

University of Wales Trinity Saint David

**Data Analysis and
Visualization
ACCB4002**

Assignment 2

Lecturer: Anupam Mazumdar

Student: Daniel Manea

Student ID: 1817998

Table of Contents

Introduction	3
Description of the dataset.....	3
The Dataset	3
Description.....	3
Hypotheses	4
Hypothesis Development	4
The null hypothesis	4
The alternative hypothesis	4
Type I and Type II errors	4
t Test	5
Analysis	5
$\alpha = 0.05$	5
Deviation and normal distribution scores	6
$\alpha = 0.01$	8
$\alpha = 0.02$	8
$\alpha = 0.1$	9
Conclusion	9
References	10

Introduction

This report is meant to analyse a dataset in order to identify if the infection rate differs between white and BAME patients. The dataset contains 50 white patients and 50 BAME patients.

Statistical concepts such as null hypothesis, alternative hypothesis, standard deviation, t critical, p-value etc. will be defined and applied.

Description of the dataset

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

Description

In this dataset, we can see that the values are between 20 and 55 for white patients and from 20 to 50 for BAME patients, having a range of 35, and 30, respectively. We can also identify that the most frequent values (mode) for white patients is 20, and for BAME patients is 50.

Hypotheses

Hypothesis Development

A hypothesis is a suggested idea of how to explain a phenomenon but has not yet been tested [1]. It is almost impossible to demonstrate that a hypothesis is true, but much easier to prove that one is wrong, so, to prove that a hypothesis is right, we need to demonstrate that the exact opposite of that hypothesis is wrong [2].

The null hypothesis

The null hypothesis is used to represent everything that the researcher considers to be the opposite of what they are trying to prove [2]. It usually represents the status quo.

In our case, the null hypothesis (H_0) assumes that there is no significant difference between white and BAME patients.

The alternative hypothesis

The alternative hypothesis, also called the research hypothesis and symbolised as H_A , represents what the researcher wants to demonstrate. It tries to prove the fact that there is something new or different. It challenges the status quo [3].

In our case, the H_A is that the infection rate varies from white to BAME patients.

Type I and Type II errors

There are cases in which the conclusion of a researcher is false. There is the possibility that, by chance only, the reality of the population is not reflected by the sample, leading to conclusion errors [4].

The Type I error means rejecting the null hypothesis although the null hypothesis is correct, and the type II error means failing to reject the null hypothesis when it is false. The type I error is also known as a false-positive and the type II error as a false-negative. Type I and II errors cannot be 100% prevented, but the odds of them happening can be reduced by growing the sample size [4].

t Test

The main purpose of the t test, sometimes referred as the Student's t test, is to determine if the means of two groups of scores are different or similar to each other to a statistically significant extent. The main value of t test is that it allows the researcher to test the hypothesis in order to reject or fail to reject the null hypothesis.

Analysis

H_0 : There is no statistically significant difference between white and BAME patients.

H_A : There is a statistically significant difference between white and BAME patients.

$$\alpha = 0.05$$

	<i>White Patients</i>	<i>BAME Patients</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat (~~t_{obt}~~)	-0.298879822	
P(T<=t) one-tail	0.382832305	
t Critical one-tail	1.660551217	
P(T<=t) two-tail	0.76566461	
t Critical two-tail	1.984467455	

The p-value in this t Test indicates that we failed to reject the null hypothesis, given the fact that there is a 76.6% chance for the data to be random, and only 23.4% chance for the data to be significant.

This research is susceptible to a type II (false-negative) error since the null hypothesis failed to be rejected. This could happen as a result of faulty sampling.

Another way to prove that the null hypothesis failed to be rejected is considering that t_{obt} is smaller t Critical, which is also reflected on the table above.

Deviation and normal distribution scores

No.	White (X1)	x1	x1*x1	X1	normal distribution (X1)	BAME (X2)	x2	x2*x2	x1*x2	x2	normal distribution (X2)
1	20	-16.18	261.79	20	0.0097	23	-13.7	188.79	222.31	20	0.0085
2	22	-14.18	201.07	20	0.0097	34	-2.74	7.51	38.85	20	0.0085
3	25	-11.18	124.99	20	0.0097	44	7.26	52.71	-81.17	22	0.0123
4	26	-10.18	103.63	20	0.0097	45	8.26	68.23	-84.09	23	0.0144
5	29	-7.18	51.55	22	0.0136	29	-7.74	59.91	55.57	24	0.0168
6	27	-9.18	84.27	22	0.0136	27	-9.74	94.87	89.41	26	0.0220
7	48	11.82	139.71	25	0.0209	35	-1.74	3.03	-20.57	27	0.0248
8	43	6.82	46.51	25	0.0209	45	8.26	68.23	56.33	27	0.0248
9	55	18.82	354.19	26	0.0236	44	7.26	52.71	136.63	28	0.0276
10	43	6.82	46.51	27	0.0263	30	-6.74	45.43	-45.97	28	0.0276
11	27	-9.18	84.27	27	0.0263	40	3.26	10.63	-29.93	29	0.0303
12	34	-2.18	4.75	28	0.0290	50	13.26	175.83	-28.91	29	0.0303
13	44	7.82	61.15	29	0.0317	50	13.26	175.83	103.69	30	0.0330
14	50	13.82	190.99	30	0.0342	50	13.26	175.83	183.25	30	0.0330
15	37	0.82	0.67	30	0.0342	41	4.26	18.15	3.49	30	0.0330
16	35	-1.18	1.39	32	0.0384	43	6.26	39.19	-7.39	30	0.0330
17	33	-3.18	10.11	33	0.0400	47	10.26	105.27	-32.63	30	0.0330
18	40	3.82	14.59	33	0.0400	32	-4.74	22.47	-18.11	30	0.0330
19	42	5.82	33.87	33	0.0400	31	-5.74	32.95	-33.41	31	0.0354
20	48	11.82	139.71	33	0.0400	20	-16.7	280.23	-197.87	32	0.0376
21	49	12.82	164.35	34	0.0412	26	-10.7	115.35	-137.69	32	0.0376
22	30	-6.18	38.19	34	0.0412	28	-8.74	76.39	54.01	32	0.0376
23	33	-3.18	10.11	35	0.0420	30	-6.74	45.43	21.43	33	0.0395
24	43	6.82	46.51	35	0.0420	39	2.26	5.11	15.41	34	0.0410
25	50	13.82	190.99	35	0.0420	50	13.26	175.83	183.25	34	0.0410
26	50	13.82	190.99	37	0.0422	50	13.26	175.83	183.25	35	0.0421
27	32	-4.18	17.47	37	0.0422	33	-3.74	13.99	15.63	39	0.0416
28	20	-16.18	261.79	38	0.0416	45	8.26	68.23	-133.65	40	0.0403
29	30	-6.18	38.19	38	0.0416	47	10.26	105.27	-63.41	40	0.0403
30	40	3.82	14.59	40	0.0390	20	-16.7	280.23	-63.95	40	0.0403
31	44	7.82	61.15	40	0.0390	32	-4.74	22.47	-37.07	41	0.0386
32	35	-1.18	1.39	40	0.0390	22	-14.7	217.27	17.39	43	0.0342
33	28	-8.18	66.91	40	0.0390	28	-8.74	76.39	71.49	44	0.0316
34	40	3.82	14.59	42	0.0350	30	-6.74	45.43	-25.75	44	0.0316
35	43	6.82	46.51	43	0.0326	34	-2.74	7.51	-18.69	44	0.0316
36	38	1.82	3.31	43	0.0326	32	-4.74	22.47	-8.63	44	0.0316
37	33	-3.18	10.11	43	0.0326	44	7.26	52.71	-23.09	45	0.0289
38	34	-2.18	4.75	43	0.0326	47	10.26	105.27	-22.37	45	0.0289
39	50	13.82	190.99	44	0.0300	30	-6.74	45.43	-93.15	45	0.0289
40	22	-14.18	201.07	44	0.0300	44	7.26	52.71	-102.95	46	0.0261
41	37	0.82	0.67	44	0.0300	29	-7.74	59.91	-6.35	47	0.0234
42	45	8.82	77.79	45	0.0273	40	3.26	10.63	28.75	47	0.0234

43	20	-16.18	261.79	48	0.0193	30	-6.74	45.43	109.05	47	0.0234
44	44	7.82	61.15	48	0.0193	40	3.26	10.63	25.49	50	0.0156
45	33	-3.18	10.11	49	0.0168	24	-12.7	162.31	40.51	50	0.0156
46	20	-16.18	261.79	50	0.0144	50	13.26	175.83	-214.55	50	0.0156
47	38	1.82	3.31	50	0.0144	46	9.26	85.75	16.85	50	0.0156
48	35	-1.18	1.39	50	0.0144	30	-6.74	45.43	7.95	50	0.0156
49	25	-11.18	124.99	50	0.0144	27	-9.74	94.87	108.89	50	0.0156
50	40	3.82	14.59	55	0.0058	50	13.26	175.83	50.65	50	0.0156
Total	1809		4347.38			1837		4253.6		308.34	
Average	36.18					36.74					

$$\bar{x}_1 = \frac{\sum x_1}{n} = \frac{1809}{50} = 36.18$$

$$\text{Var}_{x_1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{4347.38}{49} = 88.72$$

$$\text{SD}_{x_1} = \sqrt{\text{Var } x_1} = \sqrt{88.72} = 9.42$$

$$\bar{x}_2 = \frac{\sum x_2}{n} = \frac{1837}{50} = 36.74$$

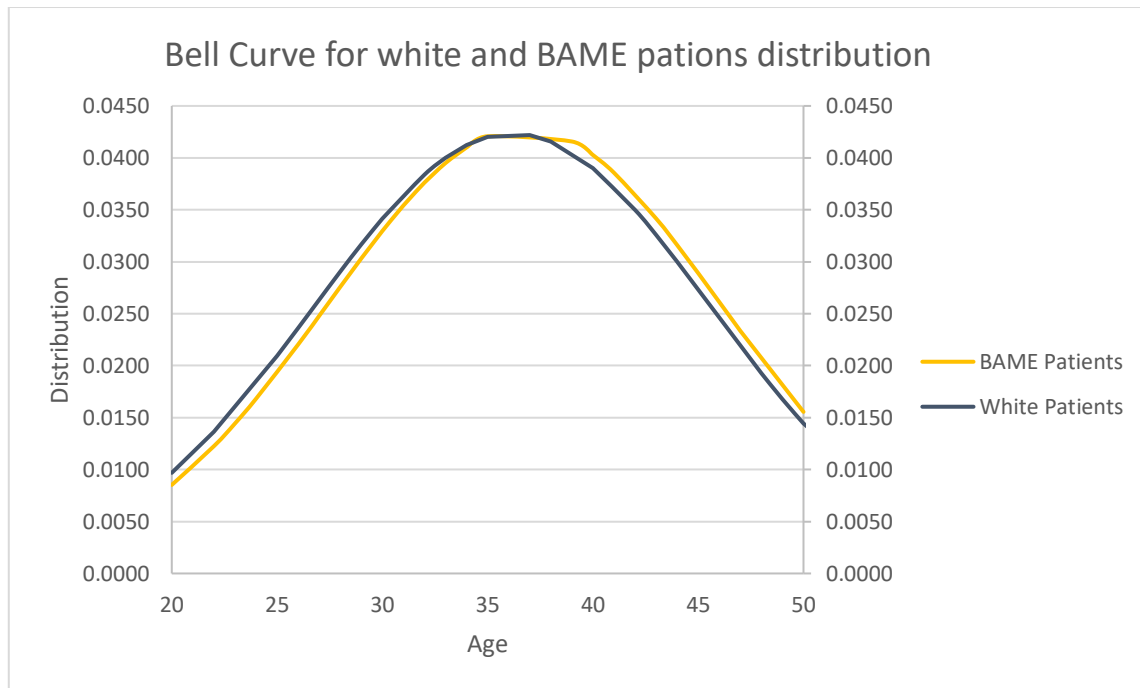
$$\text{Var}_{x_2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{4253.6}{49} = 86.8$$

$$\text{SD}_{x_2} = \sqrt{\frac{\sum x_2^2}{n-1}} = \sqrt{\frac{4253.62}{49}} = 9.32$$

For a better understanding of the dataset, I have also determined the following statistical indicators: mean, deviation scores, squared deviation scores, the variance, standard deviation, and normal distribution parameters in order to create the bell curve for visualising the distribution ratio of the two groups of patients.

The dataset indicated an average age for the white patients of 36.18 years and the average age of the BAME patients at 36.74 years. The variance of the first group of people has registered a value of 88.72 and for the second group of people a value of 86.8, therefore, the standard deviation was determined as 9.42 for the first group and 9.32 for the second group.

Moreover, as showed above, I have determined the normal distribution parameters for X_1 and X_2 , which helped in the visualisation of the distribution ratio between the white and BAME patients' ages. It can be observed that, as proved above, the difference between the two groups of patients is not substantial and are rather following the same patterns.



$\alpha = 0.01$

	<i>White Patients</i>	<i>BAME Patients</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.298879822	
P(T<=t) one-tail	0.382832305	
t Critical one-tail	2.36500241	
P(T<=t) two-tail	0.76566461	
t Critical two-tail	2.626931096	

$\alpha = 0.02$

	<i>White Patients</i>	<i>BAME Patients</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.298879822	
P (T<=t) one-tail	0.382832305	
t Critical one-tail	2.081444921	
P (T<=t) two-tail	0.76566461	
t Critical two-tail	2.36500241	

$$\alpha = 0.1$$

	<i>White Patients</i>	<i>BAME Patients</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.298879822	
P (T<=t) one-tail	0.382832305	
t Critical one-tail	1.290249904	
P (T<=t) two-tail	0.76566461	
t Critical two-tail	1.660551217	

After calculating the significance level of $\alpha = 0.02$, $\alpha = 0.01$ and $\alpha = 0.1$, we can see that even changing the α values, the null hypothesis still cannot be rejected.

Conclusion

In this report, a dataset of 50 white and 50 BAME patients has been analysed in order to determine if the BBC piece of news that suggested that the rate of Covid-19 infection is varied between white and BAME patients is or not statistically correct.

In order to suggest a statistically significant result, t tests with different significance levels were applied, and by all parameters used, we could not fail to reject the null hypothesis.

References

- [1] "Collins Dictionary," [Online]. Available:
<https://www.collinsdictionary.com/dictionary/english/hypothesis>. [Accessed 6 January 2020].
- [2] J. H. Kranzler, *Statistics for the Terrified*, London: Rowman & Littlefield, 2018.
- [3] "AP®/College Statistics," Khan Academy, [Online]. Available:
<https://www.khanacademy.org/math/ap-statistics>. [Accessed 07 January 2021].
- [4] A. Banerjee, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar and S. Chaudhury, "Hypothesis testing, type I and type II errors," *Industrial Psychiatry Journal*, vol. II, no. 18, pp. 127-131, 2009.