis is more likely to be rejected. That implies you're less likely to accept the Null Hypothesis when it's untrue; in other words, you're less likely to make a Type II mistake. As a result, the test's power improves.

3. The parameter being the tester's "true" value. As the gap between the "real" value of a parameter and the value indicated in the null Hypothesis rises, so does the test's power. In other words, the bigger the effect size, the stronger the test's ability.

Furthermore, the likelihood of making a Type II error rises as a Type I error decreases. Thus, it isn't easy to reduce the possibility of both a Type I and Type II test simultaneously [8].

9. Results – Investigation.

From the three observations, It's obvious. Both cases of t Stat have supplied. In three situations, the Null Hypothesis will be rejected. Consequently, the values of infection in both populations of varied ages (20-100) may be considered equivalent, with no significant differences, and Alternative hypotheses are possible to embrace. The null hypotheses can illustrate the belief that the infection rate fluctuates and the previous evidence and perceptions.

Outliers were found in the data and influenced the hypothesis test results in the data set.

The outcome demonstrates the integrity of the data at each significance level, with $p=10$ being the most significant.

Bibliography

[1]

J. Frost, "Alternative hypothesis," *Statistics By Jim*, Nov. 24, 2019.
<https://statisticsbyjim.com/glossary/alternative-hypothesis/> (accessed Aug. 29, 2021).

[2]

J. Fernandez, "The statistical analysis t-test explained for beginners and experts," *Medium*, Jun. 08, 2020. <https://towardsdatascience.com/the-statistical-analysis-t-test-explained-for-beginners-and-experts-fd0e358bbb62> (accessed Aug. 29, 2021).

[3]

D. Jansen, "What Is A Research Hypothesis? A Simple Definition," *Grad Coach*, Jun. 08, 2020.
<https://gradcoach.com/what-is-a-research-hypothesis-or-scientific-hypothesis/> (accessed Aug. 29, 2021).

[4]

A. Birkett, "How to Deal with Outliers in Your Data," *CXL*, Aug. 24, 2019.
<https://cxl.com/blog/outliers/> (accessed Aug. 30, 2021).

[5]

D. Mishra, "Handling Outliers in Datasets," *Medium*, Sep. 16, 2020.
<https://medium.com/@dhirajmishra57/handling-outliers-in-datasets-3740b04b52a2> (accessed Sep. 01, 2021).

[6]

W. Kenton, "How T-Tests Work," *Investopedia*, Mar. 22, 2020.
<https://www.investopedia.com/terms/t/t-test.asp> (accessed Sep. 01, 2021).

[7]

P. Bhandari, "Type I & Type II Errors | Differences, Examples, Visualizations," *Scribbr*, Jan. 18, 2021. <https://www.scribbr.com/statistics/type-i-and-type-ii-errors/> (accessed Sep. 01, 2021).

[8]

Maths and Physics tuition, "Factors Affecting the Power of a Hypothesis Test," *astarmathsandphysics.com*, Sep. 01, 2021. <https://astarmathsandphysics.com/a-level-maths-notes/s4/3804-factors-affecting-the-power-of-a-hypothesis-test.html> (accessed Sep. 01, 2021).