

University of Wales Trinity Saint David

Data Analysis and Visualisation

ACCB4002

Assignment 2

Lecturer: Anupam Mazumdar

Student: Andreea Filip - 1817999

Contents

List of tables.....	2
List of figures	2
Introduction	3
Dataset.....	3
Statistical analysis	4
T-test	4
Testing the hypothesis at significant levels of 0.10, 0.01, 0.02.....	9
Factors that can affect the results.....	10
Conclusion	11
References	12
Appendix I. z-distribution	13

List of tables

Table 1. The dataset

Table 2. Deviation scores for white patients and BAME patients.

List of figures

Figure 1. White patients' age distribution curve.

Figure 2. BAME patients' age distribution curve

Figure 3. The distribution curve of white and BAME patients

Introduction

In this report, I will analyse the probability of existing varied infection rates with the COVID-19 between white and BAME people. In order to verify this, a sample of 100 patients will be used, 50 patients from each mentioned category. To analyse the possibility of a significant difference in the ratio between the two categories, I will use a two-tail T-test assuming unequal variance for independent samples. The report will also include supplementary observations and factors that might influence the results.

Dataset

Table 1. The dataset

Patient No	White Patient	BAME	Patient No	White Patient	BAME
1	20	23	26	50	50
2	22	34	27	32	33
3	25	44	28	20	45
4	26	45	29	30	47
5	29	29	30	40	20
6	27	27	31	44	32
7	48	35	32	35	22
8	43	45	33	28	28
9	55	44	34	40	30
10	43	30	35	43	34
11	27	40	36	38	32
12	34	50	37	33	44
13	44	50	38	34	47
14	50	50	39	50	30
15	37	41	40	22	44
16	35	43	41	37	29
17	33	47	42	45	40
18	40	32	43	20	30
19	42	31	44	44	40
20	48	20	45	33	24
21	49	26	46	20	50
22	30	28	47	38	46
23	33	30	48	35	30
24	43	39	49	25	27
25	50	50	50	40	50

It can be observed that the dataset is compounded of two sets of data, representing patients with ages between 20 and 55 years. The first one presents 50 patients

characterised by being in the white ethnic group, and the second one presents 50 patients that are part of the BAME ethnic group. All patients have been infected with the new Coronavirus, COVID-19. Moreover, it can be noticed that the two sets of data are not dependent on one another in any way.

Statistical analysis

In order to verify if the infection rate is varied between white and BAME people, the sample will be analysed using T-test for independent samples.

T-test

T-test or a student's T-test is an important statistical test used to verify if there is a significant level of difference between the means of two groups of datasets [1], eliminating the bias and providing an objective interpretation of the difference.

For verifying if a set of data is significantly different, the T-test provides a critical value, a probability value that demonstrates if the data is random or significant. In scientific research, this value should be less than 0.05, to demonstrate that there is a significant difference between the means of the two data sets. A P value of 0.05 indicates that there is a 5% chance that the data is random, but a 95% chance that the data is significant.

T-test can be used for paired and unpaired sets of data, paired sets of data are dependent, or linked [2]. In addition, depending on the hypothesis, the T-tests can be one-tailed and two-tailed.

For this set of data, an independent T-test will be conducted, as mentioned, the data is not linked, the values in one sample do not affect the values in the second sample.

The T-Test will be conducted using two hypotheses, H_0 , the null hypothesis and H_A , the alternative hypothesis, as following:

H_0 - There is no significant difference between the ratio of infection rate with Covid-19 virus of white and BAME patients.

H_A - There is a significant difference between the ratio of infection rate with Covid-19 of white and BAME patients.

Table 2. Deviation scores for white patients and BAME patients.

Patient No	White Patient (X)	x	x ²	BAME Patients (Y)	y	y ²	x*y
1	20	-16.18	261.7924	23	-	-	222.313
2	22	-14.18	201.0724	34	13.74	188.7876	2
3	25	-11.18	124.9924	44	-2.74	7.5076	38.8532
4	26	-10.18	103.6324	45	7.26	52.7076	-81.1668
5	29	-7.18	51.5524	45	8.26	68.2276	-84.0868
6	27	-9.18	84.2724	29	-7.74	59.9076	55.5732
7	48	11.82	139.7124	27	-9.74	94.8676	89.4132
8	43	6.82	46.5124	35	-1.74	3.0276	-20.5668
9	55	18.82	354.1924	45	8.26	68.2276	56.3332
10	43	6.82	46.5124	44	7.26	52.7076	136.633
11	27	-9.18	84.2724	30	-6.74	45.4276	2
12	34	-2.18	4.7524	40	3.26	10.6276	-45.9668
13	44	7.82	61.1524	50	13.26	175.8276	-29.9268
14	50	13.82	190.9924	50	13.26	175.8276	-28.9068
15	37	0.82	0.6724	41	4.26	18.1476	103.693
16	35	-1.18	1.3924	43	6.26	39.1876	2
17	33	-3.18	10.1124	47	10.26	105.2676	183.253
18	40	3.82	14.5924	32	-4.74	22.4676	2
19	42	5.82	33.8724	31	-5.74	32.9476	3.4932
20	48	11.82	139.7124	20	-	-	-7.3868
21	49	12.82	164.3524	26	16.74	280.2276	-32.6268
22	30	-6.18	38.1924	28	-	-	-18.1068
23	33	-3.18	10.1124	30	-8.74	76.3876	-33.4068
24	43	6.82	46.5124	39	-6.74	45.4276	-
25	50	13.82	190.9924	50	2.26	5.1076	15.4132
26	50	13.82	190.9924	50	13.26	175.8276	183.253
							2
							183.253
							2

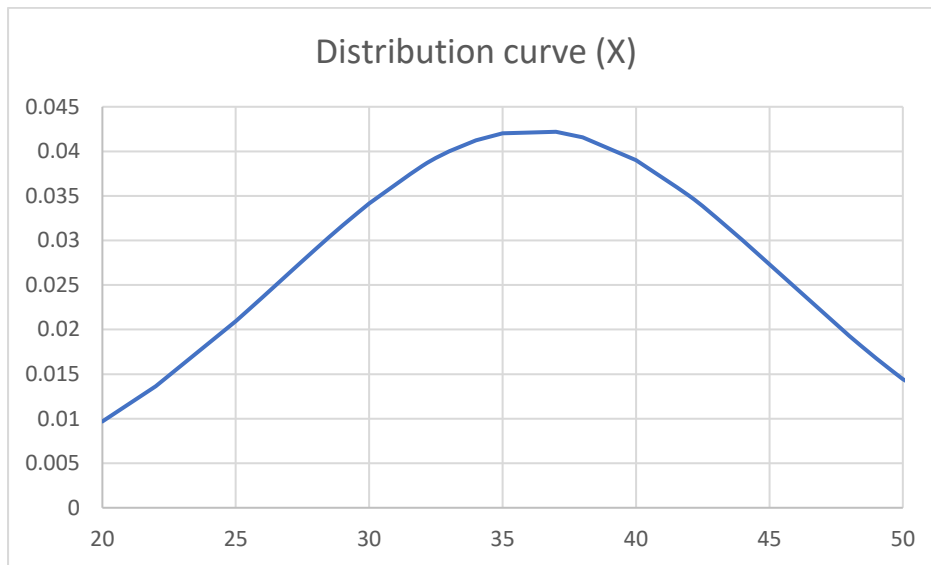
			190.9924				
27	32	-4.18	17.4724	33	-3.74	13.9876	15.6332
28	20	-16.18	261.7924	45	8.26	68.2276	-133.647
29	30	-6.18	38.1924	47	10.26	105.2676	-63.4068
30	40	3.82	14.5924	20	16.74	280.2276	-63.9468
31	44	7.82	61.1524	32	-4.74	22.4676	-37.0668
32	35	-1.18	1.3924	22	14.74	217.2676	17.3932
33	28	-8.18	66.9124	28	-8.74	76.3876	71.4932
34	40	3.82	14.5924	30	-6.74	45.4276	-25.7468
35	43	6.82	46.5124	34	-2.74	7.5076	-18.6868
36	38	1.82	3.3124	32	-4.74	22.4676	-8.6268
37	33	-3.18	10.1124	44	7.26	52.7076	-23.0868
38	34	-2.18	4.7524	47	10.26	105.2676	-22.3668
39	50	13.82	190.9924	30	-6.74	45.4276	-93.1468
40	22	-14.18	201.0724	44	7.26	52.7076	-102.947
41	37	0.82	0.6724	29	-7.74	59.9076	-6.3468
42	45	8.82	77.7924	40	3.26	10.6276	28.7532
43	20	-16.18	261.7924	30	-6.74	45.4276	109.053 2
44	44	7.82	61.1524	40	3.26	10.6276	25.4932
45	33	-3.18	10.1124	24	12.74	162.3076	40.5132
46	20	-16.18	261.7924	50	13.26	175.8276	-214.547
47	38	1.82	3.3124	46	9.26	85.7476	16.8532
48	35	-1.18	1.3924	30	-6.74	45.4276	7.9532 108.893
49	25	-11.18	124.9924	27	-9.74	94.8676	2
50	40	3.82	14.5924	50	13.26	175.8276	50.6532
50	36.18	1.42109E-14	4347.38	36.74		4253.62	308.34

$$\bar{x} = \frac{\sum x}{n} = \frac{1809}{50} = 36.18$$

$$SD_x = \sqrt{\frac{\sum x^2}{n-1}} = \sqrt{\frac{4347.38}{50-1}} = 9.4192$$

The average age of the white patients is 36.18, with a standard deviation of 9.42. For a visual representation of these findings I have used the bell curve graph.[3]

Figure 1. White patients' age distribution curve.

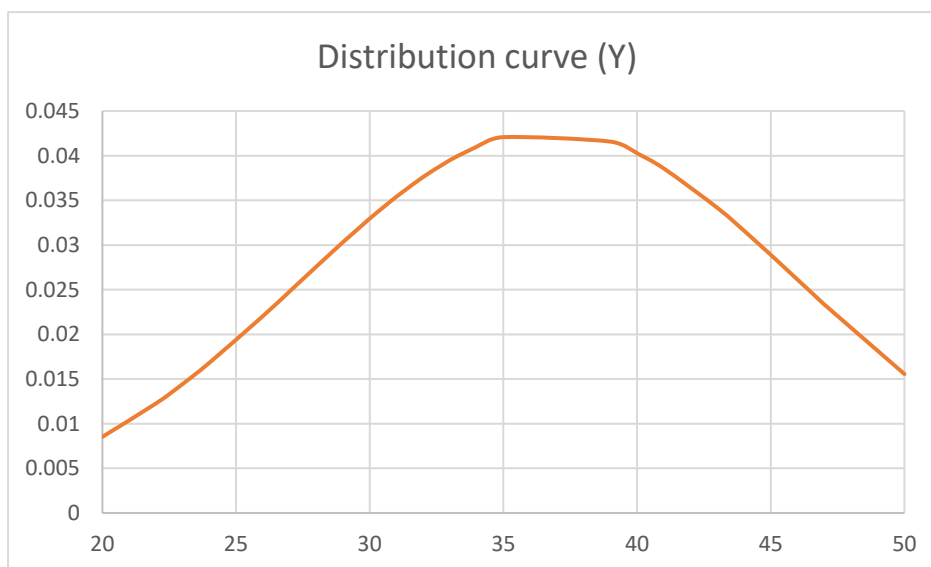


$$\bar{y} = \frac{\sum y}{n} = \frac{1837}{50} = 36.74$$

$$SD_y = \sqrt{\frac{\sum y^2}{n-1}} = \sqrt{\frac{4253.62}{50-1}} = 9.3171$$

The BAME patients' average age is 36.74 with a standard deviation of 9.32.

Figure 2. BAME patients' age distribution curve.



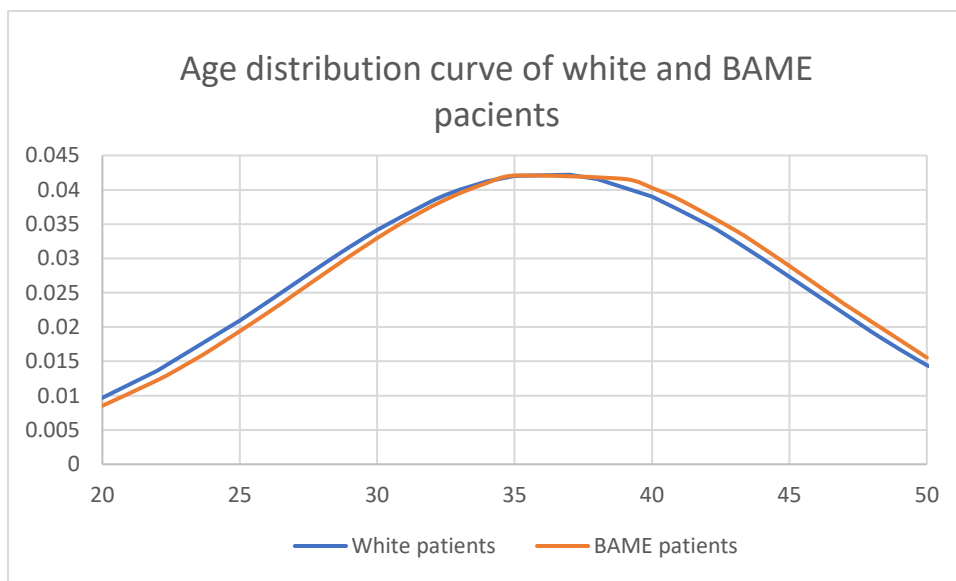
From these observations, it can be noticed that the difference between the average age of white and BAME patients is not substantial. However, for a more precise interpretation, the critical value is needed.

$$t_{\text{obt}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{SD_x^2 + SD_y^2}{n}}} = \frac{36.18 - 36.74}{\sqrt{\frac{9.42^2 + 9.32^2}{50}}} = -0.29888$$

Degree of freedom [4] is $n-2 = 98$, therefore the t_{crit} is 1.98[5] at an alpha level of 0.05.

In the analysis it has been registered a t obtain value of -0.29888, and a t critical value of 1.98, therefore the null hypothesis can be accepted, proving that there is no significant difference between the ratio of the white patients and BAME patients. This result can be observed in the below chart as well.

Figure 3. The distribution curve of white and BAME patients.



Additionally, using excel the formula for T-test [6], assuming unequal variances it can be noticed that it has been registered a P value of 0.77, indicating that there is 77% chance that the data is random and only 23% chance that the data difference is significant, clearly supporting the null hypothesis.

T-Test: Two-Sample Assuming Unequal Variances

	X	Y
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.29888	
P(T<=t) one-tail	0.382832	
t Critical one-tail	1.660551	
P(T<=t) two-tail	0.765665	
t Critical two-tail	1.984467	

Testing the hypothesis at significant levels of 0.10, 0.01, 0.02.

On the following tests it has been registered a P value of 0.765665, indicating a high percentage of chance that the difference is not significant. Moreover, at a significant level of 0.10, the t-critical for two-tail testing reaches a value of 1.66, lower than it was at an alfa value of 0.05, however, still failing to reject the null hypothesis. The tests realised at an alfa level of 0.01 and 0.02 prove, as expecting, an even stronger argument for accepting the null hypothesis. Demonstrating that there is no significant difference between the ratio infection of the two groups of patients.

t-Test: Two-Sample Assuming Unequal Variances
using an alfa level of **0.10**

	X	Y
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.29888	
P(T<=t) one-tail	0.382832	
t Critical one-tail	1.29025	
P(T<=t) two-tail	0.765665	
t Critical two-tail	1.660551	

t-Test: Two-Sample Assuming Unequal Variances
using an alfa level of **0.01**

	X	Y
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.29888	
P(T<=t) one-tail	0.382832	
t Critical one-tail	2.365002	
P(T<=t) two-tail	0.765665	
t Critical two-tail	2.626931	

t-Test: Two-Sample Assuming Unequal Variances, with
an alfa level of **0.02**

	X	Y
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.29888	
P(T<=t) one-tail	0.382832	
t Critical one-tail	2.081445	
P(T<=t) two-tail	0.765665	
t Critical two-tail	2.365002	

Factors that can affect the results

The analysis has been made considering patients with ages between 20 and 50 years, however it can be observed that there is one patient that is 55-year-old, therefore, the sample does not accurately represent the objective of this experiment.

In addition, there is no information about the total population, thus, is hard to know if the selected sample does accurately represent the entire population, what was the method of selection, confidence level and the margin of error. Moreover, the decision of rejecting or not the null hypothesis can be biased, as not all the members of the population have been studies and only a sample of it, thus sampling errors can occur.[7]

Conclusion

The objective of the report is to analyse whether the finding of the BBC UK news is true or not. They claim that the infection rate between white patients and black, Asian and other minority ethnic groups is varied. To determine this, a sample of 100 white and BAME patients has been examined. The analysis was based on the following statistical indicators, the mean between the two groups, the variance, the standard deviation, and the value P , as well as the t obtain and t critical.

The analysis started with a null hypothesis stating that there is no significant difference between the two groups of patients. Following the examination, it was shown that P value has registered 0.765665, a lot higher than the significance level of 0.05, meaning that there is 77% chance that the means between the two groups does not have a significant difference. This has also been concluded by having t obtain value lower than the t critical value. Consequently, the null hypothesis failed to be rejected, demonstrating that there is no significant variance between the infection rate of white patients and BAME patients. However, as only a sample of the population have been studied, the findings can be corrupted due to sampling error.

References

- [1] G. Upton, I. Cook, *Introducing Statistics*. Oxford: Oxford University Press, 2001
- [2] D. Ruppert, *Statistics and Finance*. New York: Springer, 2004.
- [3] Automat Excel, *How to Create a Normal Distribution Bell Curve in Excel*. [Online]. Available: <https://www.automateexcel.com/charts/bell-curve-plot-template>. [Accessed] 03.01.2021
- [4] N. J. Salkind, *Statistics for People Who Think They Hate Statistics*. Los Angeles: Sage Publications, 2010.
- [5] Mera Caulculator, *T Value (Critical Value) Calculator*. [Online]. Available: <https://www.meracalculator.com/math/t-distribution-critical-value-table.php>. [Accessed 02.01.2021]
- [6] M. Harmon, *Two-Independent-Sample Unpooled t-Test Effect Size in Excel*. [Online]. Available: <http://blog.excelmasterseries.com/search?q=t+test>. [Accessed 28-Dec.2020]
- [7] J. H. Kranzler, *Statistics for the Terrified*. Lanham: Rowman & Littlefield, 2017.

Appendix I. z-distribution

A	B	C	D	E	F	G	H	I
X	White patients	Y	BAME patients	X mean	36.18	stnd dev	9.41924	
20	0.009686286	20	0.0085242	y mean	36.74	stnd dev	9.31711	
20	0.009686286	20	0.0085242					
20	0.009686286	22	0.0122502					
20	0.009686286	23	0.014434					
22	0.013638591	24	0.0168122					
22	0.013638591	26	0.022034					
25	0.020939936	27	0.0247926					
25	0.020939936	27	0.0247926					
26	0.023618589	28	0.0275771					
27	0.026341322	28	0.0275771					
27	0.026341322	29	0.0303231					
28	0.029048664	29	0.0303231					
29	0.03167523	30	0.0329605					
30	0.034152176	30	0.0329605					
30	0.034152176	30	0.0329605					
32	0.038382265	30	0.0329605					
33	0.040007757	30	0.0329605					
33	0.040007757	30	0.0329605					
33	0.040007757	31	0.035417					
33	0.040007757	32	0.0376207					
34	0.041234697	32	0.0376207					
34	0.041234697	32	0.0376207					
35	0.042022937	33	0.0395039					
35	0.042022937	34	0.0410061					
35	0.042022937	34	0.0410061					
37	0.042193797	35	0.042078					
37	0.042193797	39	0.0415769					
38	0.041570687	40	0.0402758					
38	0.041570687	40	0.0402758					
40	0.03901031	40	0.0402758					
40	0.03901031	41	0.0385686					
40	0.03901031	43	0.0341667					
40	0.03901031	44	0.031607					
42	0.034993839	44	0.031607					
43	0.03258781	44	0.031607					
43	0.03258781	44	0.031607					
43	0.03258781	45	0.0289042					
43	0.03258781	45	0.0289042					
44	0.030007083	45	0.0289042					
44	0.030007083	46	0.0261298					
44	0.030007083	47	0.0233511					
45	0.027321049	47	0.0233511					
48	0.019272947	47	0.0233511					
48	0.019272947	50	0.0155527					
49	0.016774196	50	0.0155527					
50	0.014435782	50	0.0155527					
50	0.014435782	50	0.0155527					
50	0.014435782	50	0.0155527					
55	0.005754508	50	0.0155527					