

DATA ANALYSIS AND VISUALISATION Assessment 2

Mariana Angelova

Student number: 2008309

Lecturer name: Anupam Mazumdar

Module code: ACCB4002

Programme: BSc Cloud Computing

University of Wales Trinity Saint David

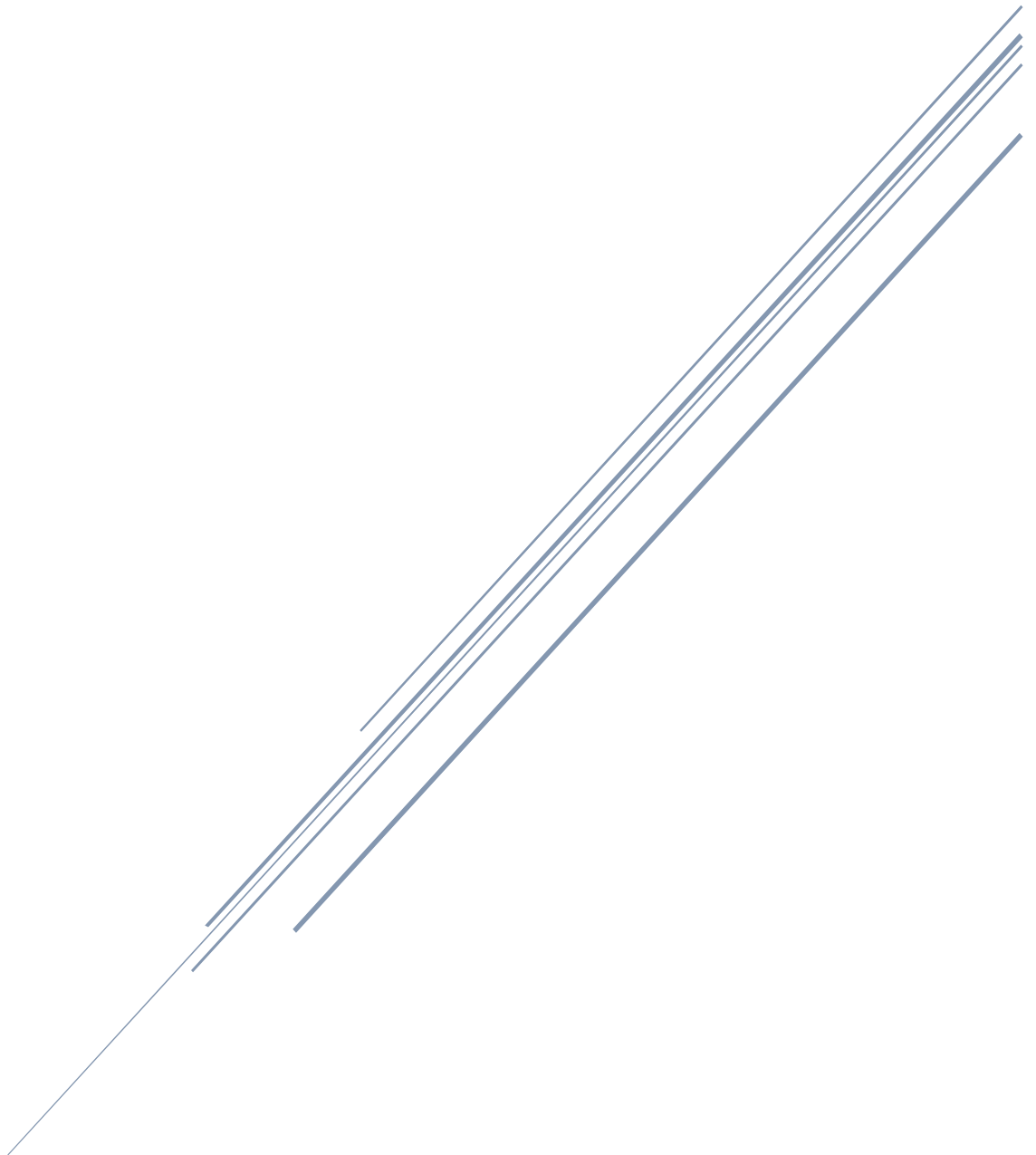


Table of Contents

Introduction	2
Hypotheses	2
What is the t-test?	2
The result of T-test	3
Test the hypothesis to significance levels of 0.1, 0.01 and 0.02	5
Factors that might affect the result	7
Conclusion	8
References	9

Introduction

In this assignment paper I will demonstrate my knowledge about statistic and will use different statistical methods to analyse given me data.

In BBC(UK) news has been said that the infection rate of Covid-19 is varied between white people and BAME in the UK. In this task I must analyse the ratio of infection of COVID-19 between white patients and BAME in the UK and have find this is true or not. All the patients are between 20 and 50 years old. I will use given me data to make my hypotheses.

Hypotheses

According to Gerald Kranzler [1]:when researchers were interested to demonstrate the truth of something – that one population is bigger, stronger, smarter, and so on than another population, they tested their hypothesis. They found that if they want to prove that something is true, the best tool is to prove that is false.

I will need to make my hypothesis as compare the values of white patients and BAME patients. For that I need to find are there statistically significant difference or they are not significant different, it is true or false that infection rate varied between white and BAME patients, can or cannot reject null hypothesis (H_0). What mean the null hypothesis?

"A null hypothesis is a type of conjecture used in statistics that proposes that there is no difference between certain characteristics of a population or data-generating process. [2]" "A primary use of inferential statistic is that of attempting to reject H_0 " [1]

H_0 (null hypothesis) = The infection rate is the same for white and BAME patients

H_1 (alternative hypothesis) = The infection rate is different between white patients and BAME patients

The alternative hypothesis(H_1) states that difference is true.

T-test will help me to decide can I reject null hypothesis or no.

What is the t-test?

T- tests are hypothesis tests which use sample data to infer properties of a population.

As Will Kenton say: “A *t*-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features” [3]. According to him *t*-test is hypothesis testing tool.

According to Gerard Kranzler [1] there have two kinds of *t* tests:

- *T*- test for independent samples- groups which members are independent from each other
- *T*-test for nonindependent samples- two groups whose members are paired in some way

In our case we have two groups which members are independent from each other. I have decided to use MS Excel for my *t*- test.

I have found that MS Excel has two-sample *t*-tests- used to compare two groups of population. One of them is that assumes equal variances and other that assumes unequal variances. I will use first – that assumes equal variances.

The result of *T*-test

In Figure 1 I will show my result of *t*-test made by MS Excel. In statistics characteristics are named variables. First variable show data for white patients. Second variable represent data for BAME patients.

	Variable 1	Variable 2
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Pooled Variance	87.76531	
Hypothesized Mean Difference	0	
df	98	
<i>t</i> Stat	-0.29888	
<i>P</i> (<i>T</i> ≤ <i>t</i>) one-tail	0.382832	
<i>t</i> Critical one-tail	1.660551	
<i>P</i> (<i>T</i> ≤ <i>t</i>) two-tail	0.765665	
<i>t</i> Critical two-tail	1.984467	

Figure 1: *t*-Test: Two-Sample Assuming Equal Variances

Different way to represent the result is with chart.

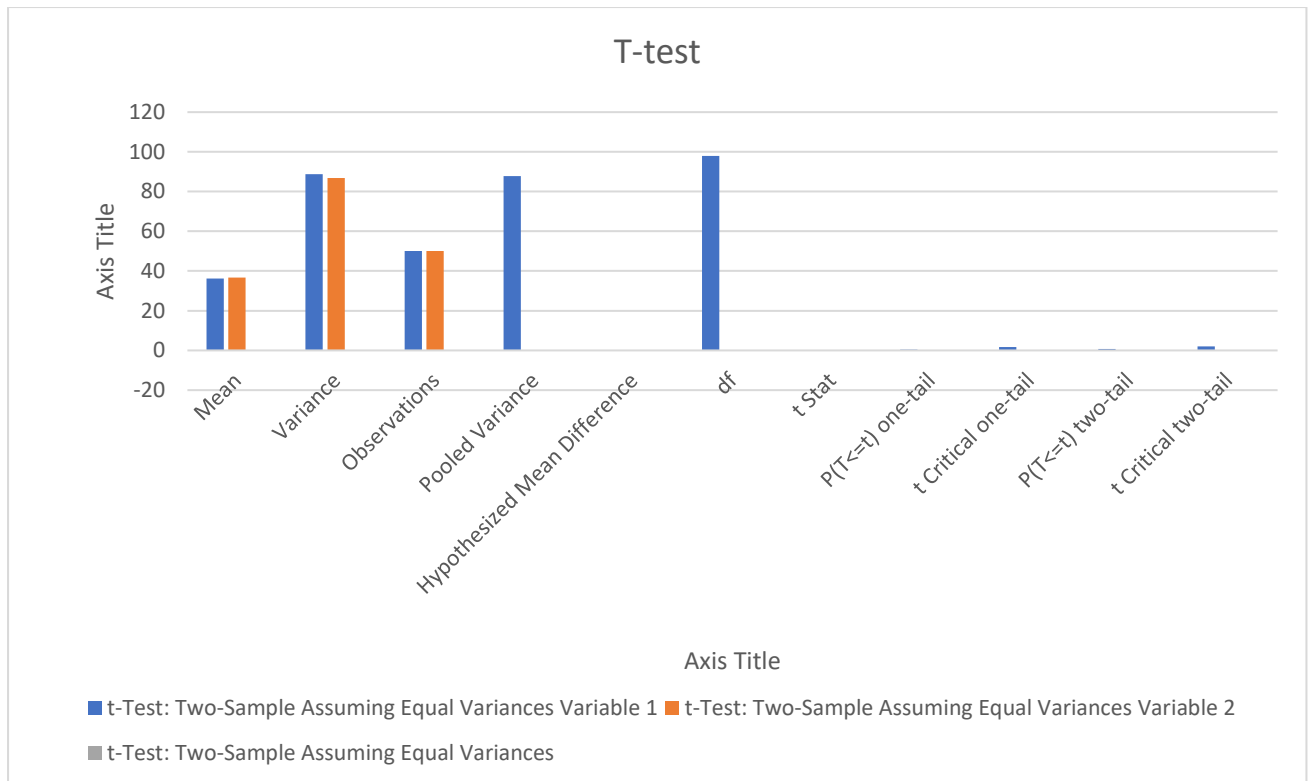


Figure 2

In the result I can see t-test represent three types of data values.

- First is mean value of each data set. The mean is often used measure. We find the mean as adding up all scores and divide them by the number of scores.
- Second row show variances of each variable. I can see that Variance is not exactly equal, but values are close to each other
- Third is number of observations for each group.

What is t-value and df means? Will Kenton say: *“The t-test produces two values as its output: t-value and degrees of freedom. The t-value is a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets.”* [3] and *“A large t-score indicates that the groups are different. A small t-score indicates that the groups are similar.”* [3]

Formula to find df (degrees of freedom) is total number of observations minus 2. In my case one of the variables is 50, another is 50. That mean $50+50-2=98$

T critical is determined by df and selected level of significance. MS Excel find it automatically, but if I don't have MS Excel I must use a table for t critical values.

What is the p value? In our result I can see $P(T \leq t)$ which is the p-value for two tailed form for the t-test. P values are calculated based on the data and assuming that the null hypothesis is true. I will use p value when I test my hypothesis.

My result includes lot of data, but I really look for the two things: T-scores and alpha level. When I did my t-test I wrote my alpha level (or significance level) on 0.05.

I will do two things, to compare my result and decide- can I reject null hypothesis or no.

1. I will compare the significance level which I choose (0.05) to the p-value in the output. I can see that my p-value is bigger than my alpha level. That mean I cannot reject null hypothesis.
2. I will compare the t-critical two-tail value in the output (Figure 1) with t start. I can see that t critical value (1.984467) is bigger than t start (-0.29888) That mean I cannot reject null hypothesis again.

Test the hypothesis to significance levels of 0.1, 0.01 and 0.02

Let see what will happen if I change the significance level in my t- test. First, I will made t-test on significance level of 0.1, after that with 0.01 and 0.02.

	<i>Variable</i> <i>1</i>	<i>Variable</i> <i>2</i>
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Pooled Variance	87.76531	
Hypothesized Mean Difference	0	
df	98	
t Stat	-0.29888	
$P(T \leq t)$ one-tail	0.382832	
t Critical one-tail	1.29025	
$P(T \leq t)$ two-tail	0.765665	
t Critical two-tail	1.660551	

Figure3 (t test with significance level 0.1)

I can see on this output that only t Critical values are change. When I compare my t-start value with new t Critical value I can see that t critical value are bigger than t start. That mean on significance level of 0.1 I cannot reject null hypothesis.

Let see what will be result for significance level of 0.01.

	<i>Variable</i>	<i>Variable</i>
	<i>1</i>	<i>2</i>
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Pooled Variance	87.76531	
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	

Figure 4(t test with significance level of 0.01)

I can see that only t Critical values are change again. Because they are bigger than t start value, I cannot reject null hypothesis again.

Let see what happen with result if I change significance level again on 0.02.

	<i>Variable</i>	<i>Variable</i>
	<i>1</i>	<i>2</i>
Mean	36.18	36.74
Variance	88.72204	86.80857
Observations	50	50
Pooled Variance	87.76531	

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

Figure 5 (t test with significance level of 0.02)

The result is the same, only T critical values are change, but I cannot reject null hypothesis because t Critical value is bigger than t- start value.

According to Gerald Kranzler [1] when I write my conclusion about H_0 I may be make two type of errors known as Type 1 error and Type2 error.

- Type 1 is rejection of true H_0
- Type 2 is failure to reject a false null hypothesis

When I made my test with significance level of 0.01 and 0.02 (1% and 2%) maybe I increase a chance to do Type 2 error, because I insist more stringent conditions. Reducing the probability of Type 1 error to only 1% or 2% and that mean I will fall to reject H_0 . The level of 0.05 is set as significance minimum standard of the null hypothesis. What about significance level of 0.1? With the level of 0.1 the result will be not statistically significant, because it does not represent real differences between population.

Factors that might affect the result

The result of the statistic test can change. That can be happen if we change the number of observation and how variable are observations.

What will happen if I change the number of observations? For example, if I take only 5 numbers of my table with white and BAME patients and make another t-test. Is the result will change?

	Variable 1	Variable 2
Mean	24.4	35
Variance	12.3	90.5

Observations	5	5
Pooled Variance	51.4	
Hypothesized Mean Difference	0	
df	8	
t Stat	-2.33773	
P(T<=t) one-tail	0.023792	
t Critical one-tail	1.859548	
P(T<=t) two-tail	0.047584	
t Critical two-tail	2.306004	

Figure 6 (change the number of observations)

First, I can see the mean of the variables are change. Second, variance change and now have a big difference between variables. Third, p value is smaller than alpha value of 0.05. That mean I can reject null hypothesis, but that do not mean that alternative hypothesis is true.

This result show how result of the test change when I change the number of observations.

According to Gerald Kranzler [1] my decision to reject or no null hypothesis can be error. That is because I cannot observe all members of population and because statistics are subject to sampling error, I can be never sure whether null hypothesis is true or not.

If I make another observation with more people from the same groups, I am sure result will change again. If there is big difference between these two groups – white and BAME I will make error if I say there is no difference. On other side I cannot reject and null hypothesis because that will be subject of error.

Conclusion

According to Gerald Kranzler [1] the purpose of inferential statistics is help us to take right decision to reject or no null hypothesis and estimate the probability of errors when we take our decisions. Inferential statistics cannot tell us for sure whether we made errors when we make our decisions but can tell us how likely is that we have made one.

In conclusion I can say that the result of my research was very interesting for me. I read in the media about this difference of ratio of infection between white patients and BAME patients. The result of my research show that there

no significance difference between these two groups. It shows that we are all humans, we can get sick and it does not depend on skin colour.

References

- [1] J. M. Gerald Kranzler, in *Statistics for the terrified*, New Jersey, USA, Prentice-Hall, Inc. Upper Saddle River, 1999, pp. 74-100.
- [2] A. Hayes, "www.investopedia.com," 24 October 2020. [Online]. Available: https://www.investopedia.com/terms/n/null_hypothesis.asp.
- [3] W. Kenton, "Investopedia," [Online]. Available: <https://www.investopedia.com/terms/t/t-test.asp>. [Accessed 22 March 2020].
- [4] B. Beers, "www.investopedia.com," 19 October 2020. [Online]. Available: <https://www.investopedia.com/terms/p/p-value.asp>. [Accessed 23 Dec 2020].