



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

BSc Cloud Computing (Term 1)

Data Analysis and Visualisation /ACCB4002/

Assessment 2

/Analyze the ratio of infection rate and truthfulness of the findings/

Author:

Petar Ivanov

Lecturer:

Anupam Mazumdar

Student ID:

2013025

Date:

03.01.2021

Table of Contents

Introduction	2
What is a hypothesis?	2
Collecting the data	3
Statistical testing	3
Accept or Reject	4
Presenting our findings	4
T-test	5
Analysis	5
Test #1 (Significance level (p) / Alpha value / = 0.01)	7
Test #2 (Significance level (p) / Alpha value / = 0.02)	7
Test #3 (Significance level (p) / Alpha value / = 0.05)	8
Test #4 (Significance level (p) / Alpha value / = 0.1)	8
Conclusion	9
References	10

Introduction

When we are given a certain set of data and are required to analyze that data in order to export some hidden information, we often struggle to pick the right approach to achieve our goal. Scientists have figured out some ways to analyze data that are pretty accurate. In statistics, hypothesis testing is the terminology used to describe the procedure for investigating everything about the world. It's mostly used to test hypothesis – predictions arising from theories [1]. Hypothesis testing mainly consists of 5 stages:

- Stating the “research hypothesis” as a null (H_0), and the alternative (H_a or H_1);
- Collecting the data;
- Statistical testing;
- Decide if the H_0 is accepted or rejected;
- Presenting the findings and discussing results.

What is a hypothesis?

Shona McCombes gave a definition of a hypothesis in her article “How to write a hypothesis”, stating “A hypothesis states your predictions about what your research will find” [2]. She added that it should not be just a guess, but it should be based on existing theories and knowledge, and, of course, the hypothesis has to be testable – supported or refused by scientific research methods. Hypothesis consists of a relationship between two or more variables, mostly one being called an “independent variable” – changeable by the researcher to observe how it affects the test – and another one called “dependent variable” – the variable that's being observed and measured. In order to develop a hypothesis, we must firstly have the research question that needs to be answered. Logically, the question must be researchable, specific and focused. Next, we can try and give an answer based on what we already know. We can even look for theories and previous studies, answering the question, and try to give an initial answer in a clear and a straightforward sentence. Next step is refining and rephrasing the hypothesis, making sure it contains the specific group and relevant variables that are being studied, and

the predicted outcome of the experiment or analysis. Phrasing can be done in three ways by using the if...then form. When it comes to academic researches, most of the times, hypothesis is phrased in terms of correlations or effects. When it comes to comparing two groups, the difference expected to be found is phrased in the hypothesis. Finally, we can write the null hypothesis – the default position – stating that there is no significant association between the dependent and independent variables, and the alternative hypothesis stating that there is a link (association) between the two variables.

Collecting the data

Collecting the right samples means that you can make an accurate test and the results will be satisfying and not misleading. For example, if you were to test the difference between average height in men and women, you should consider using the same age group, ethnicity, country of origin, etc. because, if for example you use different people from all across the globe, your results will vary as average male height in Guatemala is 160.9 cm (5 ft 3 1/2 in) [3], but in Sweden, the average male height is 181.5 cm (5 ft 11 1/2 in) [4], and by using this type of data, you may have misleading results, depending on what you are testing for. Overall – your data should be representable in order for it to be accurate.

Statistical testing

The variety of statistical tests makes it hard for one to know how to approach and what to do with the data he has collected. The choice may be narrowed down due to the type of collected data. What all tests have in common is how spread out the data is within a certain category (or we can say this is the within group variance), and how the categories differ from one-another (or we can say this is the between group variance). The statistical result low “p-value” means that there is no overlapping between groups, with the between group variance being large enough. Meaning, that the differences between groups coming about by chance is highly unlikely. On the other hand, having high “p-value” means that the difference we measure

between groups is due to chance thus having a low between and a high within group variance.

Accept or Reject

Based on the result, we have to either accept or reject our null hypothesis. Most of the times the p-value guides our decision. In most cases our alpha value is 0.05 (also referred as cutoff) – equal to less than 5% chance to see these results if the null hypothesis were to be true. For instance, using the same example for the average height between man and women, we have a p-value result equal to 0.004. Based on the difference the numbers have, we decide to reject the null hypothesis of no difference (0.004 being less than our alpha value of 0.05).

Presenting our findings

Giving a brief summary of the data in the “results” section could be the difference between the mean values, or the p-values, or anything we consider a weight in our decision. In the discussion, we discuss the action we took with the null hypothesis. However, there is a difference if we are writing a statistical report or an academic one. In the statistical report, we say that we either accept or reject the null hypothesis, but if we are talking about an academic report, the results in our research paper will be referring to our alternative hypothesis – meaning, that if we reject the null hypothesis, we write what the alternative hypothesis is stating. We do not write “accept or reject”, because hypothesis testing is not designed to prove or disapprove anything. It’s a test, showing whether a pattern we measure could have arisen by chance.

T-test

Using a t-test, we can determine whether or not there is a significant difference between the means of two groups. It's being used when datasets have unknown variances and the data would follow normal distribution. A T-test, being a type of inferential statistics tool is also used as a hypothesis testing tool based on assumptions [5]. Essentially, the T-test gives us an idea if the two compared groups' average values are equal and they come from the same population. Assuming a null hypothesis of the two means being equal, the t-test takes a sample from each group, applies mathematical formulas and by calculating the data, accepts or rejects the hypothesis.

In order to calculate a t-test, we require three key values: one being the mean difference between the two groups, the second being the standard deviation of each group, and the last one – the number of data values. To determine the chance alone, we use the calculated t-value from the outcome of the t-test, compared to the value obtained from the critical value table – available in one and two – tail formats. The one-tail format is greatly used when having a fixed range or clear direction (positive/negative). The two-tail format is used for range bound analysis. Apart from the t-value, we also get degrees of freedom from the output of the calculations.

Analysis

I'm going to find out the truthfulness of the dataset sample consisting of 50 white and 50 BAME patients aged between 20 and 50 years old, stating that the infection rate of COVID – 19 is varied between white people and BAME in the UK. I will be using a t-test, to compare the two means of the datasets. Our sample consists of 50 White and 50 BAME patients, which leads to a conclusion that we should use Equal Variance T-Test (Pooled). This test is used when the number of samples in each group is the same, or the variance of both groups is similar.

Firstly, we need to define the Null and Alternative hypothesis (as we already know, the null hypothesis is the one we hope to reject):

- H_0 : There is no significant Covid-19 infection rate variation between White and BAME patients in the UK, aged between 20 and 50;
- H_1 : There is a significant Covid-19 infection rate variation between White and BAME patients in the UK, aged between 20 and 50;

In order to calculate the T-Values and Degrees of Freedom, we use this formula:

$$T\text{-Value} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{n}}}$$

$$df = n1 + n2 - 2$$

Where:

$\overline{X}_1$ = Mean of White Patients = 36.18;

$\overline{X}_2$ = Mean of BAME Patients = 36.74;

SD_1^2 = Variance of White Patients = 88.72;

SD_2^2 = Variance of BAME Patients = 86.80;

n = Number of items in sample (n1+n2)

n1=n2=50

Or we can simply use the Excel's built-in function in Data>Data Analysis and get these results:

Test #1 (Significance level (p) / Alpha value / = 0.01)

t-Test: Two-Sample Assuming Equal Variances

	<i>X - WP (White Patients)</i>	<i>Y - BAME</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Pooled Variance	87.76530612	
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	

Test #2 (Significance level (p) / Alpha value / = 0.02)

t-Test: Two-Sample Assuming Equal Variances

	<i>X - WP (White Patients)</i>	<i>Y - BAME</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Pooled Variance	87.76530612	
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	

Test #3 (Significance level (p) / Alpha value / = 0.05)

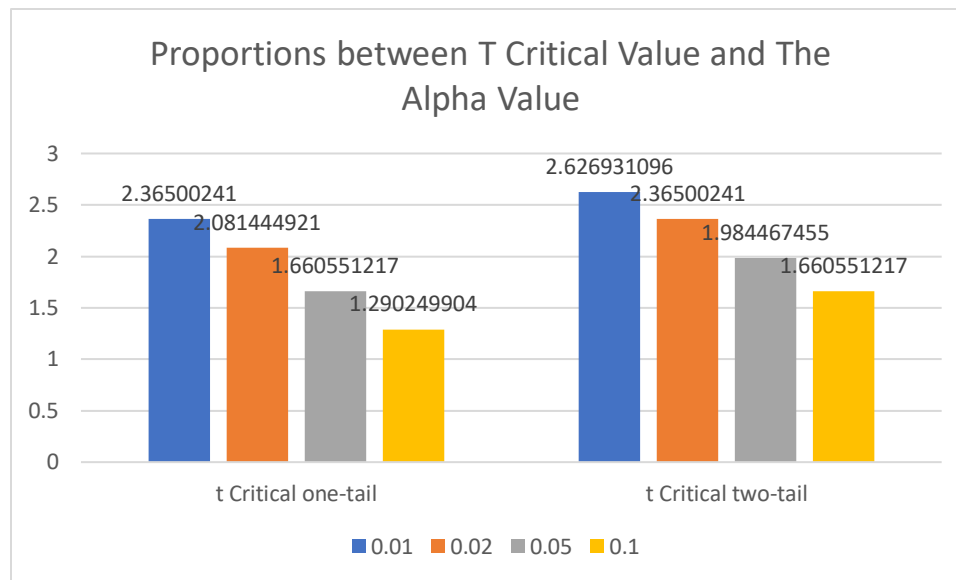
t-Test: Two-Sample Assuming Equal Variances

	<i>X - WP (White Patients)</i>	<i>Y - BAME</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Pooled Variance	87.76530612	
Hypothesized Mean Difference	0	
df	98	
t Stat	-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	

Test #4 (Significance level (p) / Alpha value / = 0.1)

t-Test: Two-Sample Assuming Equal Variances

	<i>X - WP (White Patients)</i>	<i>Y - BAME</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857143
Observations	50	50
Pooled Variance	87.76530612	
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	



Conclusion

The tests show that by changing the significance value (alpha value) everything but the t Critical one-tail and t Critical two-tail values stays the same. As we can see, the more we increase the significance value (alpha value) the more the t Critical one and two tail values are decreased.

In all of the cases, the t-stat value is less than the t-critical one tail value so we cannot reject the null hypothesis.

The second piece of information we need is the P-Value (for the one-tail value), and again, in all of the cases, the p-value is greater than the alpha value (significance value) and this means that we cannot reject the null hypothesis.

References

- [1] R. Bevans, (September 25 2020), *A step-by-step guide to hypothesis testing*, Scribbr, Available: <https://www.scribbr.com/methodology/hypothesis-testing/>
- [2] S. McCombes, (October 22 2020). *How to write a hypothesis*. Scribbr. Available: <https://www.scribbr.com/research-process/hypotheses/>
- [3] Informe Final, *V Encuesta Nacional de Salud Materno Infantil*. 2008-2009 National Institute of Statistics (Guatemala) (in Spanish). 2011. pp. 16, 337, 358. Available: <https://www.ine.gob.gt/sistema/uploads/2014/01/22/LYk4A1kGJAO7lvfS0Aq6tezcUa9tQh35.pdf>
- [4] "Svenskarna längre och tyngre". Dagens Nyheter. (Feb 29 2008). Available: <http://www.dn.se/nyheter/sverige/svenskarna-langre-och-tyngre>
- [5] W. Kenton, (Mar 22 2020), *T-Test*, Tools for fundamental analysis, Investopedia, Available: <https://www.investopedia.com/terms/t/t-test.asp#what-is-a-t-test>